RNA-seq Analysis: RNA expression profiling of CD13/CD33-positive and CD13/CD33-negative B-ALL patients

1. Introduction

A transcriptome refers to the sum of all RNA transcripted by a particular tissue or cell at a certain time or state, including primarily mRNA and non-coding RNA. In a narrow sense, it refers to the sum of all mRNAs. Transcriptome sequencing in this project specifically refers to mRNAs sequencing. Transcriptome research is the basis of studying gene function and structure and plays an important role in the development of organisms and the occurrence of diseases. With the development of gene sequencing technology and the reduction of sequencing cost, RNA-seq has become the main method for transcriptome research due to its advantages of high throughput, high sensitivity and wide application range. In this project, RNA-seq analysis was performed on 4 CD13/CD33-positive and 4 CD13/CD33-negative B-ALL patients. The RNA-Seq data used for this analysis was gotten from NCBI-Gene Expression Omnibus Database. The dataset of interest was identified with the unique identifier (ID) GSE197178. The data was submitted by Liao H et al., 2023.

The data set contain 8 samples, and all were used for the RNA-seq analysis. The 8 samples are:

Samples	Label
GSM5910793	CD13/CD33-negative
GSM5910794	CD13/CD33-negative
GSM5910795	CD13/CD33-negative
GSM5910796	CD13/CD33-negative
GSM5910797	CD13/CD33-positive
GSM5910798	CD13/CD33-positive
GSM5910799	CD13/CD33-positive
GSM5910800	CD13/CD33-positive

2. Difference Gene Statistics

The statistics of the number of difference genes (including up-regulation and down-regulation) for each group and the threshold for screening are shown in the table below.

Group	Total	Down	Up
Negative vs Positive	18	3	15

- Group: Comparison group name.
- Total: The total number of difference genes in the comparison group.
- Down: The down-regulation number of difference genes in the comparison group.
- Up: The up-regulation number of difference genes in the comparison group.

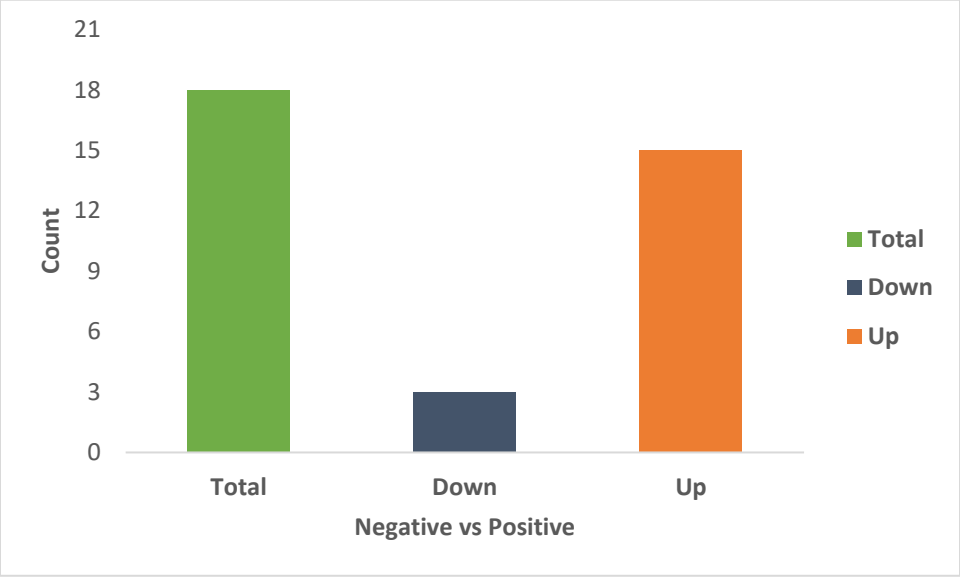


Figure 1.0 Differential Gene Number Statistics Histogram

Note: Orange and Dark blue represent the differential genes for up- and down-regulation, respectively, and the numbers on the columns indicate the number of differential genes.

3. Difference Gene Table

The difference significance analysis is shown in the table below. The table shows all the rows of the difference significance results for all the comparison group.

