

# Final Project

Microarray Series GSE48558 Differential Expression Analysis

Sharif University of Technology

Student: Hossein Jafarinia(400211557)

Professors: Dr. Somayyeh Koohi, Dr. Ali Sharifi-Zarchi



## An Introduction to Microarray

A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time. DNA microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence or gene. Often, these slides are referred to as gene chips or DNA chips. The DNA molecules attached to each slide act as probes to detect gene expression, which is also known as the transcriptome or the set of messenger RNA (mRNA) transcripts expressed by a group of genes.

To perform a microarray analysis, mRNA molecules are typically collected from both an experimental sample and a reference sample. For example, the reference sample could be collected from a healthy individual, and the experimental sample could be collected from an individual with a disease like cancer. The two mRNA samples are then converted into complementary DNA (cDNA), and each sample is labeled with a fluorescent probe of a different color. For instance, the experimental cDNA sample may be labeled with a red fluorescent dye, whereas the reference cDNA may be labeled with a green, fluorescent dye. The two samples are then mixed together and allowed to bind to the microarray slide. The process in which the cDNA molecules bind to the DNA probes on the slide is called hybridization. Following hybridization, the microarray is scanned to measure the expression of each gene printed on the slide. If the expression of a particular gene is higher in the experimental sample than in the reference sample, then the corresponding spot on the microarray appears red. In contrast, if the expression in the experimental sample is lower than in the reference sample, then the spot appears green. Finally, if there is equal expression in the two samples, then the spot appears yellow. The data gathered through microarrays can be used to create gene expression profiles, which show simultaneous changes in the expression of many genes in response to a particular condition or treatment. [1]



Figure 1. Microarray fills holes [1]

## Introduction to Differential Expression Analysis

There are many databases for microarray datasets and one of the most popular ones is Genome Expression Omnibus (GEO) database [2] and their microarray information can be used and analyzed with different methods and one of the most popular ones is Differential Expression analysis which can be done with different computer programs and one of the most popular ones is with R programming language. The method is generally composed of two parts. First is quality control the second is the analysis itself. More details are as follows:

### Quality Control:

Quality control should always be done because there may be cases that the experiment is not done with good care or as precise as it should and many different research can be done based on the wrong results of this experiment. For example, if the samples aren't cleaned from a certain bacteria called mycoplasma, we essentially will analyze the bacteria's gene too which gives us wrong results.

In this part we should make some plots like box plot on samples gene data and see if the data are reasonable or not. For example, if there is a sample that its box plot is highly different than others it probably had a lot more mRNA it's better to delete that sample from analysis. And before all these the data should have been normalized (for example with quantile normalizing method) and scaled in  $\log_2$  factor so results be more meaningful and realistic.

### Dimensionality Reduction:

This part is essentially another part of the quality control. Usually there are many genes being experimented in these kinds of experiments. For example in this microarray experiment there are 32321 genes experimented so our data is going to be 32321 dimension which is impossible for human to comprehend so we need to do dimensionality reduction with methods like Principal Component Analysis (PCA) which can show us how clustered and differentiated are our data and if they match the overall design section of dataset explanation we can be sure of two things. First, we can be sure that meaningful genes are selected to be experimented and secondly the experiment is done well. This is the most important part of quality control.

#### Correlation Analysis:

This part can still be considered another part of quality control. In this part we should compute the probabilistic correlation between samples gene data then cluster them see if the clusters are meaningful or not. For example, aside from each sample with itself Normal samples should have the highest correlation with each other than AML samples should have more but less than Normal samples because a tissues cancer cells are bit different from each other.

#### Differential Expression Analysis:

This is the most important part. In this part based different methods we should analyze if the higher expression of the more expressed genes and lower expression of less expressed genes has meaningful effect on AML or not. We generally must define a test statistic and based on that test statistics calculate the P. Value. In biological studies if P. Value is  $< 0.05$  or  $< 0.02$  we consider it meaningful. Secondly based on biological studies we need to calculate logFC which tells us how many folds the higher expression is. And based on our knowledge in biological studies we want it to be at least  $> 1$  for higher expressed genes and  $< -1$  for lower expressed genes.

#### Pathway and Gene Ontology Analysis:

After finding the meaningful and valuable genes in the disease we can use the studies and analysis of databases like Enrichr [3]. We usually choose ~250-300 genes and give those gene names to Enrichr [3] and after their analysis we can find gene ontology and pathways related to our gene list.

## Dataset

The dataset is Series GSE48558 [4]

Status: Public on Jul 06, 2013

Title: Expression data from normal and Malignant hematopoietic cells

Organism: Homo sapiens

Experiment type: Expression profiling by array

Summary: This data was used to determine levels of BRCA1 and BRCA2 in primary human leukemia samples. Samples were determined to be high BRCA1 and/or BRCA2 or low BRCA1 and/or BRAC2.

This data was used to determine levels of BRCA1 and BRCA2 in primary human leukemia samples. Samples were determined to be high BRCA1 and/or BRCA2 or low BRCA1 and/or BRAC2.

Overall design: AML cell lines and patient samples, B ALL cell lines and Patient samples, T ALL cell lines and patient samples, normal B cells, normal granulocytes, normal monocytes, normal T cells and normal CD34+ cells were used for RNA extraction and hybridization on Affymetrix microarrays. All the AML, B ALL, T ALL cell lines were cultured in vitro under appropriate culture conditions and harvested in their log phase growth for RNA extraction. AML, B ALL, and T ALL patient samples were collected...(I assume these are the PBMCs from either peripheral blood or bone marrow from patients, please confirm). Normal B cells, granulocytes, monocytes, and T cells were purified from human peripheral blood of normal healthy donors.

Contributor: Civin CI, Civin CI

Submission date: Jul 05, 2013

Last update date: Jul 26, 2018

E-mails: ysun@som.umaryland.edu

Organization name: U. of Maryland, Baltimore

[4]

## Method

### Quality Control:

Based on the GSE48558Analysis.r which is available both in the end of this report and in GSE48558Analysis.r file I have checked max and min values for gene expression values which are respectively 13.76154 and 1.611473 tells us data is in log2 scale and the resulted boxplot in Figure 2 shows that the data are normalized and experimented well and there are no unusual outliers on samples.

