RNA-seq Analysis Using PyDESeq

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, I intend to use **PyDESeq for analyse RNA-seq**. The dataset of interest was identified with the unique identifier (ID) GSE197178. The data was submitted by Kumar SA et al., 2021.

The data set contain 12 samples, and all were used for the RNA-seq analysis. The 12 samples are:

Samples	Label
GSM5720043	Control_16hpi_rep1
GSM5720044	Control_16hpi_rep2
GSM5720045	Control_16hpi_rep3
GSM5720046	△AP2-MRPKO_16hpi_rep1
GSM5720047	△AP2-MRPKO_16hpi_rep2
GSM5720048	△AP2-MRPKO_16hpi_rep3
GSM5720049	Control_40hpi_rep1
GSM5720050	Control_40hpi_rep2
GSM5720051	Control_40hpi_rep3
GSM5720052	△AP2-MRPKO_40hpi_rep1
GSM5720053	△AP2-MRPKO_40hpi_rep2
GSM5720054	△AP2-MRPKO_40hpi_rep3

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
Control vs Test	331	184	147

- 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.

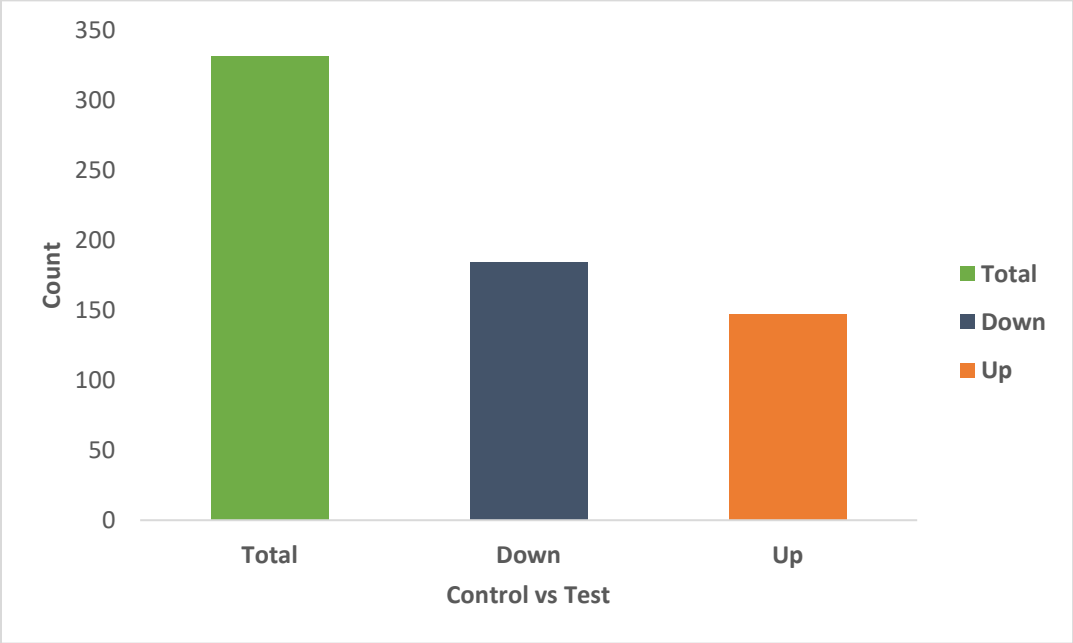


Figure 1.0 Differential Gene Number Statistics Histogram

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.