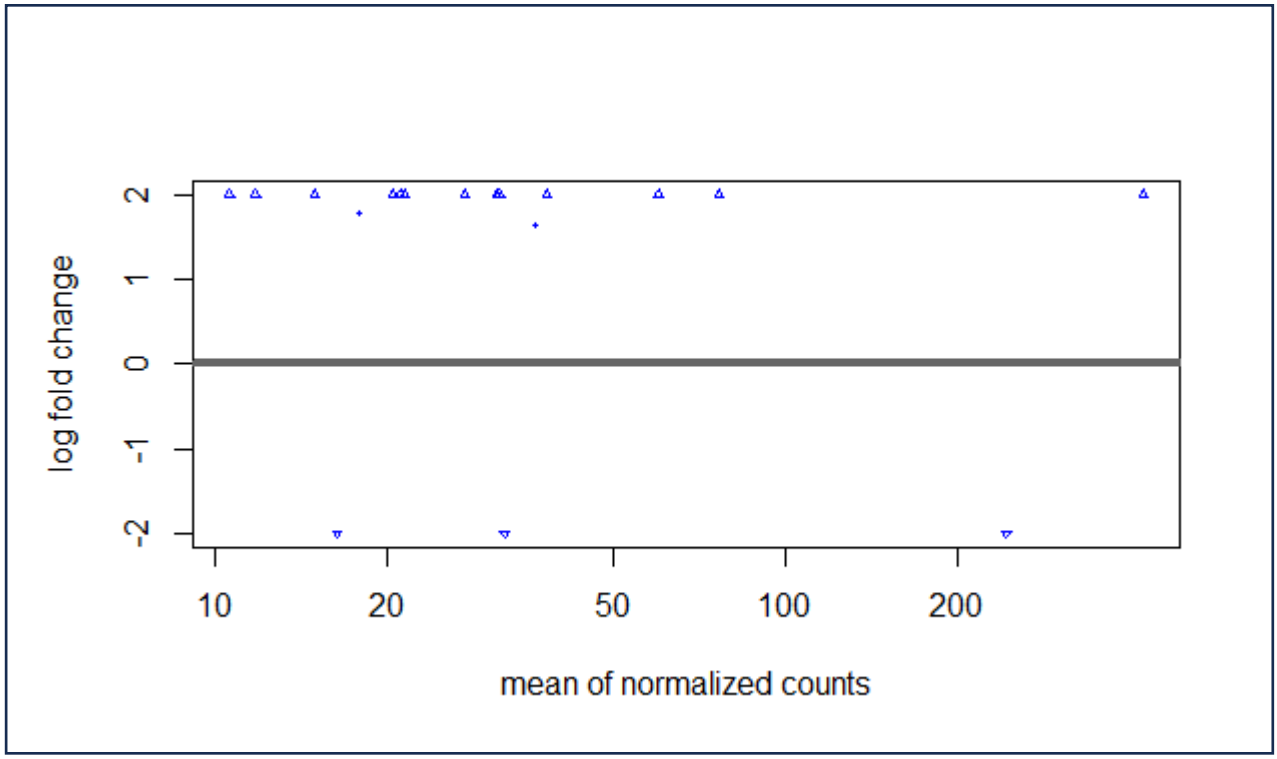
gene_id	baseMean	log2FoldC	lfcSE	stat	pvalue	padj	regulation	gene_nam	gene_locu	ALL_1_1	ALL_1_2	ALL_1_3	ALL_1_4	ALL_2_1	ALL_2_2	ALL_2_3	ALL_2_4
ENSG0000	14.96139	2.945	0.823438	3.576466	0.000348	0.096532	up	SPARC	5:1516610	2.492803	4.040721	3.030834	5.59269	34.66285	25.67475	2.193042	40.90396
ENSG0000	76.5798	2.749073	0.598911	4.590119	4.43E-06	0.006958	up	ID2	2:8678845	16.42624	7.865644	36.16955	21.4052	81.64028	147.2506	70.97458	201.3286
ENSG0000	244.1966	-2.66595	0.695876	-3.83108	0.000128	0.066797	down	PRDX1	1:4551103	991.9314	506.5234	117.6626	161.4653	115.8742	74.69392	26.28819	51.3131
ENSG0000	32.15289	-4.27779	1.203225	-3.55527	0.000378	0.098846	down	NPY	7:2428418	149.7379	1.722699	67.41058	22.36434	1.261032	0	7.055689	2.534833
ENSG0000	11.78459	3.74487	1.019612	3.672838	0.00024	0.086944	up	TNFSF9	19:653102	0.826602	3.174857	1.969373	1.070822	32.9622	13.57528	36.1752	1.460825
ENSG0000	17.94423	1.786371	0.472382	3.781626	0.000156	0.073416	up	FOSB	19:454679	8.355768	7.753748	10.81642	7.311561	40.70284	22.25595	20.15142	24.5567
ENSG0000	16.31839	-4.53275	1.093254	-4.14611	3.38E-05	0.026558	down	IRF4	6:391752-	79.85324	29.7703	12.85976	5.954365	1.356441	0.308335	0.914576	3.441836
ENSG0000	21.53817	6.507048	1.540917	4.222842	2.41E-05	0.022735	up	IFI44L	1:7861992	0.130632	1.106547	0.760114	0.20513	52.82999	111.8618	13.57267	0.054339
ENSG0000	31.60382	2.623278	0.66763	3.929242	8.52E-05	0.050191	up	IFI44	1:7864979	0.929699	12.63641	12.774	11.26521	60.09067	100.7975	35.51768	20.86994
ENSG0000	59.86588	2.087746	0.564974	3.695296	0.00022	0.086242	up	IRF8	16:858991	9.86765	9.726089	24.93314	54.36985	91.42636	89.4529	109.0825	77.2299
ENSG0000	27.49488	6.329109	1.198678	5.280072	1.29E-07	0.000304	up	SCN3A	2:1650875	0.016217	0.008502	0.06449	3.292608	38.69994	51.42978	9.245695	105.9656
ENSG0000	10.61141	6.940628	1.728686	4.014974	5.95E-05	0.04002	up	TDRD9	14:103928	0.275943	0.037109	0.159359	0.036525	34.00635	0.277326	40.06492	4.674281
ENSG0000	20.52651	2.542664	0.685232	3.710661	0.000207	0.086242	up	ATF3	1:2125653	3.66907	7.14927	9.792403	3.930758	27.15562	17.61001	62.74054	18.61558
ENSG0000	424.645	2.138592	0.298818	7.156828	8.26E-13	3.89E-09	up	FOS	14:752788	160.1503	170.2272	152.6898	188.1984	791.9412	455.9698	605.883	765.7665
ENSG0000	36.34955	1.654601	0.457623	3.615645	0.0003	0.094115	up	PTTG1IP	21:448495	11.71931	35.30859	12.22411	18.06002	58.32588	66.36377	60.17768	28.62244
ENSG0000	31.26158	3.219586	0.896989	3.589328	0.000332	0.096532	up	MAFF	22:382007	9.616405	2.449826	10.8792	1.239964	71.67829	22.37569	13.24021	107.127
ENSG0000	38.17	2.943412	0.682235	4.314368	1.60E-05	0.018855	up	CEBPA	19:332999	4.636776	22.0272	1.339128	12.00552	126.1674	60.79933	28.63078	56.47732
ENSG0000	21.11634	4.880602	1.33584	3.653582	0.000259	0.08704	up	SELENOP	5:4279988	0.372856	0.990723	1.686523	3.122809	0.777952	147.8849	5.259474	14.02884