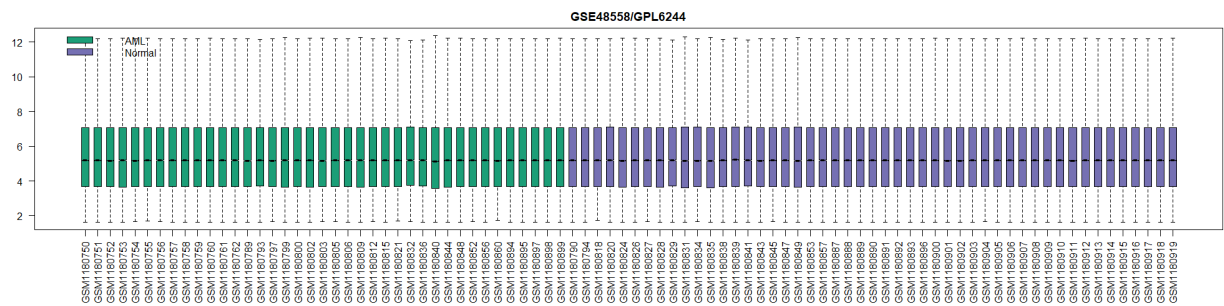


Figure 2. Box plot for each samples genes expression

### Dimensionality Reduction:

For dimensionality reduction I used PCA method and since the best thing we can show on 2D monitor I just used PC1 and PC2. The code for these two parts can be found in GSE48558Analysis.r.

Figure 3 shows the plot for genes. In this one an extra job of expression value – mean is also done so we get more meaningful result of difference in gene expression and not just expression itself.

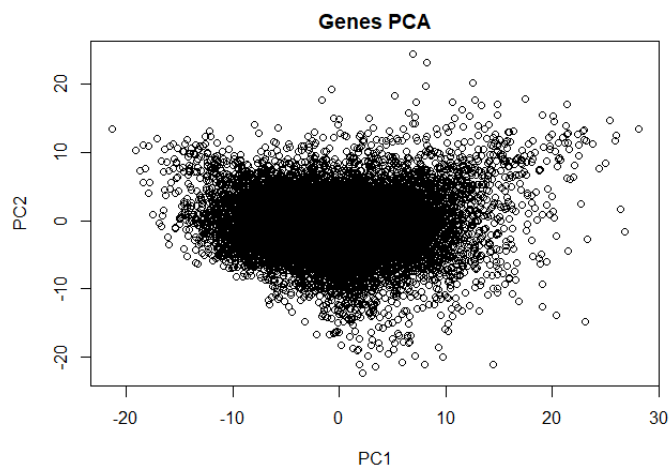
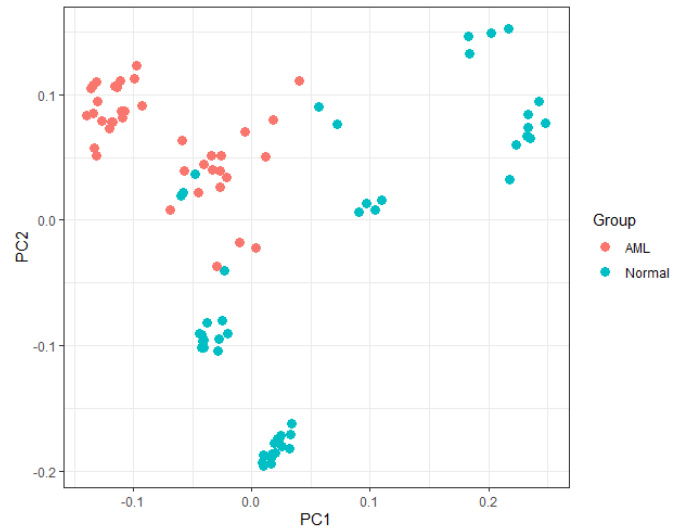


Figure 3. Genes PCA plot. Each Circle is a gene

Figure 4 shows the plot for samples. As we can see there are generally five clusters of Normal samples and that matches the experiment description's overall design which says there are five groups of normal samples.





## Correlation Analysis:

I calculated the correlation between each sample and clustered more close ones and plotted the values in a heat map as we can see in GSE48558Analysis.r and the result is in Figure 5. As we can see Normal samples have the highest correlation, Normal samples have the lowest correlation with AML samples and AML samples have some amounts of correlation as we know from biological studies that even cancer cells of a specific tissue can have some levels of difference with each other.

