Introduction to Bioinformatic

Dr.Sharifi - Dr.Koohi

Project Report

Kahbod Aeini
98101209

Purpose of the project is to find genes with the most influence on Leukemia (Adult Acute Myeloid Leukemia) infection. By analysing the given dataset of Microarrays (Series GSE48558 / Platform GPL6244) we are able to pin-point genes with the highest probability of being the cause od the infection.

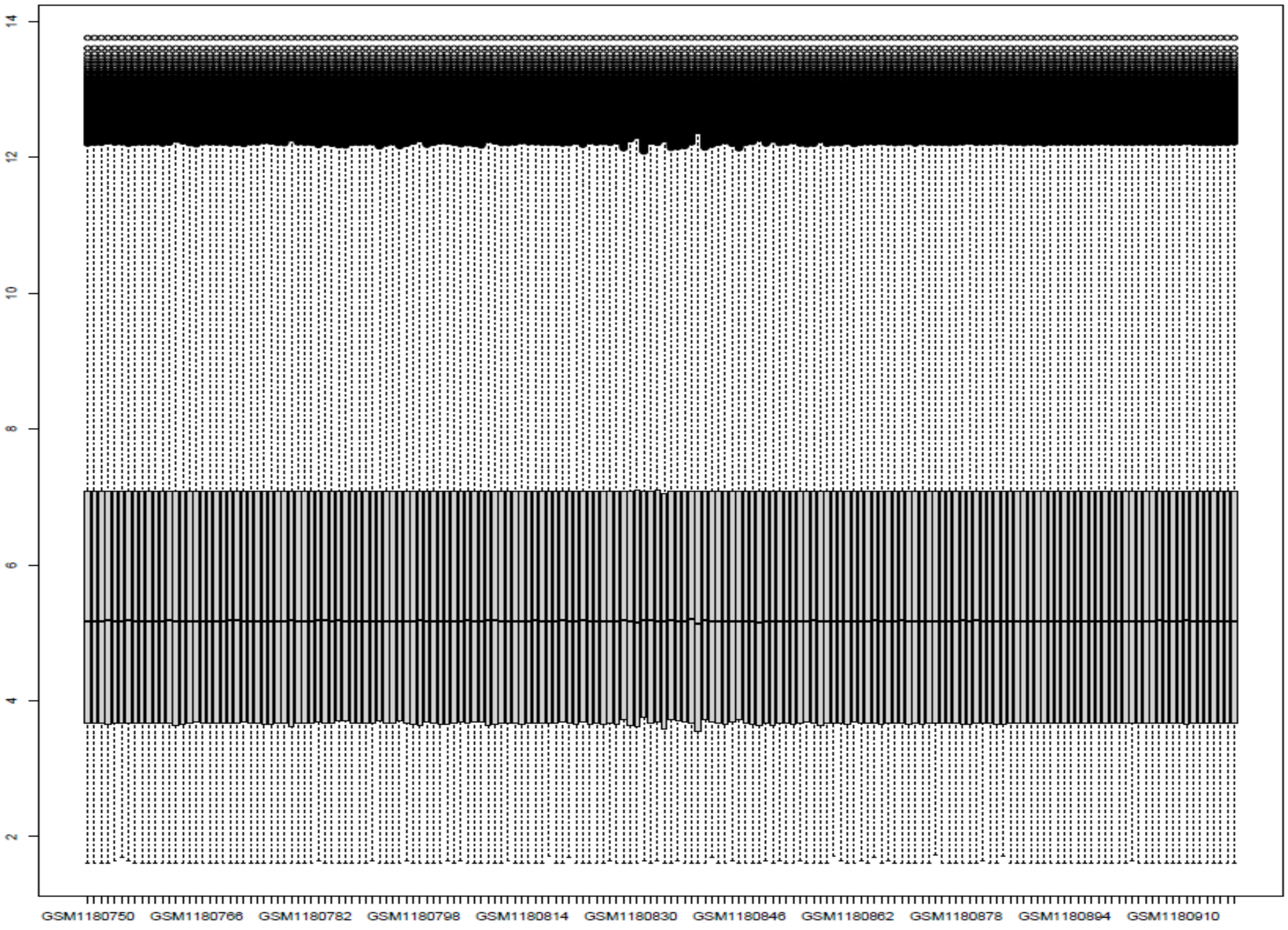
The main approach is to find mentioned Genes based on Differential Expression Analysis. The Approach and a vision to it is explained step by step and given as follows. Note that the comments in the code may also give you an idea of what is described in the report.

Data Fetching And Source Name Grouping

The data is fetched by a built-in function in GEOquery library and generates a set which will be behaved with later. Then we define an array of Source Names in the default order of the dataset. (Line 12-26 of the code)