- **gene_id**: Gene number
- **baseMean**: The average expression level of the gene across all negative and positive samples.
- **log2FoldChange**: The value is a ratio of gene expression levels between the negative group and the positive group, and then take the logarithm of 2.
- **lfcSE**: The standard error of the log2 fold change estimate.
- **stat**: The test statistic is used to assess the significance of the log2 fold change.
- **pvalue**: The p-value associated with the test statistic.
- **padj**: The adjusted p-value, which accounts for multiple testing.
- **gene_name**: The name of the gene
- **gene_locus**: The genomic locus or location of the gene.

- **ALL_1_1** to **ALL_2_4**: Expression values for the gene across negative and positive samples.

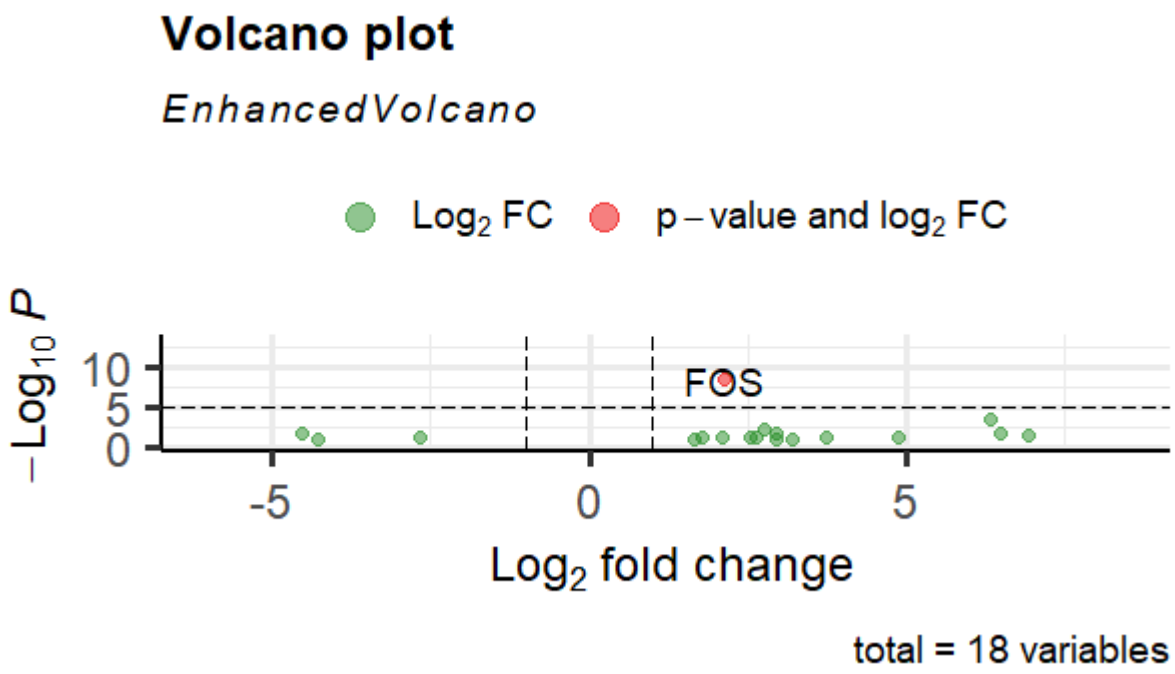
4. MA plots

The MA plot can visually show the overall distribution of gene expression levels and differential multiples, as shown in the figure below.