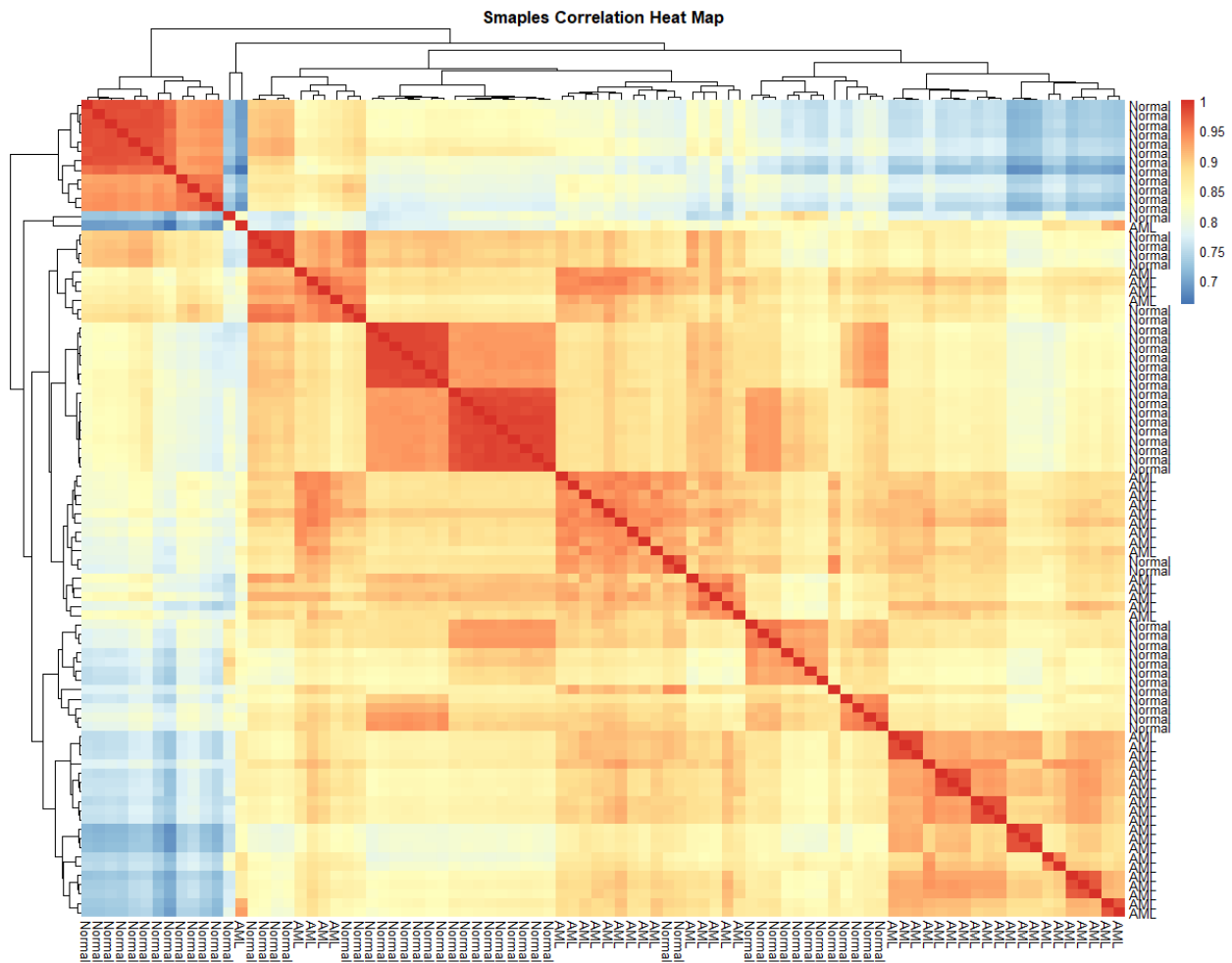


Figure 5. Correlation Heat Map

## Differential Expression Analysis:

In this part I used library limma which uses a really complex line fitting method and uses a test statistic named B to calculate P. Values and for the final gene list only took the ones with lower than 0.05 P. Value and  $> 1$  or  $< -1$  logFC(AML / Normal). The code for it can be found in GSE48558Analysis.r and most important genes which are adj.P.Value  $< 0.05$  and logFC  $> 1$  -> Up Genes adj.P.Value  $< 0.05$  and logFC  $< -1$  -> Down Genes can be found in Results folder named AMLDownGenes.txt and AMLUpGenes.txt respectively. Note that some genes are different names of the same gene so we can find in any gene database with ease.

Top 10 of these some of these genes can be seen in table 1.

Gene.symbol	Gene.ID	adj.P.Val	logFC
MPO	4353	3.617813e-19	5.5635012
FLT3	2322	4.835716e-19	5.2500645
KIAA0101	9768	6.308160e-19	4.5591352
BUB1B	701	1.664043e-18	2.7565536
SUCNR1	56670	1.938573e-18	2.9968155
MCM10	55388	3.712137e-18	2.3188477
TPX2	22974	4.695529e-18	3.1564149
CIT	11113	1.147946e-17	2.3707507
CDC45	8318	1.658665e-17	2.2875007
IQGAP3	128239	1.775540e-17	1.6696974

Table 1. Effective genes with their adj.P.Value and logFC

## Pathway and Gene Ontology Analysis:

I took ~250-300 top genes of the Up Genes list and submit that to Enrichr [3] and chose the most meaningful ones with adj.P.Value < 0.05.

here are the top 10 results of Pathway analysis from WikiPathway 2021 Human (the full list can be found at WikiPathway\_2021\_Human\_table.txt (the full list can be found at WikiPathway\_2021\_Human\_table\_up/down.txt):

WikiPathway 2021 Human					
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	Retinoblastoma gene in cancer WP2446	1.589e-25	3.973e-23	28.90	1650.30
2	Cell cycle WP179	3.254e-19	4.068e-17	16.84	716.95
3	G1 to S cell cycle control WP45	2.422e-14	2.018e-12	21.18	664.12
4	DNA Replication WP466	7.409e-13	4.631e-11	27.42	765.75
5	DNA IR-damage and cellular response via ATR WP4016	1.280e-11	6.400e-10	14.61	366.54
6	Gastric Cancer Network 1 WP2361	1.832e-7	0.000007632	21.44	332.65
7	DNA damage response WP707	7.526e-7	0.00002688	10.33	145.68
8	miRNA regulation of DNA damage response WP1530	0.000001094	0.00003418	9.83	134.93
9	Fluoropyrimidine Activity WP1601	0.000008372	0.0002325	14.92	174.44
10	DNA Repair Pathways Full Network WP4946	0.00001308	0.0003269	6.16	69.30

Table 2. Top 10 meaningful UP results from WikiPathway 2021 Human

WikiPathway 2021 Human					
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	Type II interferon signaling (IFNG) WP619	0.0002127	0.02515	10.45	88.39
2	Glioblastoma signaling pathways WP2261	0.0002270	0.02515	6.27	52.64
3	DNA damage response (only ATM dependent) WP710	0.0002490	0.02515	5.28	43.83
4	Immune response to tuberculosis WP4197	0.0003464	0.02543	14.05	111.92
5	Interactions between immune cells and microRNAs in tumor microenvironment WP4559	0.0007552	0.02543	11.12	79.91
6	Rett syndrome causing genes WP4312	0.0007307	0.02543	7.77	56.15
7	Hepatitis B infection WP4666	0.0004804	0.02543	4.24	32.43
8	Ebola Virus Pathway on Host WP4217	0.0007261	0.02543	4.45	32.15
9	Insulin Signaling WP481	0.0006959	0.02543	4.02	29.21
10	Type I interferon induction and signaling during SARS-CoV-2 infection WP4868	0.001120	0.03395	9.88	67.13

Table 3. Top 10 meaningful Down results from WikiPathway 2021 Human

It is also worth to note results from Kegg 2021 Human (the full list can be found at KEGG\_2021\_Human\_table\_up/down.txt):

**KEGG 2021 Human** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	Cell cycle	4.186e-20	7.116e-18	17.11	763.28
2	p53 signaling pathway	1.298e-7	0.00001104	10.79	171.03
3	Cellular senescence	0.00002367	0.001341	5.15	54.87
4	Progesterone-mediated oocyte maturation	0.0001286	0.004809	5.86	52.50
5	Oocyte meiosis	0.0001414	0.004809	5.06	44.89
6	DNA replication	0.0001862	0.005275	10.79	92.68
7	Homologous recombination	0.0003482	0.008457	9.29	73.97
8	Folate biosynthesis	0.0005645	0.01200	12.13	90.72
9	Transcriptional misregulation in cancer	0.0006501	0.01228	3.71	27.23
10	MicroRNAs in cancer	0.0008391	0.01426	2.97	21.03

