THE UNIVERSITY of EDINBURGH
School of Biological Sciences

[Analysis of NCOR1-LYN gene microarray data using R]

[Functional Genomic Technologies]

Student Exam Number: ……B196466……

The file name of your uploaded document **must** include your exam number

Analysis of NCOR1-LYN gene microarray data using R

1.Introduction

NCOR1-LYN was identified as a novel fusion gene in paediatric acute lymphoblastic leukaemia patients, and leukaemic cells expressing NCOR1-LYN were found to be sensitive to dasatinib in vivo.[1] To reveal which pathways are affected by NCOR1-LYN, this trial used microarrays for gene expression analysis.[2]

2.Data

NCOR1-LYN-expressing Ba/F3 cells were created by inducing NCOR1-LYN expression in Ba/F3 cells using a tetracycline-inducible gene expression system.[1] Finally, Ba/F3 cells expressing DOX-induced NCOR1-LYN were grown in triplicate alongside uninduced Ba/F3 cells.[1]

3.Method

3.1Quality control

Quality control checks the distribution of expression values to avoid abnormal chips. In the present trial, quality control is performed by examining a a density plot , a boxplot and a MA plot.

3.2 RMA Normalisation

The unification of data between distinct groups is known as normalisation. This is due to the background correction of the data between groups, which removes part of the noise.[4] However, due to internal or external reasons, it is impossible to eliminate overall variations in a group. If this is not done, it will be impossible to compare groups since general increases and decreases in a group will cause the size relationship between groups to be disrupted.[4]

3.3 Hierarchical clustering of normalised data

Cluster analysis is mainly used to understand unknown functions. By clustering analysis, genes are grouped into categories based on certain characteristics. Genes aggregated into the same type have similar patterns of expression, and they are likely to have some similar functions. The function of other unknown genes can then be further analyzed by understanding the known processes of genes in a particular category. Clustering refers to the grouping of genes according to different functions or the same expression behavior based on gene expression data from gene chips.

3.4 Principal Components Analysis of normalised data

The principal component analysis method allows visualization of the clustering of samples on the gene microarray, allowing visualization of the distribution of samples between experimental and control groups, thus facilitating the detection and removal of abnormal samples, the presence of which would otherwise adversely affect subsequent analysis such as the identification of differential genes.[5] The purpose of the principal component analysis is to project high-dimensional data into a lower dimensional space. For a complex thing described by several variables, the main aspects of the item are reflected in a few main variables, which are then separated out and analyzed in detail.[5]

3.5 Statistical analysis using Limma

Establish a suitable design matrix first; there are generally multiple equivalent procedures that can be utilized to create an acceptable design matrix for a given experiment.[3] The capacity of the limma linear model technique to tolerate arbitrary levels of experimental complexity is one of its main strengths. The comparison matrix then specifies which sample comparisons should be made. The lmFit and contrasts are then used to carry out the linear modeling of limma. Fit functions, with each gene's expression data being fitted to a different model. Then, using information from the entire gene, an empirical Bayesian correction is conducted, allowing for a more precise estimation of gene variability. After that, a summary table is created.

3.6 Functional Enrichment Analysis using Limma

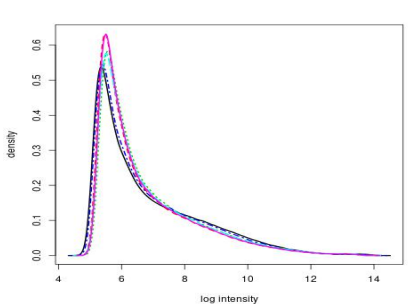
Enrichment analysis refers to statistical analysis with the help of various databases and analytical tools to mine the database for functional categories of genes that have significant relevance to the biological questions we want to study and is a method of differential expression analysis by using the limma package to analyze up-or down-regulation of gene functions and expression pathways.