Expression Matrix

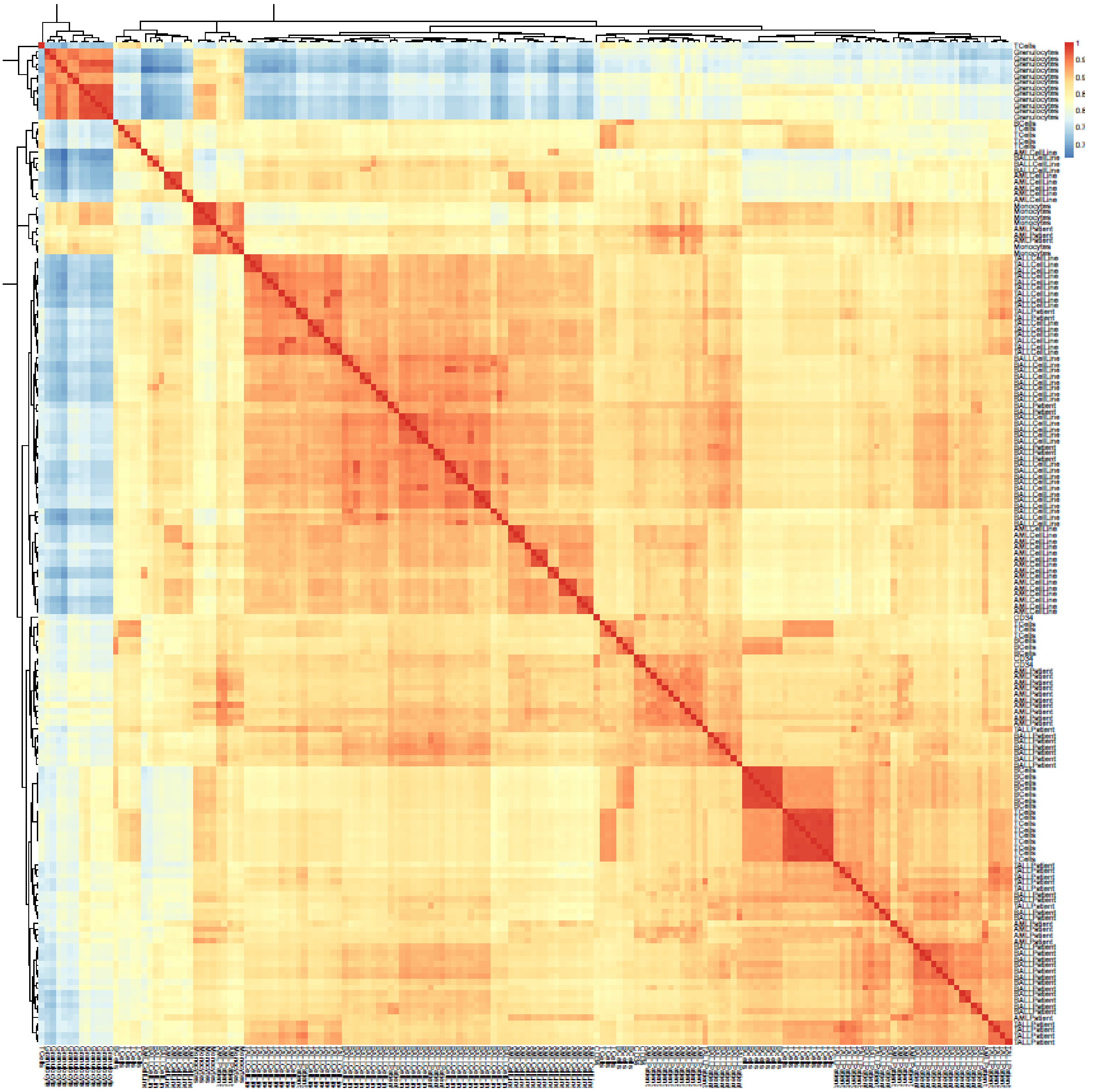
Expression Matrix of the samples can be generated by a built-in function in GEOquery library. The Matrix is plotted in the boxplot.pdf and also shown below.



In the upper table we can see the difference of the samples by the quarter-based table. This difference is determinant for the rest of the algorithm. (Line 28-37 of the code)

Correlation Heatmap

In this section we generate a Correlation Heat Map in order to visualise the correlation and proximation between samples. The more the common square between to genes is near to Red, the more Correlation they have and less unlikely to be the desired gene. The hierarchical Clustering has also been applied to the dataset and the order of the samples in the map is based on the clustering showed on the border.