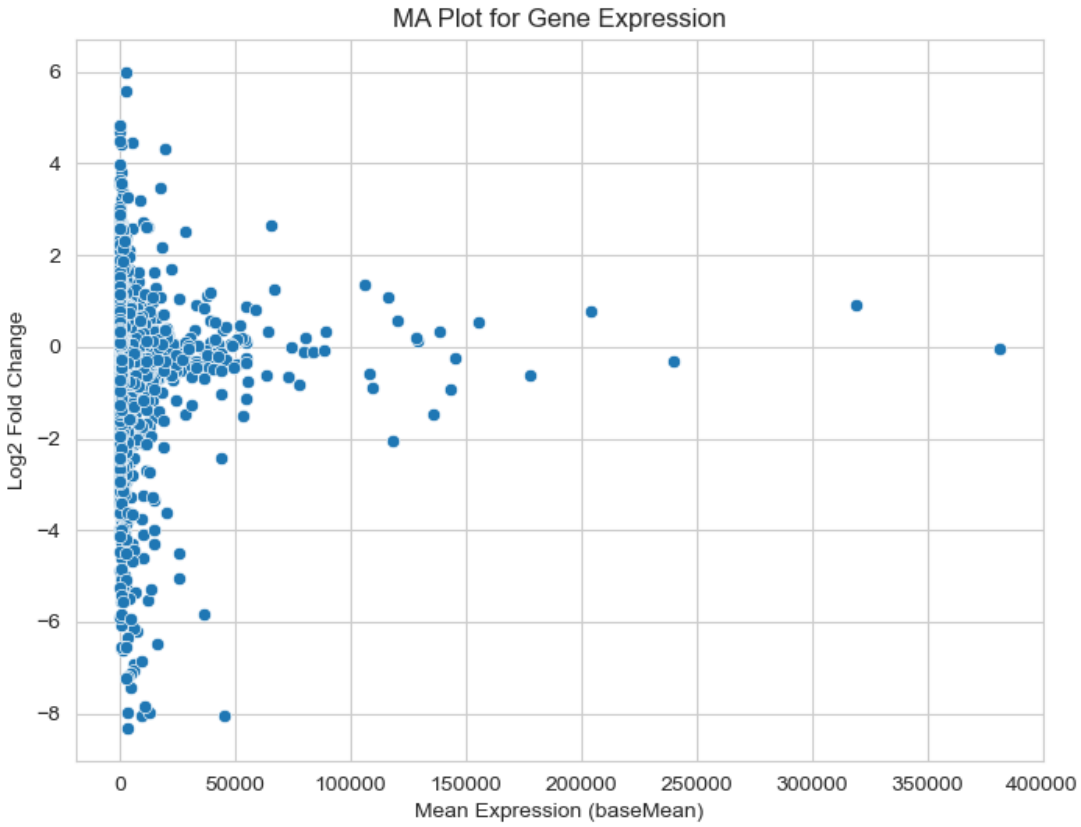
baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	regulation	Gene ID
2356.856	2.37428	0.174417	13.61269	3.37E-42	1.81E-38	up	PF3D7_0114000
2740.704	-1.35192	0.104872	-12.8912	5.05E-38	1.36E-34	down	PF3D7_1442400
517.07	2.731742	0.23208	11.77067	5.53E-32	9.91E-29	up	PF3D7_1016500
146.5257	1.861662	0.160982	11.56444	6.24E-31	8.39E-28	up	PF3D7_1126800
1161.414	2.541198	0.224264	11.3313	9.18E-30	9.88E-27	up	PF3D7_1134600
1304.435	2.525837	0.22551	11.20056	4.05E-29	3.63E-26	up	PF3D7_1038800
4247.473	-1.91597	0.177021	-10.8234	2.67E-27	2.05E-24	down	PF3D7_1131800
468.8212	1.307628	0.123312	10.60424	2.85E-26	1.91E-23	up	PF3D7_0804700
80.71419	2.974157	0.291866	10.19013	2.19E-24	1.31E-21	up	PF3D7_1201300
208.2032	2.790608	0.297628	9.37615	6.84E-21	3.68E-18	up	PF3D7_1038700
4175.516	2.118374	0.226218	9.364303	7.66E-21	3.74E-18	up	PF3D7_1467600
17.40865	-5.25377	0.582304	-9.02239	1.84E-19	8.25E-17	down	PF3D7_1036400
8774.445	3.182595	0.355846	8.943746	3.76E-19	1.56E-16	up	PF3D7_0801900
45238.54	-8.03646	0.932348	-8.61959	6.72E-18	2.58E-15	down	PF3D7_0731500
6370.883	1.239937	0.145848	8.50156	1.87E-17	6.71E-15	up	PF3D7_1142100
1852.846	0.951185	0.114669	8.295058	1.09E-16	3.65E-14	up	PF3D7_1453900
9254.761	-8.05796	1.016401	-7.92794	2.23E-15	7.05E-13	down	PF3D7_1125800
10393.58	1.149368	0.147246	7.805746	5.92E-15	1.77E-12	up	PF3D7_0617800
12638.4	-7.9723	1.044125	-7.63539	2.25E-14	6.37E-12	down	PF3D7_0102500
10395.66	-7.82861	1.03406	-7.57075	3.71E-14	9.98E-12	down	PF3D7_1301600

- **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.