4. Result

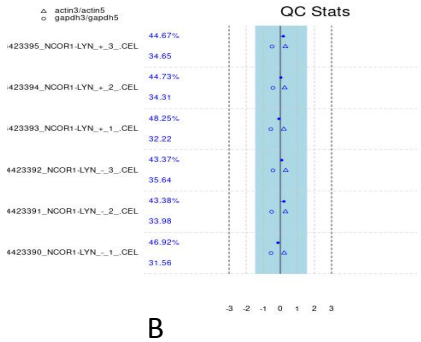
Before analysing, it is important to note in advance that because there was one factor, there were two sets of replicates, each containing three samples. And one sample from each array was hybridised, so there were six samples in total. And they have been named B-1, B-2, B-3, B+1, B+2 and B+3 respectively.

4.1 Quality control

As shown in Figure 1 A, the first one is a density plot by quality control. Direct observation of this plot shows only one peak. Still, the peak is downward and to the left, demonstrating that the data do not differ significantly between the six gene chips, but there are differences between the two groups. It is not necessary to remove the non-conforming data, but further normalisation is required. And then, we use the average method of quality control and the figure shown in Figure 1 B; in short, all indicators appearing in blue are normal, and red may have quality problems, and it can be inferred that there are no quality issues with these samples.



A



B

Figure 1 A:Density profile before data normalisation. B: Average method: Column 1 is the name of the sample; column 2 is two numbers,

the upper is the detection rate and the lower is the mean background noise; column 3 ("QC Stats"): the bottom horizontal axis is the coordinate corresponding to the scale factor etc.

Before RMA normalisation, Figure 2A shows that the median for each sample is very close, but with some trailing and different heights, requiring normalisation. However, overall there is little difference in the distribution of expression across the samples. With RMA normalization, Figure 2B shows that the heights of all samples are neat, the median values of each sample are consistent, and the overall distribution of each sample is the same; this data is more conducive to the next step of differential expression analysis.

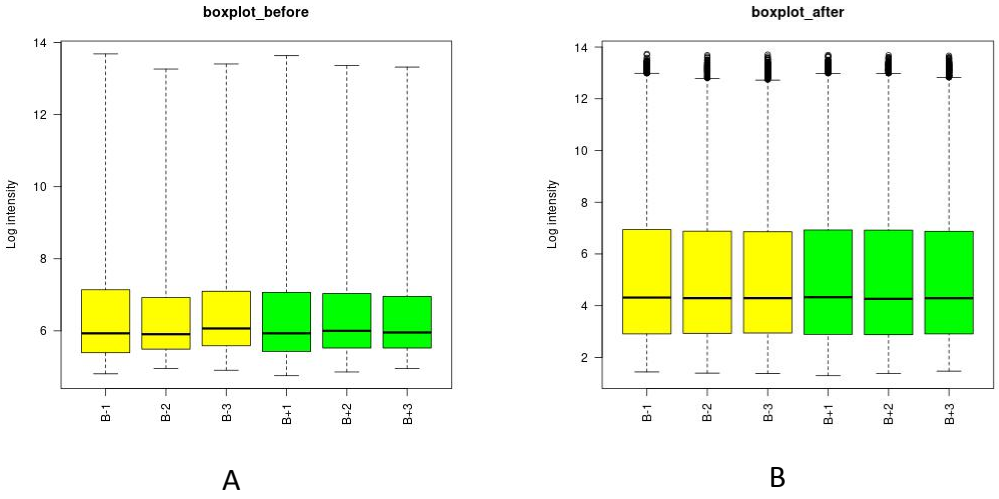


Figure 2 A:The Boxplot of data before normalization B: The Boxplot of data after normalization

MA plots reflect the distribution of gene expression differences with gene signal intensity in the compared samples and are used to show if the data expression is abnormal. And the normalisation was effective if most points in the MA plot were around 0 after normalisation, by pairing any two of the six samples after which we made a total of 15 comparisons. As Figure 3 shows, plot A is the MA plot before normalization, while plot B is the MA plot after normalization. Plots C and D are individual cases before and after normalization, respectively. From these plots, we can find that the symmetry of the upper and lower axes indicates that the number of highly expressed up-regulated and down-regulated genes is similar, and the data are normal. The majority of the points in panel B are indeed closer to zero than those in panel A, indicating a significant normalization effect. This is also evidenced by the fact that the median and IQR (interquartile range) in the right plot are smaller than those in the left plot.