Table 4. Top 10 meaningful UP results from KEGG 2021 Human

**KEGG 2021 Human** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	FoxO signaling pathway	0.000004489	0.0004467	6.23	76.74
2	Epstein-Barr virus infection	0.000002266	0.0004467	5.10	66.27
3	Cellular senescence	0.00002367	0.001570	5.15	54.87
4	PD-L1 expression and PD-1 checkpoint pathway in cancer	0.0003764	0.009788	5.74	45.22
5	Measles	0.0002484	0.009788	4.67	38.78
6	Sphingolipid signaling pathway	0.0004246	0.009788	4.85	37.67
7	Viral carcinogenesis	0.0002516	0.009788	3.88	32.16
8	Human T-cell leukemia virus 1 infection	0.0004797	0.009788	3.58	27.36
9	Endocytosis	0.0004346	0.009788	3.39	26.25
10	Human papillomavirus infection	0.0004919	0.009788	3.00	22.88

Table 5. Top 10 meaningful Down results from KEGG 2021 Human

And results from Reactom 2016 (the full result can be found at Reactome\_2016\_table\_up/down.txt):

**Reactome 2016** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	Cell Cycle, Homologous Recombination R-HSA-1640170	4.432e-37	4.432e-37	10.88	979.56
2	Cell Cycle, Mitotic Homologous Recombination R-HSA-69278	1.229e-35	3.497e-33	11.22	902.07
3	Mitotic G1-G1/S phases Homologous Recombination R-HSA-453279	4.096e-19	7.768e-17	15.26	646.27
4	G1/S-Specific Transcription Homologous Recombination R-HSA-69205	5.784e-19	8.228e-17	164.71	6916.64
5	E2F mediated regulation of DNA replication Homologous Recombination R-HSA-113510	1.303e-17	1.483e-15	50.89	1978.40
6	G1/S Transition Homologous Recombination R-HSA-69206	1.935e-17	1.835e-15	16.28	626.45
7	Cell Cycle Checkpoints Homologous Recombination R-HSA-69620	4.272e-16	3.473e-14	10.79	382.02
8	Mitotic Prometaphase Homologous Recombination R-HSA-68877	1.953e-15	1.389e-13	15.12	512.23
9	S Phase Homologous Recombination R-HSA-69242	2.445e-15	1.546e-13	13.51	454.46
10	DNA Replication Homologous Recombination R-HSA-69306	2.966e-13	1.688e-11	13.44	387.57

Table 6. Top 10 meaningful UP results from Reactom 2016

**Reactome 2016** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	Interferon alpha/beta signaling Homologous Recombination R-HSA-909733	0.000007716	0.004722	9.00	105.94
2	Antigen Presentation: Folding, assembly and peptide loading of class I MHC Homologous Recombination R-HSA-983170	0.00003002	0.009185	16.74	174.28
3	G beta:gamma signalling through PI3Kgamma Homologous Recombination R-HSA-392451	0.00007685	0.01568	9.59	90.81
4	G-protein beta:gamma signalling Homologous Recombination R-HSA-397795	0.0001086	0.01653	8.94	81.64
5	GPVI-mediated activation cascade Homologous Recombination R-HSA-114604	0.0001351	0.01653	8.56	76.30
6	Cytokine Signaling in Immune system Homologous Recombination R-HSA-1280215	0.0001732	0.01767	2.54	21.97
7	Innate Immune System Homologous Recombination R-HSA-168249	0.0002101	0.01837	2.31	19.54
8	Immune System Homologous Recombination R-HSA-168256	0.0002450	0.01874	1.92	15.96
9	ER-Phagosome pathway Homologous Recombination R-HSA-1236974	0.0004183	0.02845	6.82	53.03
10	Interferon gamma signaling Homologous Recombination R-HSA-877300	0.0004920	0.03011	5.47	41.65

Table 7. Top 10 meaningful Down results from Reactom 2016

Here are the top 10 results of Ontology Analysis from Go Cellular Component (the full result can be found at GO\_Cellular\_Component\_2021\_table\_up/down.txt):

**GO Cellular Component 2021** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	spindle (GO:0005819)	1.117e-11	1.843e-9	8.14	205.25
2	cyclin-dependent protein kinase holoenzyme complex (GO:0000307)	9.994e-9	8.245e-7	24.59	453.00
3	serine/threonine protein kinase complex (GO:1902554)	6.020e-8	0.000003311	18.65	310.04
4	intracellular membrane-bounded organelle (GO:0043231)	9.638e-8	0.000003976	1.91	30.80
5	microtubule cytoskeleton (GO:0015630)	1.442e-7	0.000004759	4.47	70.40
6	microtubule (GO:0005874)	6.525e-7	0.00001794	5.71	81.34
7	azurophil granule lumen (GO:0035578)	9.558e-7	0.00002253	8.49	117.63
8	nucleus (GO:0005634)	0.000001159	0.00002390	1.84	25.12
9	mitotic spindle (GO:0072686)	0.000001899	0.00003482	6.87	90.44
10	nuclear chromosome (GO:0000228)	0.000004127	0.00006810	8.23	102.05

Table 8. Top 10 meaningful UP results from GO Cellular Component 2021

**GO Cellular Component 2021** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

Hover each row to see the overlapping genes.

10 ▾ entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
-------	------	---------	------------------	------------	----------------

Table 9. Top 10 meaningful Down results from GO Cellular Component 2021

It is worth to note results from GO Biological Process 2021(the full result can be found at GO\_Biological\_Process\_2021\_table\_up/down.txt):

**GO Biological Process 2021** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	microtubule cytoskeleton organization involved in mitosis (GO:1902850)	1.500e-18	1.357e-15	15.55	638.29
2	mitotic spindle organization (GO:0007052)	9.808e-19	1.357e-15	13.53	560.89
3	DNA metabolic process (GO:0006259)	1.554e-15	9.377e-13	8.07	275.22
4	mitotic sister chromatid segregation (GO:0000070)	1.221e-14	5.525e-12	14.96	479.25
5	regulation of transcription involved in G1/S transition of mitotic cell cycle (GO:0000083)	6.772e-13	2.451e-10	35.79	1002.97
6	G1/S transition of mitotic cell cycle (GO:0000082)	2.061e-12	6.216e-10	14.81	398.57
7	mitotic cytokinesis (GO:0000281)	5.632e-12	1.395e-9	22.22	575.59
8	DNA replication (GO:0006260)	6.166e-12	1.395e-9	12.05	311.05
9	DNA-dependent DNA replication (GO:0006261)	9.740e-11	1.959e-8	9.80	225.92
10	DNA replication initiation (GO:0006270)	1.567e-10	2.836e-8	24.31	548.88