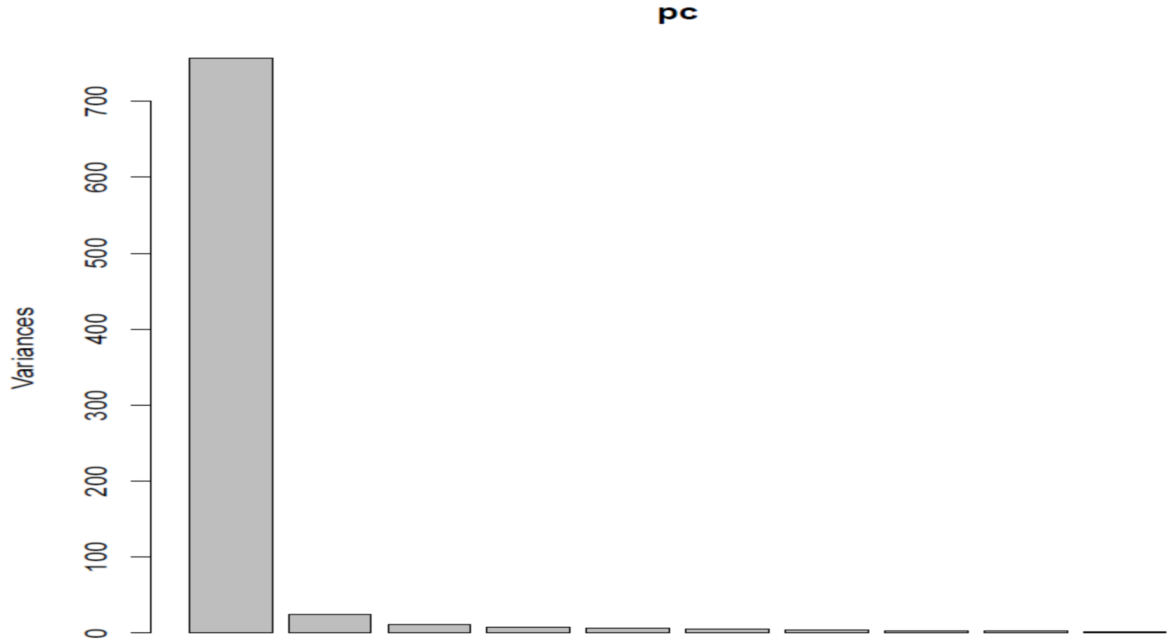


High correlation between genes indicates that they are nearly in every sample and lead us to believe that probably they are House Keeping genes in cells and are not the cause of the cancer. Yet, low correlation demonstrates rarity and impact of the gene on the cell, which probably is the cause of the infection of the cancer. Hence according to the Heat Map, Granulocytes and T Cells are more likely to be the main cause of difference and may be the Leukemia. (Lines 39-42 of the code)

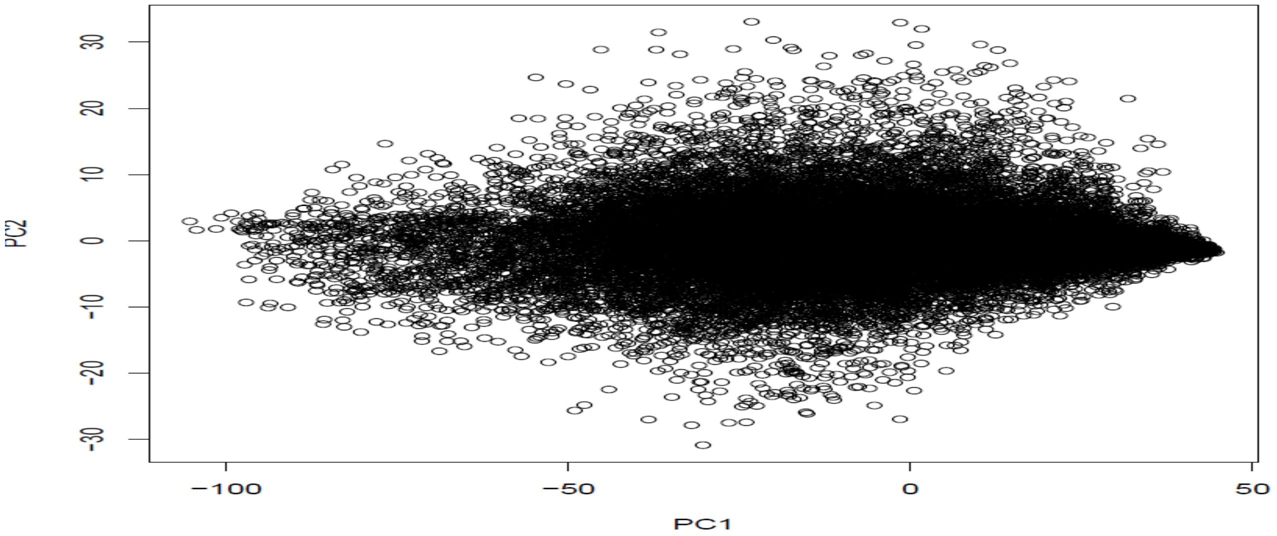
Principal Component Analysis

Now it’s time for Dimension Reduction. We use PCA algorithm to find and fetch the first two components which effect the most on the genes so we can plot it on a 2-D plane.

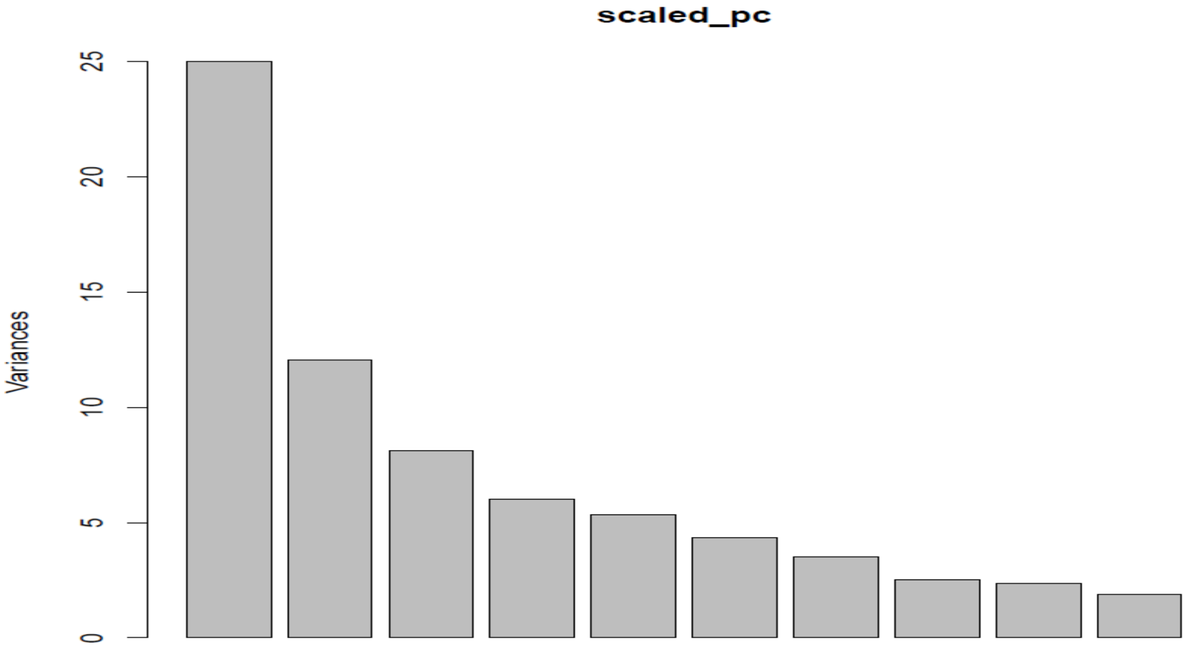
After applying PCA in the dataset we come to this Bar Chart.