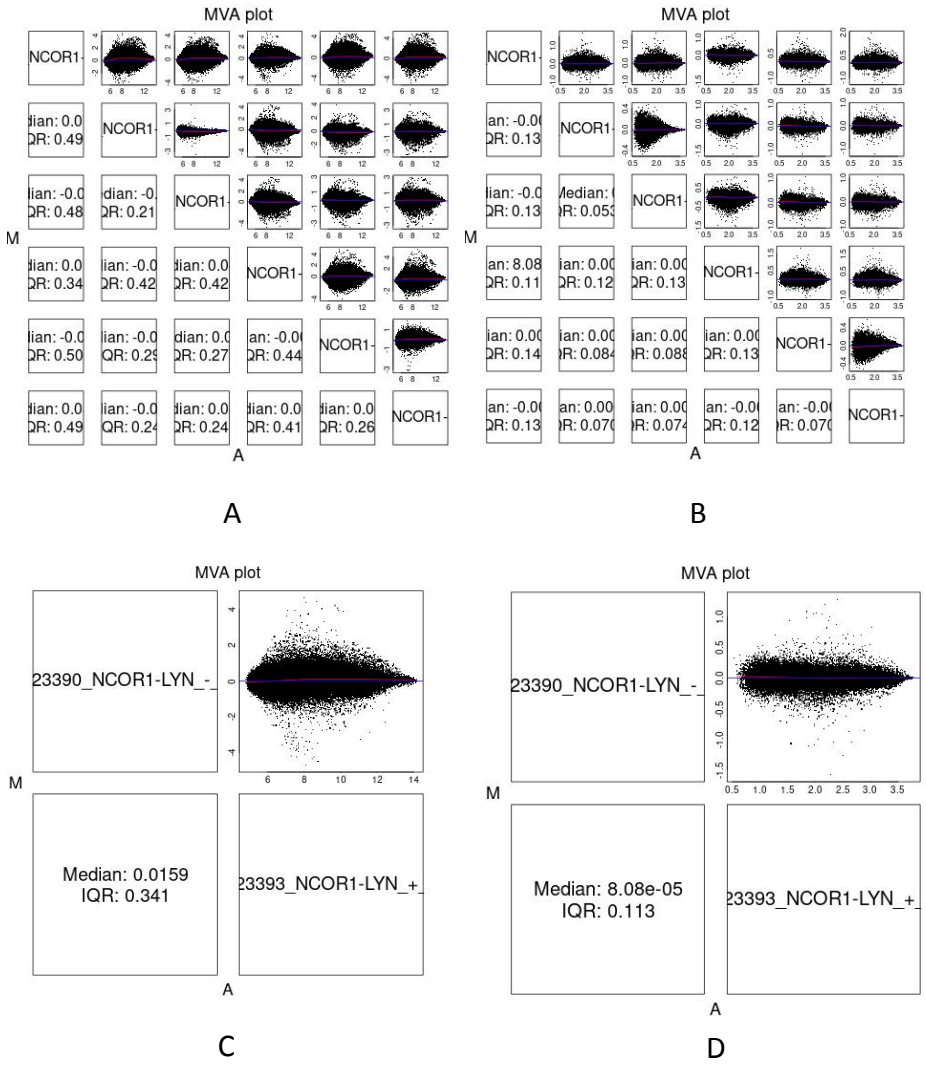


Figure 3 A: MA plot before normalisation which include all samples B:MA plot after normalisation which include all samples C: MA plot before normalisation which include one sample D: MA plot after normalisation which include one sample

Clustering measures the similarity between samples, the more similar the samples are, the closer they are together. Figure 4 indicates that the overall expression differences within groups are more significant than the expression differences between groups. There is a good similarity between B-1 and B+1 and also among B-2, B-3, B+2, and B+3.