Table 10. Top 10 meaningful UP results from GO Biological Process 2021

**GO Biological Process 2021** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	cellular response to type I interferon (GO:0071357)	0.000005477	0.002266	9.47	114.78
2	type I interferon signaling pathway (GO:0060337)	0.000005477	0.002266	9.47	114.78
3	peptidyl-serine phosphorylation (GO:0018105)	0.000004138	0.002266	5.68	70.39
4	protein phosphorylation (GO:0006468)	0.000001790	0.002266	3.39	44.83
5	peptidyl-serine modification (GO:0018209)	0.000009417	0.002843	5.20	60.24
6	phosphorylation (GO:0016310)	0.00001031	0.002843	3.44	39.51
7	protein autophosphorylation (GO:0046777)	0.0001442	0.03409	4.54	40.16

Table 11. Top 10 meaningful Down results from GO Biological Process 2021

Also, worth to note results from Go Molecular Function 2021(the full result can be found at GO\_Molecular\_Function\_2021\_table\_up/down.txt):

**GO Molecular Function 2021** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	tubulin binding (GO:0015631)	2.058e-7	0.00006810	4.57	70.42
2	microtubule binding (GO:0008017)	4.559e-7	0.00007544	5.10	74.47
3	DNA replication origin binding (GO:0003688)	8.653e-7	0.00009547	23.71	331.01
4	microtubule motor activity (GO:0003777)	0.000001732	0.0001433	11.26	149.33
5	cyclin-dependent protein serine/threonine kinase regulator activity (GO:0016538)	0.000003712	0.0002457	12.74	159.31
6	single-stranded DNA helicase activity (GO:0017116)	0.000007071	0.0003344	23.92	283.62
7	motor activity (GO:0003774)	0.000006152	0.0003344	9.31	111.71
8	DNA polymerase binding (GO:0070182)	0.0001270	0.005253	19.07	171.06
9	single-stranded DNA binding (GO:0003697)	0.0006342	0.01860	5.22	38.46
10	protein kinase regulator activity (GO:0019887)	0.0006744	0.01860	5.17	37.72

Table 12. Top 10 meaningful UP results from GO Molecular Function 2021

**GO Molecular Function 2021** Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	protein serine/threonine kinase activity (GO:0004674)	0.000004736	0.001312	3.81	46.68
2	TAP1 binding (GO:0046978)	0.00003235	0.004481	99.83	1032.08

Table 13. Top 10 meaningful Down results from GO Molecular Function 2021



And also, worth to note the results form MGI Mammalian Phenotype Level 4 2021 (the full result can be found at MGI\_Mammalian\_Phenotype\_Level\_4\_2021\_table\_up/down.txt):

**MGI Mammalian Phenotype Level 4 2021**

Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	abnormal mitosis MP:0004046	4.157e-8	0.00004211	12.36	210.07
2	small testis MP:0001147	5.050e-8	0.00004211	4.58	76.87
3	abnormal cell nucleus morphology MP:0003111	2.092e-7	0.00005815	15.45	237.58
4	chromosomal instability MP:0008866	2.092e-7	0.00005815	15.45	237.58
5	decreased testis weight MP:0004852	1.729e-7	0.00005815	5.51	85.83
6	embryonic lethality between implantation and somite formation, complete penetrance MP:0011096	1.150e-7	0.00005815	5.02	80.28
7	abnormal erythrocyte morphology MP:0002447	2.462e-7	0.00005866	9.99	152.02
8	abnormal mitotic spindle morphology MP:0009760	3.008e-7	0.00006271	19.65	295.15
9	abnormal blastocyst morphology MP:0004957	6.476e-7	0.0001200	25.19	359.01
10	abnormal cell cycle MP:0003077	9.544e-7	0.0001592	7.41	102.75

Table 14. Top 10 meaningful UP results from MGI Mammalian Phenotype Level 4 2021

**MGI Mammalian Phenotype Level 4 2021**

Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	abnormal B cell differentiation MP:0002144	0.000007380	0.01097	7.61	89.94
2	abnormal lymph node cortex morphology MP:0002343	0.00006399	0.03170	66.55	642.64
3	decreased response to antigen MP:0020001	0.00004354	0.03170	26.70	268.11
4	decreased single-positive T cell number MP:0008083	0.00009706	0.03606	9.15	84.53
5	abnormal thymus medulla morphology MP:0002375	0.0001270	0.03773	19.07	171.06
6	abnormal T cell differentiation MP:0002145	0.0001691	0.04187	5.61	48.76

Table 15. Top 10 meaningful Down results from MGI Mammalian Phenotype Level 4 2021

## Conclusion

Although this data has been analyzed before but in a hypothetical situation, I can confidently say that I did Differential Expression Analysis on a type of leukemia and found highly effective genes in which have significant effects on the disease and this result can be shared with Physicians, Biologists, Pharmacists, ... and we may be able to cure this disease with methods like inhibit some of these pathways. Although in real situations I'm actually able to analyze new data too.



For down genes there are generally two cases either for random or external reasons for example radiation the expression of some genes have been decreased and/or increase in expression of some genes caused this decrease in expression and they also have pathway.

For gene ontologies:

The Gene Ontology is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species [8]. It has three major sections which I'm going to explain them and their relation to our study

GO Cellular Component:

In this part we are concerned about the cellular component that are involved in this disease. For example, as a very significant case CMG complex (GO:0071162) is A protein complex that contains the GINS complex, Cdc45p, and the heterohexameric MCM complex, and that is involved in unwinding DNA during replication [9]. And we all know that cancer is highly related to cell division and DNA unwinding.

GO Biological Process:

In this part we are concerned about the biological processes that are involved in this disease. For example, as a very significant case we have mitotic spindle elongation (GO:0000022) is the cell cycle process in which the distance is lengthened between poles of the mitotic spindle. Mitotic spindle elongation begins during mitotic prophase and ends during mitotic anaphase B [10]. We all know that cancer is highly related to cell division and DNA unwinding.

GO Molecular Function:

In this part we are concerned about the molecular function that are involved in this disease. For example, as a very significant case as a very significant case DNA replication origin binding (GO:0003688) is binding to a DNA replication origin, a unique DNA sequence of a replicon at which DNA replication is initiated and proceeds bidirectionally or unidirectionally [11].