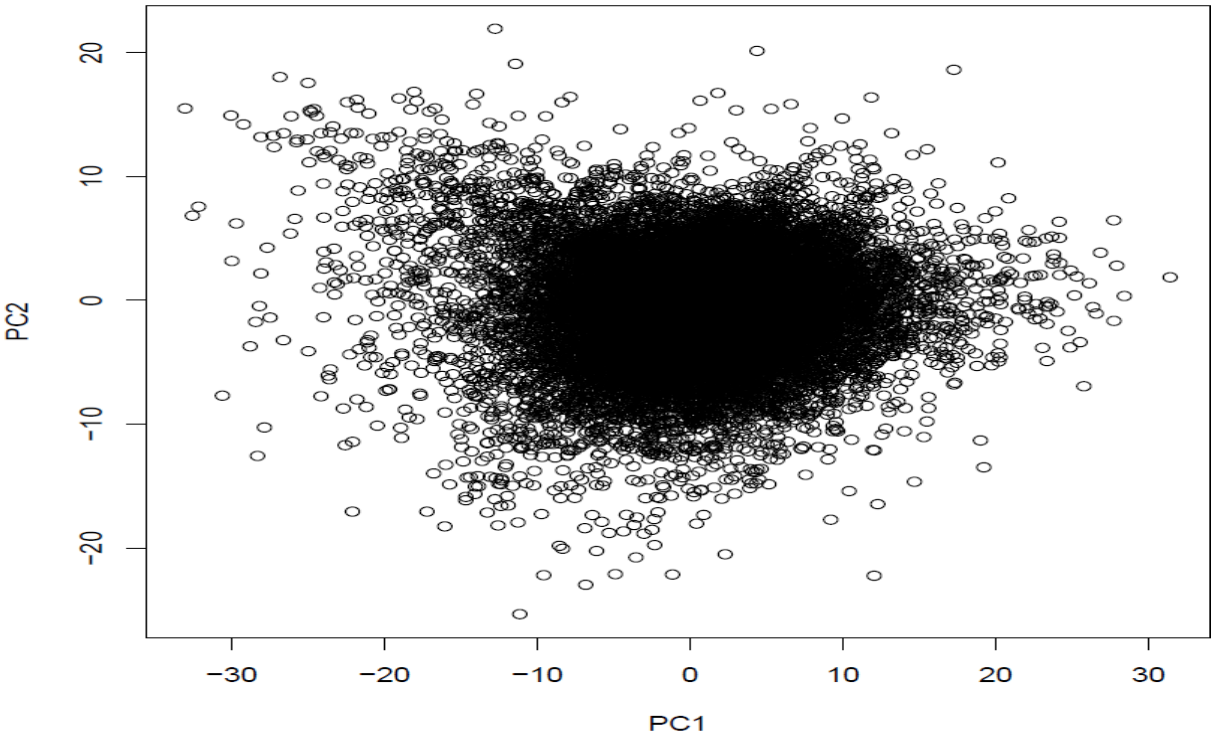


We also can plot samples on a 2-D plane after applying PCA, to visualise the segmentation this analysis represent.