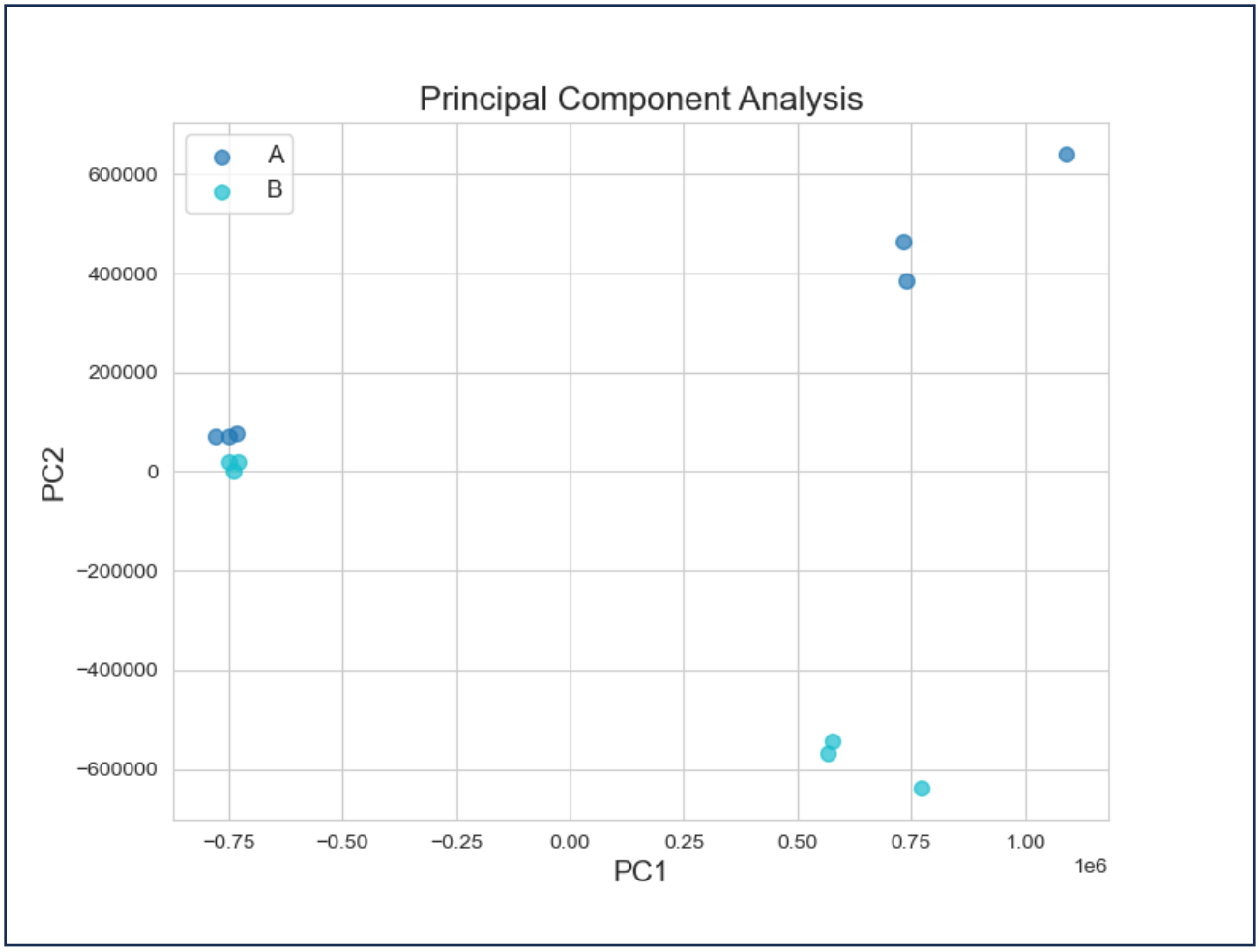
- **padj**: The adjusted p-value, which accounts for multiple testing.
- **regulation**: up-regulated or down-regulated in condition.
- **Gene ID**: Gene number