## II- Transcription Analysis

Here I analyzed the results of transcription factors related the found genes (the full result can be found at TRANSFAC\_and\_JASPAR\_PWMs\_table\_up/down.txt)

### ENCODE and ChEA Consensus TFs from ChIP-X

Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	E2F4 ENCODE	4.331e-90	4.418e-88	20.39	4195.55
2	FOXM1 ENCODE	8.212e-29	4.188e-27	30.28	1958.04
3	SIN3A ENCODE	2.435e-19	8.280e-18	4.56	195.42
4	E2F6 ENCODE	2.959e-16	7.546e-15	2.90	103.62
5	NFYB ENCODE	1.098e-12	2.240e-11	2.49	68.45
6	NFYA ENCODE	1.638e-12	2.785e-11	2.80	75.92
7	E2F1 CHEA	6.471e-10	9.429e-9	3.45	73.08
8	GATA1 CHEA	0.00003342	0.0004261	2.51	25.87
9	SALL4 CHEA	0.00009575	0.001085	3.23	29.88
10	NELFE ENCODE	0.0002217	0.002262	3.67	30.87

Table 16. Top 10 meaningful UP results from TRANSFAC and JASPAR PWMs

### ENCODE and ChEA Consensus TFs from ChIP-X

Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

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	IRF8 CHEA	2.742e-7	0.00002416	7.52	113.55
2	SPI1 CHEA	4.930e-7	0.00002416	2.67	38.81
3	STAT3 ENCODE	0.00009830	0.002813	2.48	22.85
4	RUNX1 CHEA	0.0001148	0.002813	2.07	18.80
5	TCF3 ENCODE	0.0003885	0.006757	2.21	17.35
6	SMC3 ENCODE	0.0004137	0.006757	2.00	15.62
7	VDR CHEA	0.0009656	0.01143	5.74	39.87
8	RELA ENCODE	0.001050	0.01143	2.48	17.03
9	SPI1 ENCODE	0.0008444	0.01143	1.82	12.87
10	PBX3 ENCODE	0.001400	0.01372	1.85	12.18

Table 17. Top 10 meaningful Down results from TRANSFAC and JASPAR PWMs

### III- Diseases/Drugs Analysis

Here I analyzed the results of diseases/drugs related the found genes (the full result can be found at LINC L1000\_Ligand\_Perturbations\_up\_table\_up/down.txt)

#### LINCS L1000 Ligand Perturbations up

Bar Graph **Table** Clustergram Appyter  

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	EGF-MCF10A	6.568e-51	5.714e-49	34.59	3997.02
2	BTC-MCF10A	8.494e-42	3.695e-40	27.88	2636.18
3	TGFA-MCF10A	1.218e-37	3.531e-36	25.20	2142.04
4	EPR-MCF10A	4.655e-26	1.013e-24	16.99	991.02
5	IGF1-MCF7	4.973e-8	8.653e-7	6.57	110.57
6	BFGF-HS578T	6.155e-7	0.000008924	6.26	89.54
7	HGF-BT20	0.00002994	0.0003721	5.01	52.21

Table 18. Top 10 meaningful UP results from LINC L1000 Ligand Perturbations up

#### LINCS L1000 Ligand Perturbations up

Bar Graph **Table** Clustergram Appyter  

Hover each row to see the overlapping genes.



10	▼	entries per page	Search: <input type="text"/>		
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	IFNG-BT20	1.202e-15	4.836e-14	12.87	441.99
2	IFNA-SKBR3	1.630e-15	4.836e-14	12.64	430.54
3	IFNG-MCF10A	1.630e-15	4.836e-14	12.64	430.54
4	IFNA-MCF7	3.415e-15	7.598e-14	12.12	403.82
5	IFNA-BT20	1.774e-14	3.157e-13	12.00	379.92
6	IFNG-MCF7	1.059e-13	1.570e-12	11.76	351.43
7	IFNG-SKBR3	1.832e-12	2.329e-11	10.72	289.77
8	IFNA-MDAMB231	8.975e-12	9.984e-11	9.64	245.30
9	IFNG-HS578T	2.712e-10	2.682e-9	9.07	199.87
10	IFNG-MDAMB231	1.050e-8	9.348e-8	8.22	151.06

Table 19. Top 10 meaningful Down results from LINC L1000 Ligand Perturbations up


#### IV- Cell Types Analysis

Here I analyzed the results of cell types related the found genes (the full result can be found at Human\_Gene\_Atlas\_table\_up/down.txt)

### Human Gene Atlas

Bar Graph **Table** Grid Network Clustergram Appyter  

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	CD105+ Endothelial	6.308e-22	3.722e-20	8.78	428.50
2	CD71+ EarlyErythroid	1.898e-21	5.601e-20	6.88	328.42
3	721 B lymphoblasts	1.393e-18	2.740e-17	3.93	161.61
4	Leukemia lymphoblastic(MOLT-4)	1.704e-13	2.514e-12	12.56	369.15
5	CD34+	1.955e-10	2.307e-9	4.01	89.62
6	Lymphoma burkitts(Daudi)	0.000001229	0.00001209	6.44	87.70
7	Bonemarrow	0.0001213	0.001022	8.75	78.90
8	Leukemia chronicMyelogenousK-562	0.002644	0.01950	4.01	23.82
9	Placenta	0.04196	0.2751	1.87	5.93
10	Adrenal gland	0.07349	0.4336	3.21	8.38

Table 20. Top 10 meaningful UP results from Human Gene Atlas

### Human Gene Atlas

Bar Graph **Table** Grid Network Clustergram Appyter  

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	CD8+ Tcells	1.335e-17	7.611e-16	5.75	223.49
2	CD4+ Tcells	1.293e-15	3.209e-14	5.65	193.66
3	CD56+ NKCells	1.689e-15	3.209e-14	4.76	161.90
4	CD19+ BCells(neg. sel.)	2.782e-11	3.965e-10	5.20	126.33
5	WholeBlood	9.855e-10	1.123e-8	4.26	88.25

Table 21. Top 10 meaningful Down results from Human Gene Atlas

## V- Literature Review

In this part I'm going to check real fidelity of the acquired results by literature review.