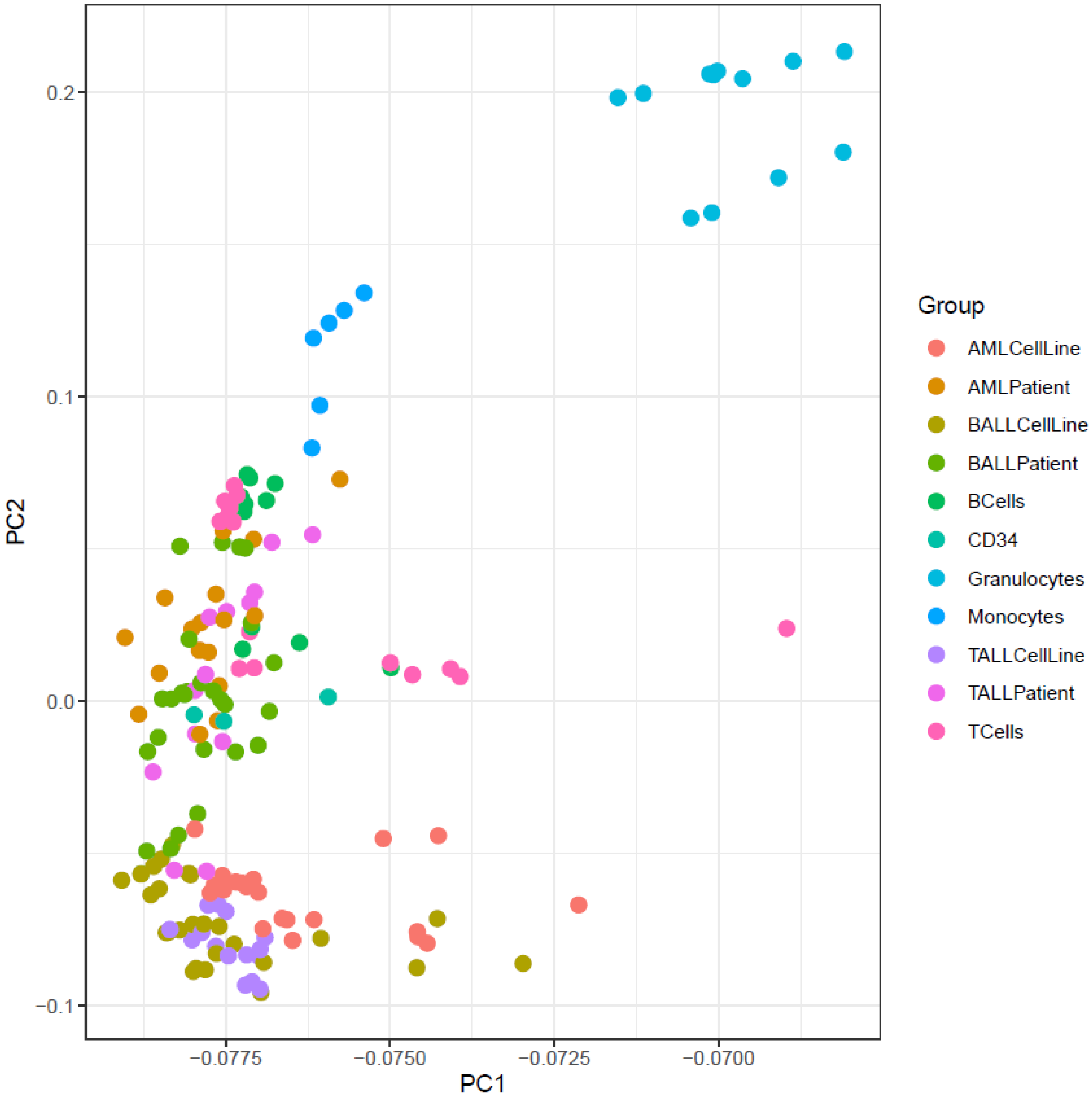


Clearly according to the two recent charts, this PCA does not segmentation well enough. First Principal Component can partition the data relatively good, but the second Principal Component does not partition it well. So we need to standardise the Data and apply PCA on it again. Below we can see Scaled PCA bar chart and samples plot on 2-D plane.





Now after dimension reduction we are able to plot the samples on a 2-D plane.



As we deduced from the Correlation Heat Map in previous section, Granulocytes genes have the most difference and least correlation with other genes and we can see it in the PCA Samples. There is a rare clustering by the PCA through the samples which is distinctive in the recent plot.

The dataset is not much clean and we may expected a better clustering or more distance between samples of different source names, but since we implemented exactly what was instructed in the course, we assume that this is the best behaviour and clustering through the data and we done the experiment well, so we shall continue the algorithm. (This section refers to the lines 44-65 of the code)

Differential Expression Analysis

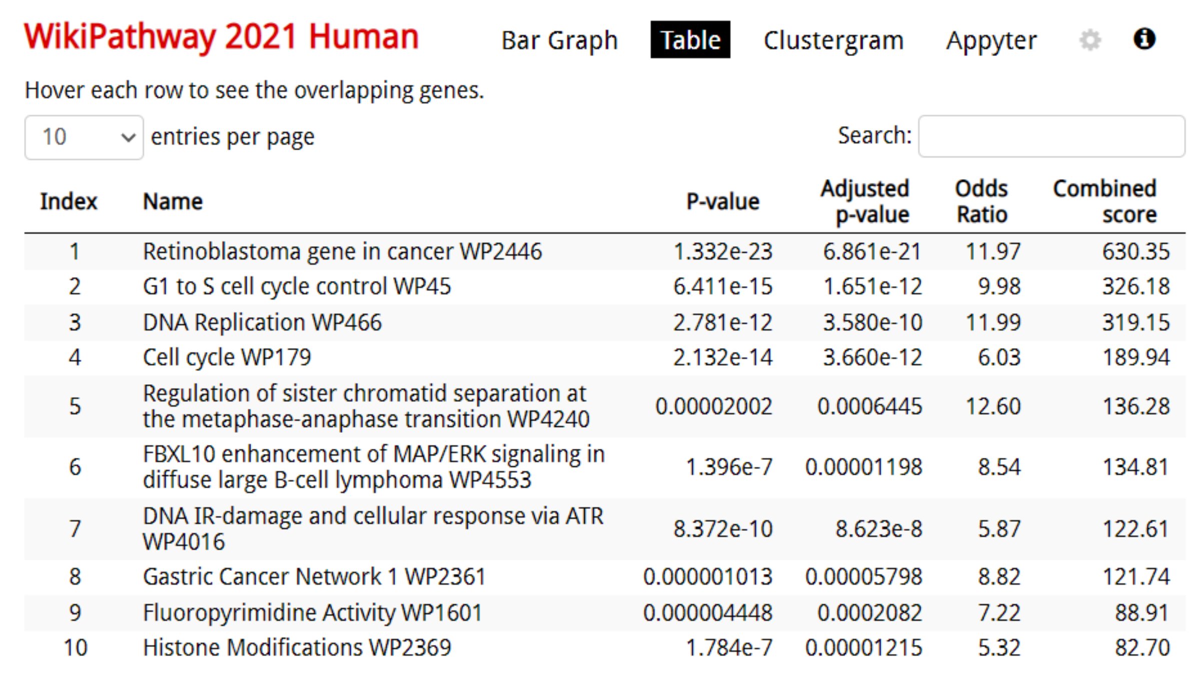
Here we come to the main part of the approach where the difference between genes are about to be analysed. Samples need to be separated to 3 groups, NORMAL, AML PATIENT and UNKNOWN as instructed in the project doc. Then we calculate cause rate of every gene using built-in functions. Discussing more detailed, we try to fit a linear model based on levels of the source name, on the matrix, using a bayesian network. Then we have the table of Symbols, ID, Adjusted P Value and logFC in the AML_Normal text file which lead us to generate list of genes which cause the AML.