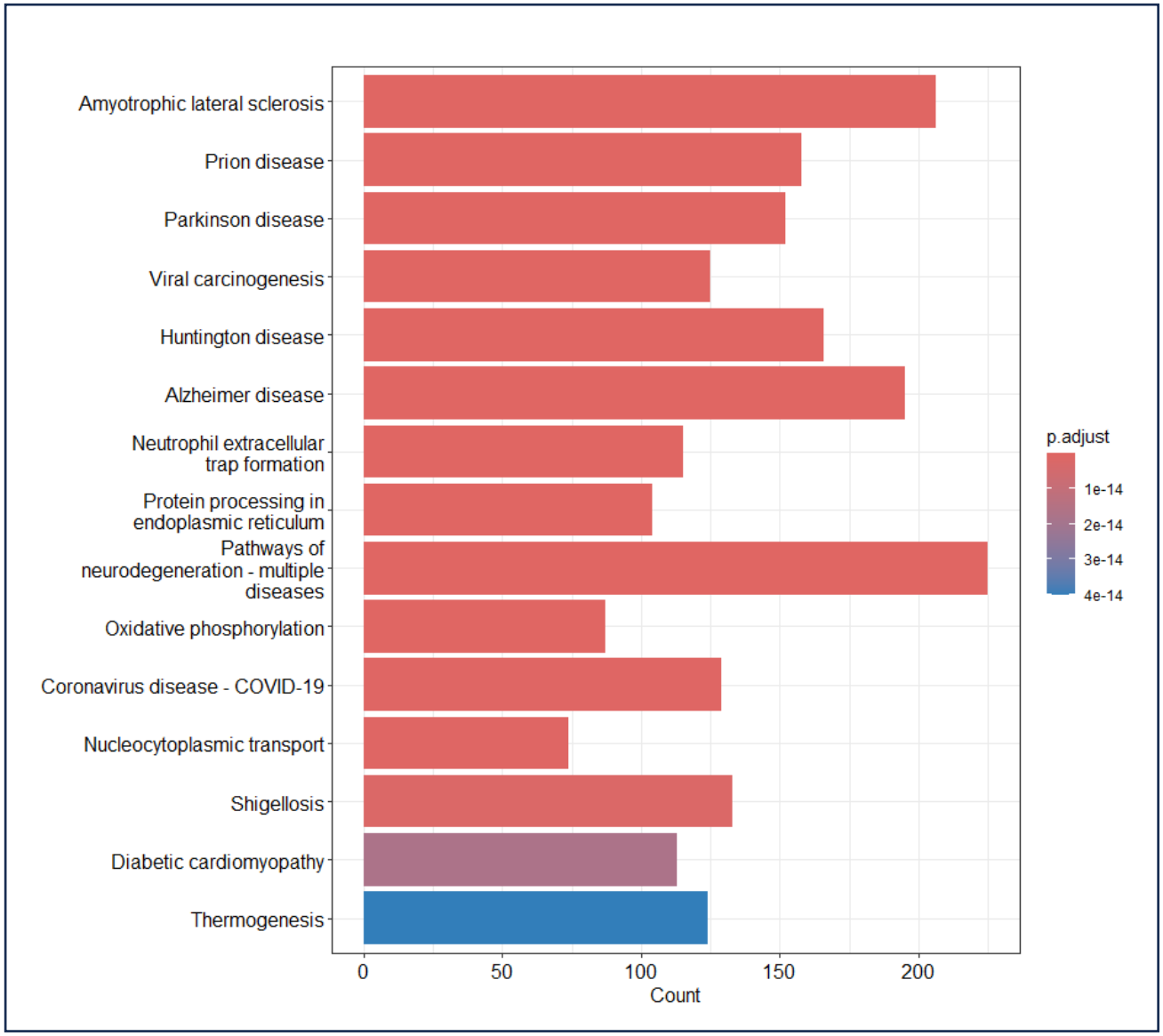
5. Volcano plots

Volcano plots can be used to infer the overall distribution of differentially expressed genes. In the figure, the x-axis shows the fold change in gene expression between different samples, and the y-axis shows the statistical significance of the differences.



6. KEGG Annotation Analysis of DEGs

After the genes were annotated into the KEGG database, the number of differential genes contained in each KEGG pathway was counted and a bar chart was drawn, as shown in the figure below.