In these results the first database name is for up gene and next one is for down gene. What comes after “->” is for the result and ==> is for papers that confirm the results.

wikipathway -> Retinoblastoma gene in cancer wp2446 ==> The retinoblastoma gene (rb1) in acute myeloid leukaemia: analysis of gene rearrangements, protein expression and comparison of disease outcome, Growth-factor stimulation reveals two mechanisms of retinoblastoma gene inactivation in human myelogenous leukemia cells

wikipathway -> type II interferon signaling (IFNG) wp619 ==> Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion

kegg 2021 human -> cell cycle ==> Cell cycle control in acute myeloid leukemia

kegg 2021 human -> foxo signaling pathway ==> AKT/FOXO Signaling Enforces Reversible Differentiation Blockade in Myeloid Leukemias

reactome 2016 -> cell cycle homo sapiens r-HSA-1640170 ==>

reactom 2016 -> interferon alpha/beta signaling homo sapiens r-hsa-909733 ==>

go cellular component 2021 -> spindle (GO:0005819) ==> Targeting aurora kinases as a potential prognostic and therapeutical biomarkers in pediatric acute lymphoblastic leukemia

go cellular component 2021 ->

go biological process 2021 -> microtubule cytoskeleton organization involved in mitosis (GO:1902850) ==> State-Transition Analysis of Time-Sequential Gene Expression Identifies Critical Points That Predict Development of Acute Myeloid Leukemia

go biological process 2021 -> cellular response to type i interferon ==> Interferon- $\alpha$  in acute myeloid leukemia: an old drug revisited, Recent Progress in Interferon Therapy for Myeloid Malignancies, Type 1 interferon to prevent leukemia relapse after allogeneic transplantation, Calreticulin promotes immunity and type I interferon-dependent survival in mice with acute myeloid leukemia, Type 1 interferon to prevent leukemia relapse after allogeneic transplantation, Chromosomal instability upregulates interferon in acute myeloid leukemia, Type I Interferon Signaling Predicts Inferior Survival in Patients with AML, Study Finds Type I Interferon May Enhance the Antileukemia Effect of Allogeneic Transplantation

go molecular function 2021 -> tubulin binding (go:00156631) ==>

go molecular function 2021 -> protein serine-threonine kinase activity ==> The serine-threonine kinase MNK1 is post-translationally stabilized by PML-RAR $\alpha$  and regulates differentiation of hematopoietic cells, Salt-inducible kinase inhibition suppresses acute myeloid leukemia progression in vivo, <https://ashpublications.org/blood/article/112/11/2655/61167/PKC412-Directly-Inhibits-the-Serine-Threonine>, Protein kinase inhibitors for acute leukemia, Therapeutic re-activation of protein phosphatase 2A in acute myeloid leukemia



mgm mammalian phenotype level 4 2021 -> abnormal mitosis ML:0004046 ==>

mgm mammalian phenotype level 4 2021 -> abnormal b cell differentiation mp:0002144 ==> Regulatory mechanisms of B cell responses and the implication in B cell-related diseases

encode and chea consensus tfs from chip-x -> e2f4 encode ==> E2F4 functions as a tumour suppressor in acute myeloid leukaemia via inhibition of the MAPK signaling pathway by binding to EZH2, E2F4 regulatory program predicts patient survival prognosis in breast cancer, E2F4 Actively Promotes the Initiation and Maintenance of Nerve Growth Factor-Induced Cell Differentiation, E2F4 Modulates Differentiation and Gene Expression in Hematopoietic Progenitor Cells during Commitment to the Lymphoid Lineage, Regulation of Trib2 by an E2F1-C/EBP $\alpha$  feedback loop in AML cell proliferation

encode and chea consensus tfs from chip-x -> irf8 chea ==> IRF8 Is an AML-Specific Susceptibility Factor That Regulates Signaling Pathways and Proliferation of AML Cells, Constitutive IRF8 expression inhibits AML by activation of repressed immune response signaling, Transcriptional plasticity drives leukemia immune escape, IRF-8 Controls Melanoma Progression by Regulating the Cross Talk between Cancer and Immune Cells within the Tumor Microenvironment

lincs l1000 ligands perturbations up -> egf-mcf10a ==> Using the MCF10A/MCF10CA1a Breast Cancer Progression Cell Line Model to Investigate the Effect of Active, Mutant Forms of EGFR in Breast Cancer Development and Treatment Using Gefitinib

lincs l1000 ligands perturbations up -> ifng-bt20 ==> Runx Transcription Factors in T Cells—What Is Beyond Thymic Development?, NPM1 upregulates the transcription of PD-L1 and suppresses T cell activity in triple-negative breast cancer

human gene atlas -> CD105+ Endothelial ==> CD105 (endoglin) as risk marker in AML patients undergoing stem cell transplantation, CD105 (Endoglin) as negative prognostic factor in AML, Functional and immunophenotypic characteristics of isolated CD105+ and fibroblast+ stromal cells from AML: implications for their plasticity along endothelial lineage, CD105 (Endoglin) Is Highly Overexpressed in a Subset of Cases of Acute Myeloid Leukemias, CD105 (endoglin) as risk marker in AML patients undergoing stem cell transplantation, CD105 (endoglin) is highly overexpressed in a subset of cases of acute myeloid leukemias, The Role of Bone Marrow Mesenchymal Stem Cells in Myelodysplastic Syndrome and Acute Myeloid Leukemia, Toll-like Receptor 4, Osteoblasts and Leukemogenesis; the Lesson from Acute Myeloid Leukemia

human gene atlas -> cd8+ tcells ==> Signatures of CD8+ T cell dysfunction in AML patients and their reversibility with response to chemotherapy, Signatures of CD8+ T cell dysfunction in AML patients and their reversibility with response to chemotherapy, CD8+ T cells expand stem and progenitor cells in favorable but not adverse risk acute myeloid leukemia, PD-1 and TIGIT Are Highly Co-Expressed on CD8+ T Cells in AML Patient Bone Marrow, Induction of leukemia-specific CD8+ cytotoxic t cells with autologous myeloid leukemic cells matured with a fiber-modified adenovirus encoding TNF- $\alpha$