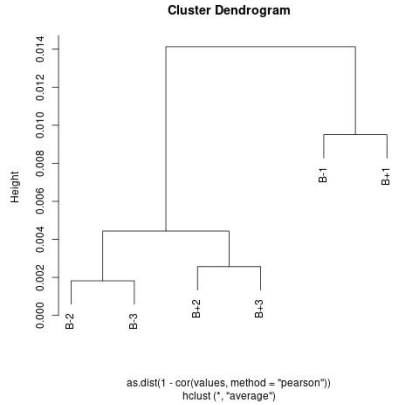


Figure 4 Sample Clustering Chart

As shown in Figure 5, we can visualize the first three principal components and examine the distribution of these six samples in the space of the first three principal components. It shows there is not much difference between B-2 and B-3 in the main ingredients. The major components of B-1 and B+1 differed significantly, although similarities existed between this gene above.

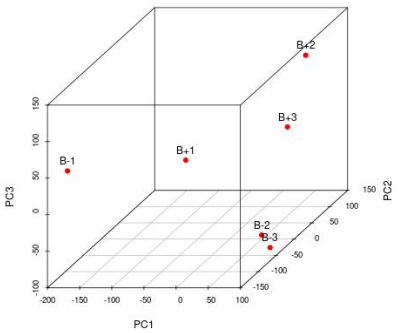


Figure 5 PCA plot

The top ten gene IDs for fold change values are shown in Table 1. Because the general screening conditions were Fold change ≥ 2 and Fold change < 0.01, and the genes screened were significantly different.

Probe id	control.mean	case.mean	Fold change(control_case)
1424727_at	107.886167	9.829117	10.97618120
1419609_at	357.062838	37.722831	9.46543061
1424542_at	1934.847847	228.178827	8.47952403
1449431_at	41.145644	5.119722	8.03669429
1435761_at	68.167410	10.104173	6.74646074
1449456_a_at	73.104654	11.722572	6.23622999
1421375_a_at	36.482867	6.459311	5.64810500
1449254_at	1822.622458	333.553351	5.46426067
1436329_at	50.162659	10.123388	4.95512566
1457105_at	42.389799	8.777904	4.82914797

Table 1 The top 10 fold change table

4.2 Statistical analysis

P value means significant difference, P value less than 0.05 means significant difference. adj. P is the FDR-adjusted P value which is more representative. logFC, FC is the up-regulation multiple of differentially expressed genes, generally logFC more than twice the gene expression is significantly different. As shown in Table 2, there are three gene IDs with P values less than 0.05 and more than twice the LogFC value, namely 1450009_at, 1421375_a_at and 1418601_at. Meanwhile, the high fold change of 1421375_a_at and its appearance in Table 1 also demonstrates that the differential expression of this gene is more significant. The functions of these three genes expressed, carrying pathways thought to influence NCOR1-LYN gene expression.

Probe ID	Symbol	Name	logFC	AveExpr	P.Value	adj.P.Val
1450009_at	Ltf	lactotransferrin	−2.851444	6.069775	1.967493e−07	0.004292482
1417266_at	Ccl6	chemokine (C−C motif) ligand 6	−1.939181	9.246508	2.417671e−07	0.004292482
1420249_s_at	Ccl6	chemokine (C−C motif) ligand 6	−1.968971	7.907373	2.855246e−07	0.004292482
1416871_at	Adam8	a disintegrin and metallopeptidase domain 8	1.788668	7.362394	4.244837e−06	0.042653961
1448898_at	Ccl9	chemokine (C−C motif) ligand 9	−1.161509	8.276769	5.614028e−06	0.042653961
1421375_a_at	S100a6	S100 calcium binding protein A6 (calcyclin)	2.459766	3.921196	5.674459e−06	0.042653961
1418601_at	Aldh1a7	aldehyde dehydrogenase family 1, subfamily A7	−2.406611	4.218004	6.703885e−06	0.043193133
1417936_at	Ccl9	chemokine (C−C motif) ligand 9	−1.237330	8.837447	1.014032e−05	0.057167296
1435657_at	Ston2	stonin 2	−2.253225	5.165244	1.396790e−05	0.058976265
1424902_at	Plxdc1	plexin domain containing 1	−1.112603	5.542212	1.423799e−05	0.058976265