7. KEGG Enrichment Analysis of DEGs

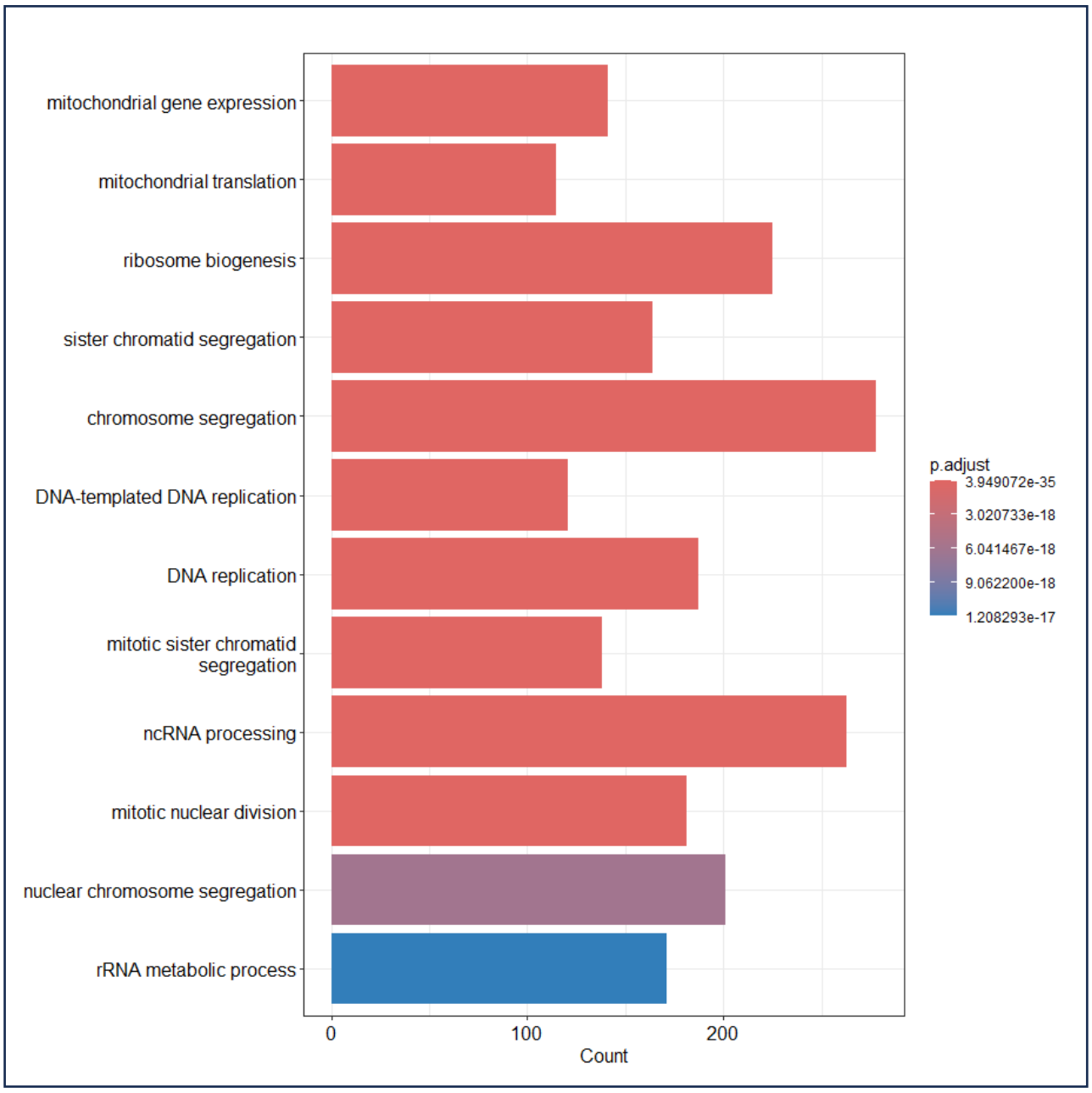
Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed genes compared with the whole genome background. Here, the KEGG pathway analysis used p.adj less than 0.1 as the threshold for significant enrichment.