Now by having the AML_Normal genes, we generate two subsets of this table. One for genes that upper the AML, where we have logFC greater than 1 and Adjusted P Value of less than 0.05. Likewise we generate the AML.down file for genes that upper the AML, where we have logFC less than -1 and Adjusted P Value of less than 0.05. These subsets are available in AML_Normal_up and AML_Normal_down text files respectively.

Gene Anthology and Pathway

The Gene Anthology and Pathway of gene set which upper or downer the AML expression can be fetched from the Enrichr website.

First we discuss the Pathway of the AML.up genes. By searching the up genes in the website, we attain couple of databases and we choose one with the highest height in the bar chart.



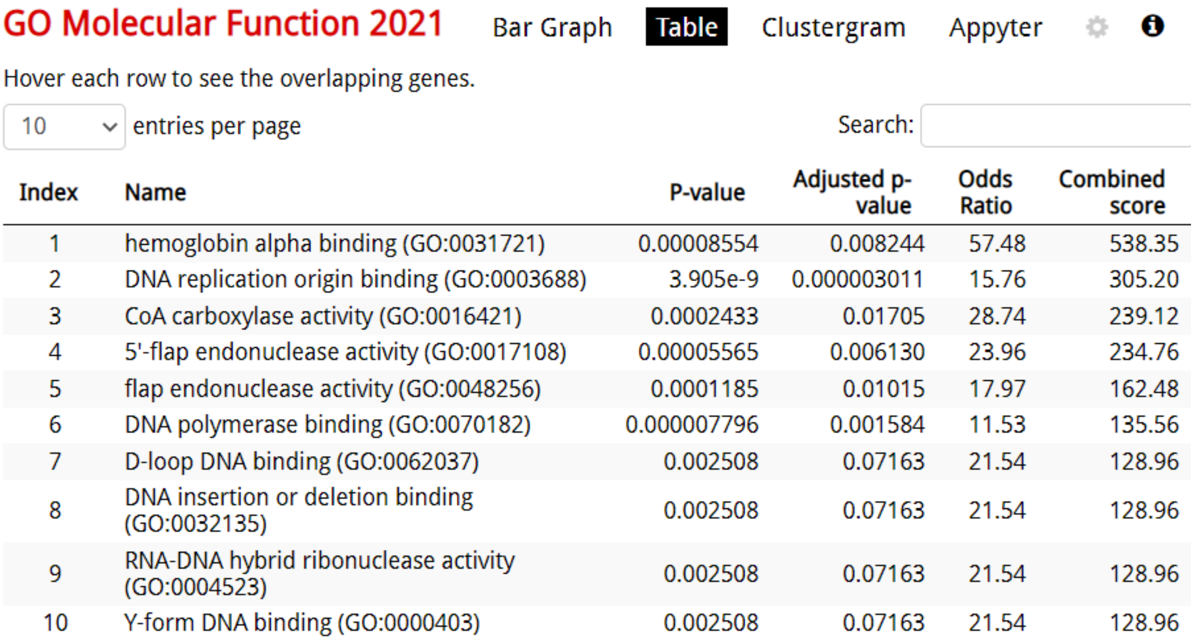
WikiPathway 2021 Human

Hover each row to see the overlapping genes.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Retinoblastoma gene in cancer WP2446	1.332e-23	6.861e-21	11.97	630.35
2	G1 to S cell cycle control WP45	6.411e-15	1.651e-12	9.98	326.18
3	DNA Replication WP466	2.781e-12	3.580e-10	11.99	319.15
4	Cell cycle WP179	2.132e-14	3.660e-12	6.03	189.94
5	Regulation of sister chromatid separation at the metaphase-anaphase transition WP4240	0.00002002	0.0006445	12.60	136.28
6	FBXL10 enhancement of MAP/ERK signaling in diffuse large B-cell lymphoma WP4553	1.396e-7	0.00001198	8.54	134.81
7	DNA IR-damage and cellular response via ATR WP4016	8.372e-10	8.623e-8	5.87	122.61
8	Gastric Cancer Network 1 WP2361	0.000001013	0.00005798	8.82	121.74
9	Fluoropyrimidine Activity WP1601	0.000004448	0.0002082	7.22	88.91
10	Histone Modifications WP2369	1.784e-7	0.00001215	5.32	82.70

Accordingly the AML.up genes Pathways are mostly related to the Retinoblastoma gene in cancer WP2446 genes from the WikiPathway 2021 Human database with Adjusted P Value of 6.861 e-21.