Table2: The top 10 differentially expressed genes identified by Limma

As shown in Figure 6A, the number of differential genes accounted for 1.6% of all genes (720/45101), indicating that whether the NCOR1-LYN protein was induced by DOX into Ba/F3 cells had a slight effect on the difference in gene expression. And as shown in Figure 6B, the horizontal coordinate represents the logarithmic value of the fold difference in gene expression levels between the two samples, i.e. log2(FC); the larger the absolute value of the horizontal coordinate, the larger the fold difference in expression between the two samples. The vertical coordinate represents the negative logarithm of the p-value, i.e. -log10 (p-value). The larger the value of the vertical coordinate, the more pronounced the differential expression and the more reliable the differentially expressed genes obtained from the screen.[3] The number of up- and down-regulated genes is similar, and both are few and far between, and are the very same three genes mentioned earlier, which are highly differentially expressed.

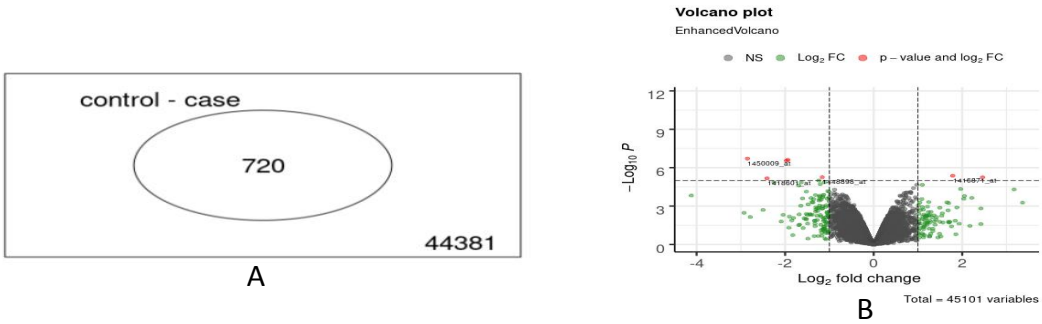


Figure 6 The situation of genes distribution (A: The Venn diagram; B: The volcano plot)

4.3 enrichment analysis

mroast table is a differential expression analysis method used to determine whether up/down regulated genes are significantly enriched in the target gene set and to show the distribution of enrichment. table 3 is arranged in descending order by FDR, all genes in this table are up regulated in expression and all have a p value less than 0. 05, the first three enrichment signals with FDR less than 0.05 are considered to be of high confidence.

enrichment signatures	NGenes	Direction	PValue	FDR
HALLMARK_APOPTOSIS	485	Up	0.0010	0.03750000
HALLMARK_UV_RESPONSE_UP	615	Up	0.0020	0.03750000
HALLMARK_KRAS_SIGNALING_UP	640	Up	0.0025	0.03750000
HALLMARK_IL2_STAT5_SIGNALING	627	Up	0.0065	0.06750000
HALLMARK_INFLAMMATORY_RESPONSE	569	Up	0.0070	0.06750000
HALLMARK_ESTROGEN_RESPONSE_EARLY	662	Up	0.0095	0.07708333
HALLMARK_COAGULATION	368	Up	0.0115	0.08035714
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	643	Up	0.0155	0.08375000
HALLMARK_MYOGENESIS	705	Up	0.0165	0.08375000
HALLMARK_HYPOXIA	663	Up	0.0170	0.08375000

Table 3: mroast table

camera is a gene function enrichment, and Table 4 shows that of the first eight enrichment signals with an FDR less than 0.05, the first six have a small FDR with the highest confidence, and all are almost always showing downregulation. There is both a plausible upward and downward signal, which is similar to the previous volcano chart results.