4. MA plots

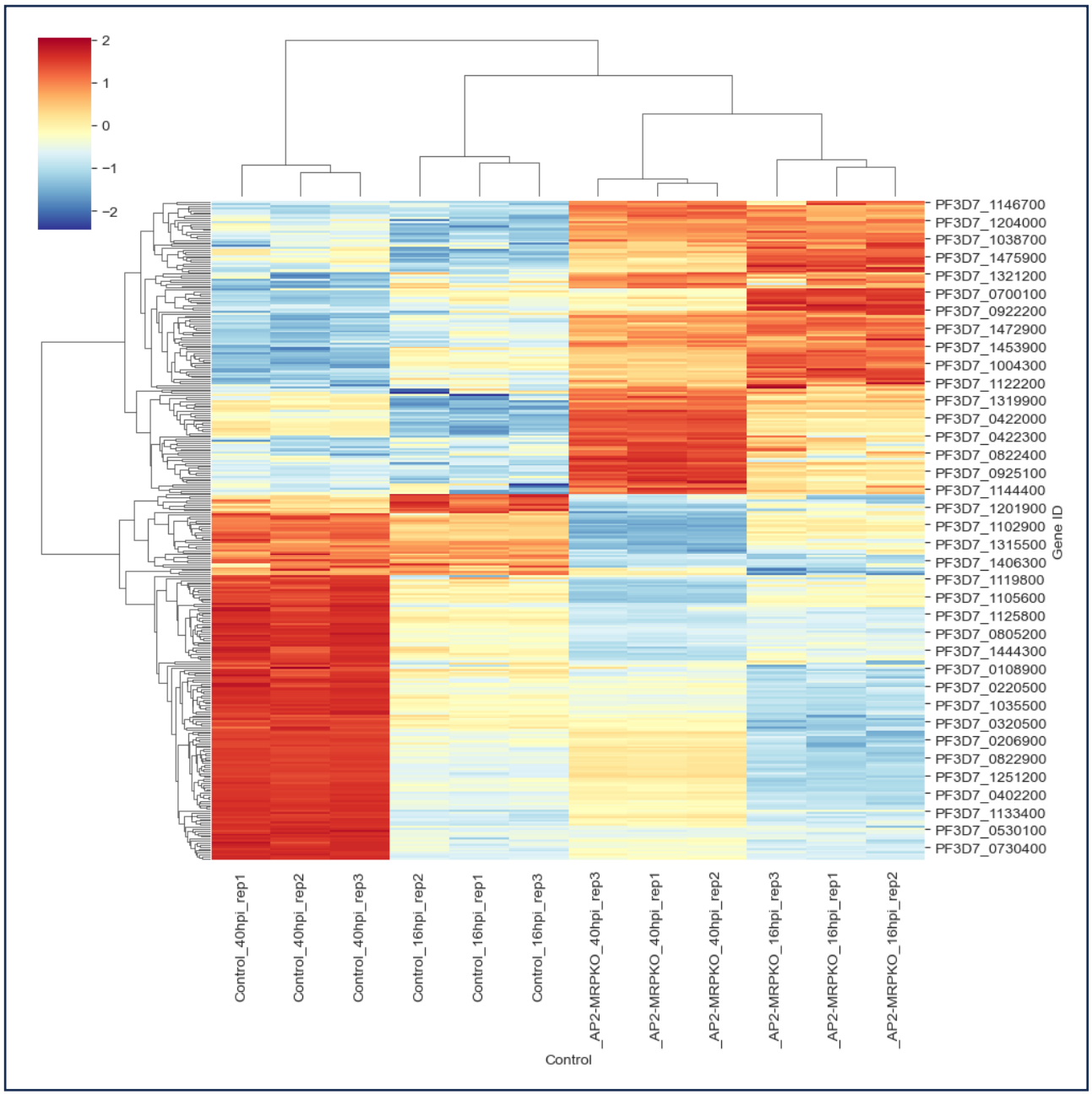
The MA plot can visually show the overall distribution of gene expression levels and differential multiples, as shown in the figure below.



5. Principal component analysis



6. Heatmap of top 37 genes



7. 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.