Likewise we now study the Ontology for these genes and attain couple of databases and we choose one with the highest height in the bar chart.



GO Molecular Function 2021

Hover each row to see the overlapping genes.

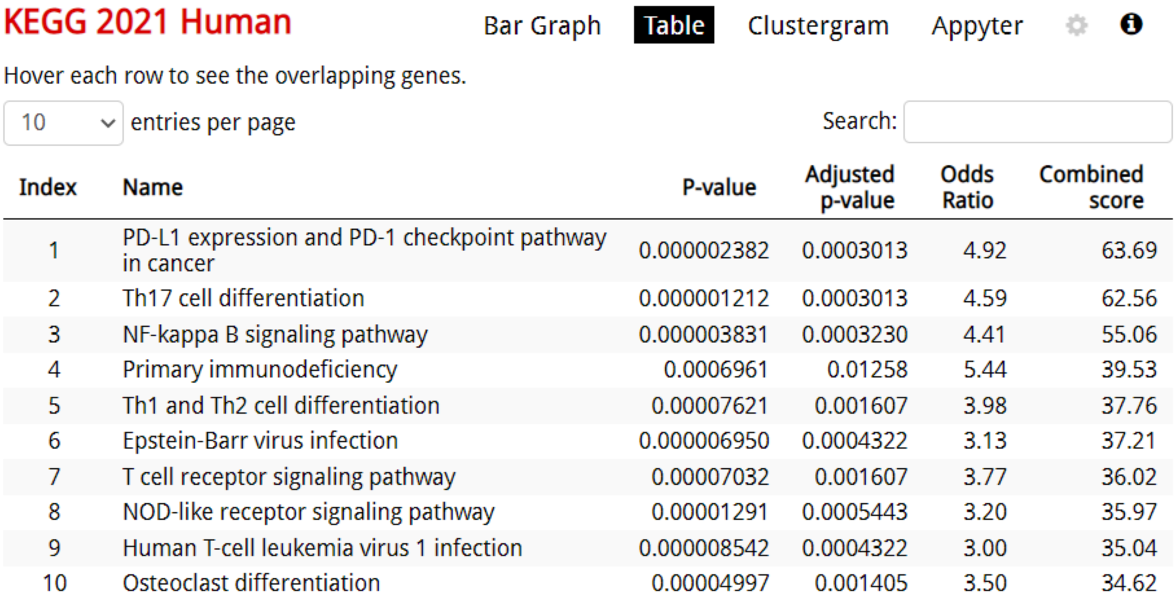
10 entries per page

Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	hemoglobin alpha binding (GO:0031721)	0.00008554	0.008244	57.48	538.35
2	DNA replication origin binding (GO:0003688)	3.905e-9	0.000003011	15.76	305.20
3	CoA carboxylase activity (GO:0016421)	0.0002433	0.01705	28.74	239.12
4	5'-flap endonuclease activity (GO:0017108)	0.00005565	0.006130	23.96	234.76
5	flap endonuclease activity (GO:0048256)	0.0001185	0.01015	17.97	162.48
6	DNA polymerase binding (GO:0070182)	0.000007796	0.001584	11.53	135.56
7	D-loop DNA binding (GO:0062037)	0.002508	0.07163	21.54	128.96
8	DNA insertion or deletion binding (GO:0032135)	0.002508	0.07163	21.54	128.96
9	RNA-DNA hybrid ribonuclease activity (GO:0004523)	0.002508	0.07163	21.54	128.96
10	Y-form DNA binding (GO:0000403)	0.002508	0.07163	21.54	128.96

Accordingly the AML.up genes Ontology are mostly related to the haemoglobin alpha binding (GO:0031721) genes from the GO Molecular Function 2021 with Adjusted P Value of 0.008244.