enrichment signatures	NGenes	Direction	PValue	FDR
HALLMARK_E2F_TARGETS	594	Down	6.197520e-06	0.0002717476
HALLMARK_MYC_TARGETS_V1	624	Down	1.086991e-05	0.0002717476
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	351	Down	8.428413e-04	0.0140473554
HALLMARK_OXIDATIVE_PHOSPHORYLATION	465	Down	1.299536e-03	0.0162441955
HALLMARK_MYC_TARGETS_V2	129	Down	2.234266e-03	0.0193850579
HALLMARK_P53_PATHWAY	621	Up	2.326207e-03	0.0193850579
HALLMARK_G2M_CHECKPOINT	722	Down	6.090273e-03	0.0435019525
HALLMARK_PANCREAS_BETA_CELLS	135	Up	7.565198e-03	0.0472824863
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	643	Up	1.518228e-02	0.0778552201
HALLMARK_TNFA_SIGNALING_VIA_NFKB	529	Up	1.648893e-02	0.0778552201

Table 4: camera table

5. Discussion

When quality control analysis is performed, it is possible to learn that there are no abnormal

data. From the PCA plots and cluster analysis plots, it was known that there was even more variation within groups than between groups, that samples with the NCOR1-LYN gene were also similar to samples without the gene, and that there was no polarisation between the two groups in terms of principal components, proving that the presence of the NCOR1-LYN gene had little effect on the samples. Although the effect was small, subsequent statistical analysis showed that several genes were differentially expressed in this sample, such as 1450009_at, 1421375_a_at, and 1418601_at, especially 1421375_a_at. These genes are associated with lactogen, S100 calcium-binding protein, and the aldehyde dehydrogenase family, respectively.[6] A search of the literature shows that aldehyde dehydrogenases are overexpressed in various types of cancer and that one of the family genes, ALDH1A2, is aberrantly expressed in more than 50% of cases of T-cell acute lymphoblastic leukaemia.[7] Finally, enrichment analysis reveals that the samples have up-regulated apoptotic capacity, up-regulated P53 pathway, and down-regulated protein response. This effect demonstrates that the NCOR1-LYN gene may activate the p53 pathway, promote growth inhibition, and promote apoptosis in leukemia patient cells.

In conclusion, the presence or absence of the NCOR1-LYN gene made no significant difference to gene expression in Ba/F3 cells. However, enrichment analysis revealed that the NCOR1-LYN gene might up-regulate the expression of aldehyde dehydrogenase, activate the p53 pathway, inhibit cell growth and promote apoptosis. It is hypothesized that this pathway is responsible for the association of the NCOR1-LYN gene with leukaemia. And this inference is confirmed by literature: "low expression of ALDH2 reduces the probability of p53 mutation".[8]

Reference

[1] Tomii, Toshihiro, et al. "Leukemic cell expressing a novel kinase fusion protein NCOR1-LYN exhibits high sensitivity to Dasatinib and Rapamycin." Blood 132 (2018): 1557.

[2] Tomii, Toshihiro, et al. "Leukemic cells expressing NCOR1-LYN are sensitive to dasatinib in vivo in a patient-derived xenograft mouse model." Leukemia 35.7 (2021): 2092-2096.

[3] Bi, Weiwei, et al. "Metabonomics analysis of flavonoids in seeds and sprouts of two Chinese soybean cultivars." Scientific Reports 12.1 (2022): 1-13.

[4] Kruger, Claudia, and Claudia Kappen. "Microarray analysis of defective cartilage in Hoxc8-and Hoxd4-transgenic mice." Cartilage 1.3 (2010): 217-232.

[5] Law, Charity W., et al. "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR." F1000Research 5 (2016).

[6] Chujing Zhang, Stella Amanda, et al. "Oncorequisite role of an aldehyde dehydrogenase in the pathogenesis of T-cell acute lymphoblastic leukemia." haematologica 106.6 (2021): 1545.

[7] Huser, C. A., et al. "Insertional Mutagenesis and Deep Profiling Reveals Gene Hierarchies and a Myc/p53." (2014).

[8] Wang, Wei-Lin, et al. "Aldehyde Dehydrogenase 2 Family Member (ALDH2) Is a Therapeutic Index for Oxaliplatin Response on Colorectal Cancer Therapy with Dysfunction p53." BioMed Research International 2022 (2022).