	category	subcategory	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
hsa05014	Human Diseases	Neurodegenerative disease	hsa05014	Amyotrophic lateral sclerosis	206/2534	364/8764	1.85E-29	6.24E-27	3.67E-27	572/4706/	206
hsa05020	Human Diseases	Neurodegenerative disease	hsa05020	Prion disease	158/2534	272/8764	2.18E-24	3.68E-22	2.16E-22	572/4706/	158
hsa05012	Human Diseases	Neurodegenerative disease	hsa05012	Parkinson disease	152/2534	266/8764	1.70E-22	1.90E-20	1.12E-20	4706/292/	152
hsa05203	Human Diseases	Cancer: overview	hsa05203	Viral carcinogenesis	125/2534	204/8764	2.60E-22	2.19E-20	1.28E-20	572/8379/	125
hsa05016	Human Diseases	Neurodegenerative disease	hsa05016	Huntington disease	166/2534	306/8764	3.27E-21	2.21E-19	1.30E-19	4706/292/	166
hsa05010	Human Diseases	Neurodegenerative disease	hsa05010	Alzheimer disease	195/2534	384/8764	2.56E-20	1.44E-18	8.44E-19	572/4706/	195
hsa04613	Organismal System	Immune system	hsa04613	Neutrophil extracellular trap	115/2534	191/8764	9.59E-20	4.62E-18	2.71E-18	292/4353/	115
hsa04141	Genetic Informatio	Folding, sorting and degradat	hsa04141	Protein processing in endopl	104/2534	170/8764	1.10E-18	4.65E-17	2.73E-17	7095/823/	104
hsa05022	Human Diseases	Neurodegenerative disease	hsa05022	Pathways of neurodegenerat	225/2534	476/8764	2.34E-18	8.77E-17	5.15E-17	572/4706/	225
hsa00190	Metabolism	Energy metabolism	hsa00190	Oxidative phosphorylation	87/2534	134/8764	3.43E-18	1.15E-16	6.78E-17	4706/7384	87
hsa05171	Human Diseases	Infectious disease: viral	hsa05171	Coronavirus disease - COVID-	129/2534	233/8764	1.11E-17	3.38E-16	1.99E-16	6224/5112	129
hsa03013	Genetic Informatio	Translation	hsa03013	Nucleocytoplasmic transport	74/2534	108/8764	1.20E-17	3.38E-16	1.99E-16	5976/2327	74
hsa05131	Human Diseases	Infectious disease: bacterial	hsa05131	Shigellosis	133/2534	247/8764	7.67E-17	1.99E-15	1.17E-15	823/960/7	133
hsa05415	Human Diseases	Cardiovascular disease	hsa05415	Diabetic cardiomyopathy	113/2534	203/8764	7.25E-16	1.75E-14	1.03E-14	4706/292/	113
hsa04714	Organismal System	Environmental adaptation	hsa04714	Thermogenesis	124/2534	232/8764	1.88E-15	4.03E-14	2.36E-14	23028/470	124
hsa05132	Human Diseases	Infectious disease: bacterial	hsa05132	Salmonella infection	130/2534	247/8764	1.91E-15	4.03E-14	2.36E-14	4074/5606	130
hsa05208	Human Diseases	Cancer: overview	hsa05208	Chemical carcinogenesis - rea	119/2534	223/8764	8.21E-15	1.63E-13	9.56E-14	572/4706/	119
hsa03010	Genetic Informatio	Translation	hsa03010	Ribosome	97/2534	170/8764	1.02E-14	1.91E-13	1.12E-13	6224/5112	97
hsa04932	Human Diseases	Endocrine and metabolic dise	hsa04932	Non-alcoholic fatty liver dise	90/2534	155/8764	2.26E-14	4.02E-13	2.36E-13	4706/7384	90

- **Category**: Broader category of the pathway.
- **Subcategory**: A more specific categorization within the broader category
- **ID**: The unique identifier for the pathway.
- **Description**: A descriptive label for the pathway.
- **GeneRatio**: Ratio of genes associated with the pathway.
- **BgRatio**: Ratio of genes from the background set that are associated with the pathway.
- **P-value**: p in hypergenometric test.
- **p.adjust**: Corrected p
- **qvalue**: Another adjusted p-value.
- **geneID**: IDs or symbols of the genes involved in the pathway.
- **Count**: The number of genes associated with the pathway.

8. GO Annotation Analysis of DEGs

GO (Gene Ontology) is a comprehensive database describing the function of genes. The GO classification statistical results of differentially expressed genes are shown in the figure below:



The GO function enrichment uses padj less than 0.05 as the threshold for significant enrichment.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0140053	GO:0140053	mitochondrial gene expression	141/6890	176/21261	5.95E-39	3.95E-35	3.37E-35	ENSG0000017	141
GO:0032543	GO:0032543	mitochondrial translation	115/6890	141/21261	2.15E-33	7.13E-30	6.08E-30	ENSG0000017	115
GO:0042254	GO:0042254	ribosome biogenesis	225/6890	365/21261	7.27E-31	1.61E-27	1.37E-27	ENSG0000018	225
GO:0000819	GO:0000819	sister chromatid segregation	164/6890	244/21261	4.87E-29	8.09E-26	6.90E-26	ENSG0000011	164
GO:0007059	GO:0007059	chromosome segregation	278/6890	493/21261	1.35E-28	1.79E-25	1.53E-25	ENSG0000011	278
GO:0006261	GO:0006261	DNA-templated DNA replication	121/6890	163/21261	5.62E-28	6.22E-25	5.30E-25	ENSG0000017	121
GO:0006260	GO:0006260	DNA replication	187/6890	298/21261	3.13E-27	2.97E-24	2.53E-24	ENSG0000017	187
GO:0000070	GO:0000070	mitotic sister chromatid segregation	138/6890	201/21261	4.69E-26	3.89E-23	3.32E-23	ENSG0000011	138
GO:0034470	GO:0034470	ncRNA processing	263/6890	490/21261	7.63E-23	5.11E-20	4.36E-20	ENSG0000012	263
GO:0140014	GO:0140014	mitotic nuclear division	181/6890	303/21261	7.71E-23	5.11E-20	4.36E-20	ENSG0000017	181

- **ID**: GO term ID.
- **Description**: Description of the biological process, cellular component, or molecular function.
- **GeneRatio**: The ratio of the number of difference genes annotated to GO number to the total number of difference genes.
- **BgRatio**: The ratio of the number of background genes annotated to GO number to the total number of background genes.
- **pvalue**: p in hypergenometric test.
- **p.adjust**: corrected p.
- **qvalue**: Another measure of statistical significance adjusted for multiple testing.
- **geneID**: Difference gene number annotated to GO number.