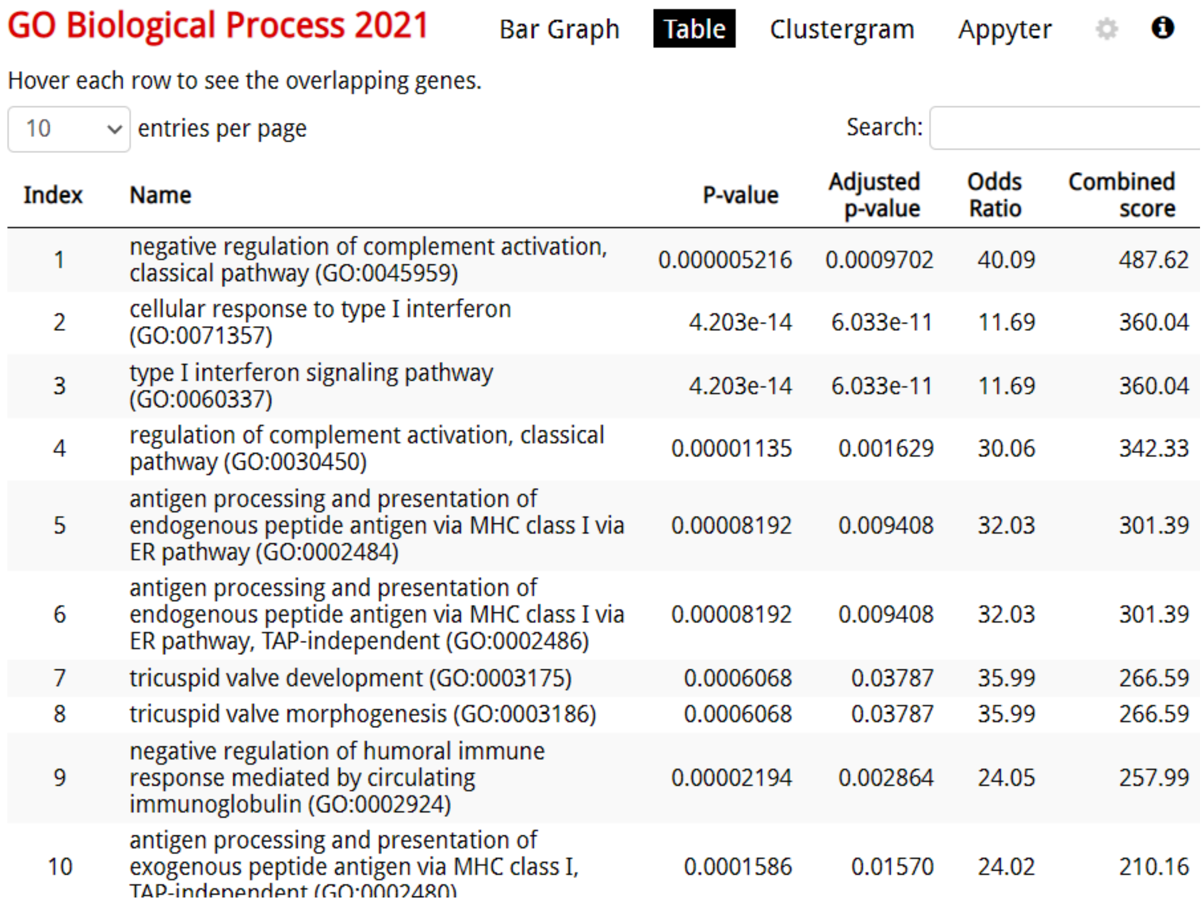
Now we discuss the Pathway of the AML.down genes. The instruction is as before. First we discuss the Pathway of the AML.down genes. By searching the up genes in the website, we attain couple of databases and we choose one with the highest height in the bar chart.



Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	PD-L1 expression and PD-1 checkpoint pathway in cancer	0.000002382	0.0003013	4.92	63.69
2	Th17 cell differentiation	0.000001212	0.0003013	4.59	62.56
3	NF-kappa B signaling pathway	0.000003831	0.0003230	4.41	55.06
4	Primary immunodeficiency	0.0006961	0.01258	5.44	39.53
5	Th1 and Th2 cell differentiation	0.00007621	0.001607	3.98	37.76
6	Epstein-Barr virus infection	0.000006950	0.0004322	3.13	37.21
7	T cell receptor signaling pathway	0.00007032	0.001607	3.77	36.02
8	NOD-like receptor signaling pathway	0.00001291	0.0005443	3.20	35.97
9	Human T-cell leukemia virus 1 infection	0.000008542	0.0004322	3.00	35.04
10	Osteoclast differentiation	0.00004997	0.001405	3.50	34.62

Accordingly the AML.down genes Pathways are mostly related to the PD-L1 expression and PD-1 checkpoint pathway genes from the KEGG 2021 Human with Adjusted P Value of 0.0003013.

Likewise we now study the Ontology for these genes and attain couple of databases and we choose one with the highest height in the bar chart.



GO Biological Process 2021

Hover each row to see the overlapping genes.

10 entries per page

Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	negative regulation of complement activation, classical pathway (GO:0045959)	0.000005216	0.0009702	40.09	487.62
2	cellular response to type I interferon (GO:0071357)	4.203e-14	6.033e-11	11.69	360.04
3	type I interferon signaling pathway (GO:0060337)	4.203e-14	6.033e-11	11.69	360.04
4	regulation of complement activation, classical pathway (GO:0030450)	0.00001135	0.001629	30.06	342.33
5	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway (GO:0002484)	0.00008192	0.009408	32.03	301.39
6	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent (GO:0002486)	0.00008192	0.009408	32.03	301.39
7	tricuspid valve development (GO:0003175)	0.0006068	0.03787	35.99	266.59
8	tricuspid valve morphogenesis (GO:0003186)	0.0006068	0.03787	35.99	266.59
9	negative regulation of humoral immune response mediated by circulating immunoglobulin (GO:0002924)	0.00002194	0.002864	24.05	257.99
10	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent (GO:0002480)	0.0001586	0.01570	24.02	210.16

Accordingly the AML.down genes Ontology are mostly related to the negative regulation of complement activation, classical pathway (GO:0045959) genes from the GO Biological Process 2021 with Adjusted P Value of 0.0003013.