VI- It's English

## Code

This is the code (GSE48558Analysis.r):

```
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
#####
# Differential expression analysis with limma
#+ Hossein Jafarinia modifications on R4.1.1 and Rstudio

#necessary libraries
library(GEOquery)
library(limma)
library(umap)
library(pheatmap)
library(ggplot2)
library(gplots)
library(reshape2)
library(plyr)
library(Biobase)
library(ggplot2)
library(reshape2)
library(plyr)
library(dplyr)

#Changing default VROOM_CONNECTION_SIZE so we can capture data
Sys.setenv("VROOM_CONNECTION_SIZE" = 131072 * 3)

#Making filing easy
curD <- dirname(rstudioapi::getActiveDocumentContext()$path)
setwd(sub(paste0("/", sub("(.)+", "", curD)), "", curD))

#load series and platform data from GEO
series = "GSE48558"
platform = "GPL6244"
gset <- getGEO(series, GSEMatrix=TRUE, AnnotGPL=TRUE, destdir = "Data/")
if (length(gset) > 1) idx <- grep(platform, attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

#make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group membership for all samples
# gsms <- paste0("111111111111XXXXXXXXXXXXXXXXXXXXXXXXX10XX10XX1X1",
#               # "1X1X11XX1XX1XX1XX0X01XX0X0000X01X001X0010X010X010",
#               # "XX10XX10XX1XXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",
#               # "00000000000000000000")
```

```

gsms <- paste0("111111111111XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0XXX0XXXX",
               "XXXXXXXXXXXXXXXXXXXXX0X0XXX0X0000X0XX00X00X0X0X0X0",
               "XXX0XXX0XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",
               "00000000000000000000")
sml <- strsplit(gsms, split="")[[1]]

# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
sml <- sml[sel]
smlWithName <- recode(sml, "1" = "AML", "0" = "Normal")
gset <- gset[, sel]

ex <- exprs(gset)

max(ex)
min(ex)
#As we can see data or in log2

#Quality Control:
#Drawing Boxplot for quality control
gs <- factor(smlWithName)
ord <- order(gs) # order samples by group
palette(c("#1B9E77", "#7570B3", "#E7298A", "#E6AB02", "#D95F02",
          "#66A61E", "#A6761D", "#B32424", "#B324B3", "#666666"))
par(mar=c(7,4,2,1))
title <- paste ("GSE48558", "/", annotation(gset), sep = "")
boxplot(ex[,ord], boxwex=0.6, notch=T, main=title, outline=FALSE, las=2,
col=gs[ord])
legend("topleft", groups, fill=palette(), legend = c("AML", "Normal"), bty="n")

corex = cor(ex)
pheatmap(corex, labels_row = smlWithName, labels_col = smlWithName, border_color
= NA, main = "Smamples Correlation Heat Map") #As we can see Normal samples of
different groups have highest correlation with each other and the lowest with AML
samples and although AMS samples have some level of correlation with each other
its not as much as normal samples due to the fact that cancer cells have some
levels of difference between each other

#Dimensionality reduction
#PCA for genes
ex.scale <- t(scale(t(ex), scale = F)) #ex - mean(ex)
pc <- prcomp(ex.scale) #Finding PCAs
plot(pc, main = "PCA", xlab = "PCs")

```

#only PC1 and PC2 are sufficient and thats generally what our 2 dimensional page can show best

```
plot(pc$x[,1:2], main = "Genes PCA") #As we can see genes have a relatively meaningful distribution
```

#PCA for samples

```
pcr <- data.frame(pc$r[, 1:3], Group = smlWithName) #r means rotation we at least mean 3 pcs
```

```
ggplot(pcr, aes(PC1, PC2, color = Group)) + geom_point(size = 3) + theme_bw() #As we can see AML and normal cells are generally separated well and we know from the examination summary says there are 5 types of Normal cells and we can see 5 clusters of normal cells in our plot which makes us conclude that the quality is good
```

#Differential Expression Analysis:

#assign samples to groups and set up design matrix

```
gset$group <- smlWithName
```

```
design <- model.matrix(~group + 0, gset) #finds pset
```

```
colnames(design) <- levels(gset) #design
```

#differential analysis main part by limma

```
fit <- lmFit(gset, design) # fit linear model its a linear model that fits a line to each model and based on the difference between line it say how much different they are
```

#set up contrasts of interest and recalculate model coefficients

```
cont.matrix <- makeContrasts(contrasts="AML-Normal", levels=design)
```

#cont.matrix and we see it wants to compare tumor and normal

```
fit2 <- contrasts.fit(fit, cont.matrix) #put the output in fit2
```

#compute statistics and table of top significant genes

```
fit2 <- eBayes(fit2, 0.01) #bayesian prior of 1% cancer related genes from biological studios
```

```
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf) # the function that calculated p.values and logFC "fdr" is our method for adjustment "B" is limmas test statistics and Inf for all genes
```

```
tT = subset(tT, select=c("Gene.symbol", "Gene.ID", "adj.P.Val", "logFC")) # a table of the columns we actually care about
```

```
write.csv(tT, "Results/tT.csv", row.names = FALSE)
```

#the genes that are expressed more in cancer samples(at least 2 times) their effect is meaningful

```

tT.Up.Gene <- subset(tT, adj.P.Val < 0.05 & logFC > 1)
upGenes = unique(tT.Up.Gene$Gene.symbol)
AML.Up.Genes.AllNames <- unique(as.character(strsplit2(upGenes, "///")))
write.table(AML.Up.Genes.AllNames, "Results/AMLUpGenes.txt", row.names = F,
col.names = F, quote = F)

#the genes that are expressed less in cancer samples(at least 2 times) their
effect is meaningful
tT.Down.Gene <- subset(tT, adj.P.Val < 0.05 & logFC < -1)
downGenes = unique(tT.Down.Gene$Gene.symbol)
AML.Down.Denes.AllNames <- unique(as.character(strsplit2(downGenes, "///")))
write.table(AML.Down.Denes.AllNames, "Results/AMLDownGenes.txt", row.names = F,
col.names = F, quote = F)

```

```

#Extra work
# Multidimensional Scaling
library(magrittr)
library(dplyr)
library(ggpubr)
# Compute MDS
mds <- ex.scale %>%
  dist() %>%
  cmdscale() %>%
  as_tibble()
plot(mds)

```

## References

- [1] "microarray," nature, [Online]. Available: <https://www.nature.com/scitable/definition/microarray-202/>.
- [2] "Genome Expression Omnibus," [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/>.
- [3] "Enrichr," [Online]. Available: <https://maayanlab.cloud/Enrichr/>.
- [4] C. CI, "Series GSE48558," U. of Maryland, Baltimore, [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48558>.
- [5] "Biological pathway," [Online]. Available: [https://en.wikipedia.org/wiki/Biological\\_pathway](https://en.wikipedia.org/wiki/Biological_pathway).
- [6] "Retinoblastoma gene in cancer (Homo sapiens)," [Online]. Available: <https://www.wikipathways.org/index.php/Pathway:WP2446>.
- [7] "Retinoblastoma protein," [Online]. Available: [https://en.wikipedia.org/wiki/Retinoblastoma\\_protein](https://en.wikipedia.org/wiki/Retinoblastoma_protein).
- [8] "Gene Ontology," [Online]. Available: [https://en.wikipedia.org/wiki/Gene\\_Ontology](https://en.wikipedia.org/wiki/Gene_Ontology).
- [9] "GO:0071162," [Online]. Available: <https://www.ebi.ac.uk/QuickGO/term/GO:0071162>.
- [10] "mitotic spindle elongation," [Online]. Available: [http://www.informatics.jax.org/vocab/gene\\_ontology/GO:0000022](http://www.informatics.jax.org/vocab/gene_ontology/GO:0000022).
- [11] D. r. o. binding. [Online]. Available: <https://www.ebi.ac.uk/QuickGO/term/GO:0003688>.