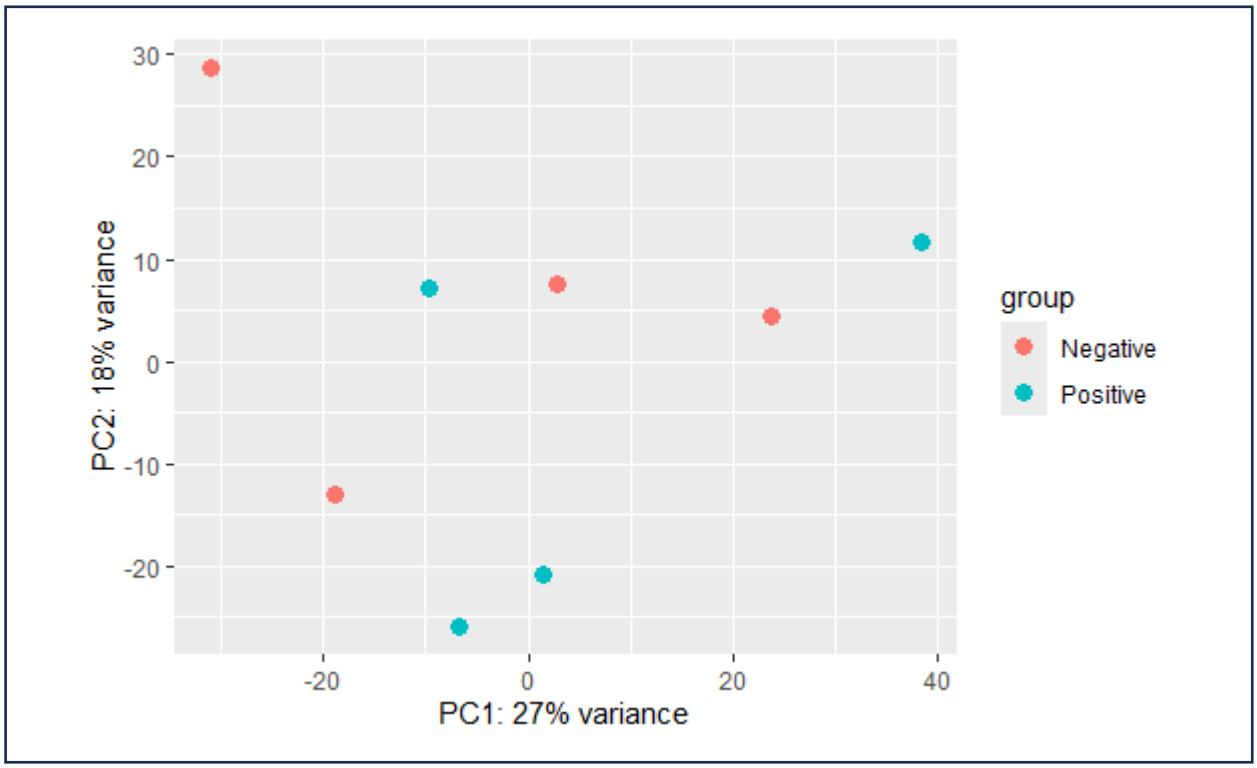
9. GSEA Enrichment Analysis

The GSEA consists of there main steps: calculating the (Enrichment Score); estimating the significance level of the enrichment score, and multiple hypothesis testing.

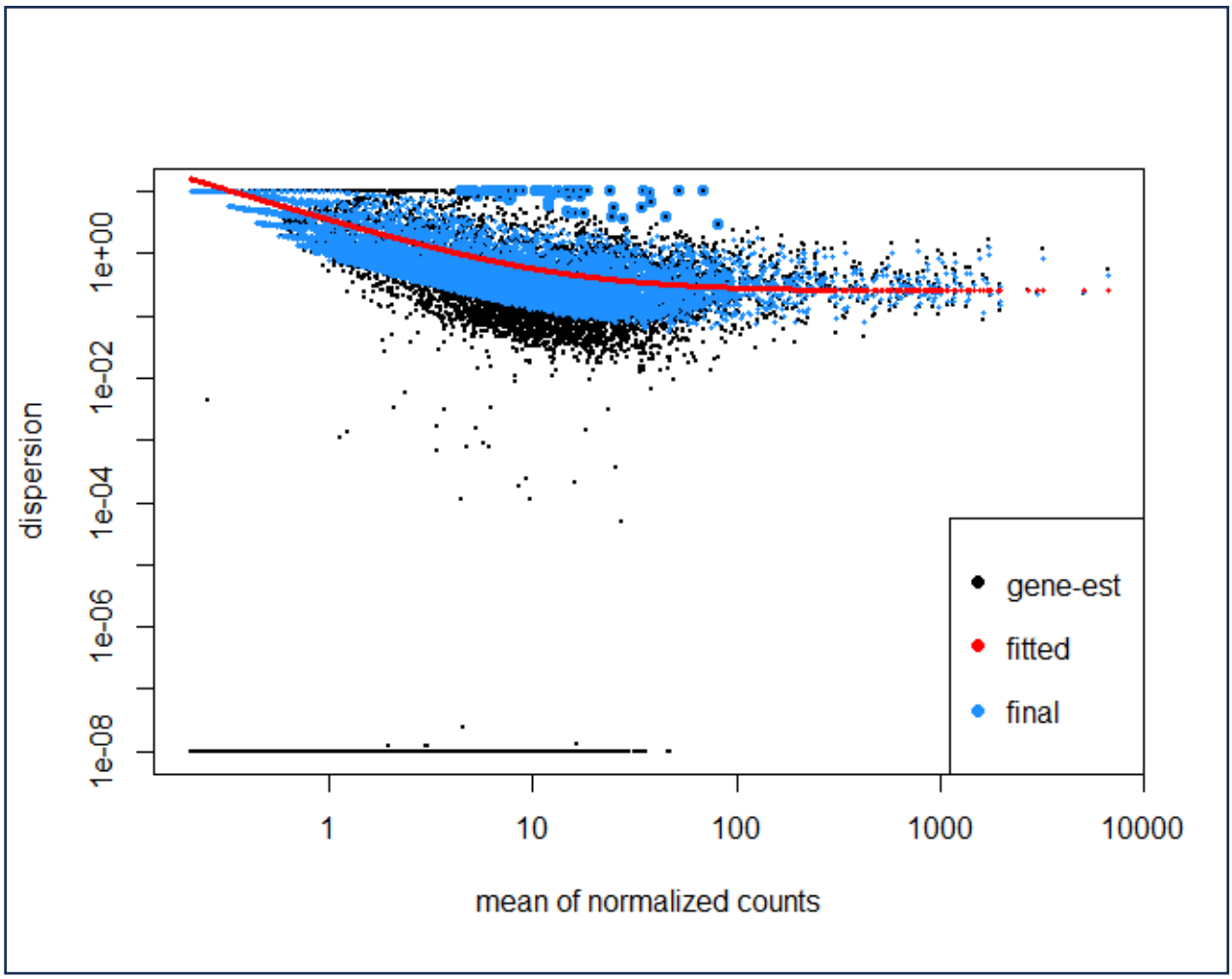
	ID	Description	setSize	enrichme	NES	pvalue	p.adjust	qvalue	rank	leading_edge	core_enrichment
GO:00192	GO:00192	cytokine-mediated signaling	366	0.456185	1.826438	9.00E-09	1.64E-05	1.41E-05	2521	tags=36%, list=19%, signal	ENSG00000263961/E
GO:00096	GO:00096	response to bacterium	459	0.433758	1.76899	8.34E-09	1.64E-05	1.41E-05	2136	tags=33%, list=16%, signal	ENSG00000138755/E
GO:00022	GO:00022	immune effector process	496	0.43238	1.768965	3.14E-09	1.64E-05	1.41E-05	2167	tags=30%, list=16%, signal	ENSG00000256660/E
GO:00516	GO:00516	defense response to virus	278	0.479641	1.869651	4.02E-08	3.14E-05	2.69E-05	2542	tags=33%, list=19%, signal	ENSG00000137959/E
GO:01405	GO:01405	defense response to symbior	278	0.479641	1.869651	4.02E-08	3.14E-05	2.69E-05	2542	tags=33%, list=19%, signal	ENSG00000137959/E
GO:00313	GO:00313	positive regulation of defens	364	0.450843	1.803842	3.91E-08	3.14E-05	2.69E-05	2580	tags=33%, list=19%, signal	ENSG00000145623/E
GO:00321	GO:00321	positive regulation of respon	467	0.420667	1.717461	3.74E-08	3.14E-05	2.69E-05	2183	tags=30%, list=16%, signal	ENSG00000145623/E
GO:00096	GO:00096	response to virus	358	0.439153	1.749544	1.27E-07	8.65E-05	7.43E-05	2596	tags=32%, list=19%, signal	ENSG00000137959/E
GO:19021	GO:19021	regulation of leukocyte differ	260	0.476331	1.851814	1.76E-07	0.000107	9.15E-05	2178	tags=33%, list=16%, signal	ENSG00000109906/E
GO:00026	GO:00026	regulation of immune effecto	301	0.457772	1.801909	3.85E-07	0.000211	0.000181	2167	tags=32%, list=16%, signal	ENSG00000256660/E

- **ID**: GO term ID.
- **Description**: Description of the GO term.
- **setSize**: The total number of genes in the gene set.
- **enrichmentScore (ES)**: A score that reflects the degree to which the genes in the gene set are overrepresented at the top or bottom of the ranked list of genes.
- **NES**: Normalized enrichment score.
- **pvalue, p.adjust, qvalue**: Statistical values indicating the significance of the enrichment.
- **rank**: The position of the gene set in the ranked list of gene sets based on their enrichment scores.
- **leading_edge**: The subset of genes within the gene set that contribute most to the enrichment signal.
- **core_enrichment**: Additional information about the contribution of different components within the gene set to the enrichment signal.

10. Principal component analysis



11. Gene dispersion and normalized mean counts



12. Heatmap of top 18 genes