Sharif University of Technology
Department of Computer Engineering

Introduction to bioinformatics course project: Analysis of microarray data for AML

Hadis Ahmadian
400211524

Course Professors:
Dr. Sharifi, Dr. Koohi

February 2022

Table of Contents

1- Instaling and Importing libraries

First we need to install and import libraries that we need for drawing the plots or statistical analysis. I tried to install some packages with command below:

```
installed.packages(c("GEOquery","limma","pheatmap","ggplot2","gplots","reshape2","plyr"))
```

However because my operating system was Windows10 and for the version of R I was using, this command didn't install the packages, so instead, I used commands below:

```
BiocManager::install(c("GEOquery","limma","pheatmap","ggplot2","gplots","reshape2","plyr"))
BiocManager::install(c("pheatmap"),force = TRUE)
BiocManager::install(c("GOexpress"),force = TRUE)
```

To load the libraries in the work space, I used code below:

```
library(GEOquery)
library(limma)
library(umap)
library(pheatmap)
library(GOexpress)
library(ggplot2)
library(reshape2)
library(plyr)
```

Also to set the work directory to the directory that my code was saved in, I used code below:

```
curD <- "F:/uni/bio_prj"
setwd(curD)
```

2- Data preparation

2-1- Downloading the dataset and excluding unwanted samples

The dataset I'm using is from GEO with title "Expression data from normal and Malignant hematopoietic cells" and it is linked here. I downloaded the dataset into my work space, by code below.we can see that the class of gset is a list:

```
series <- "GSE48558" #setting the series of dataset
platform <- "GPL6244" #setting the platform of dataset
gset <- getGEO(series,GSEMatrix = TRUE, AnnotGPL = TRUE , destdir
= "data/") #downloading the dataset and saving it in variable gset
```

gset is a list:

```
>class(gset)
[1] "list"
```

Then we have to check if the dataset we have, consists of more than one platform, and if so, extract the data of our wanted platform, save the wanted platform index in idx, and extract that index of gset:

```
if (length(gset) > 1) idx <- grep(platform, attr(gset, "names"))
else idx <- 1
gset <- gset[[idx]]
```

after the explaind step, gset is now a ExpressionSet:

```
>class(gset)
"[1] ExpressionSet"
attr(,"package")
[1] "Biobase"

>head(gset)
ExpressionSet (storageMode: lockedEnvironment)
assayData: 6 features, 170 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM1180750 GSM1180751 ... GSM1180919 (170 total)
  varLabels: title geo_accession ... phenotype:ch1 (32 total)
  varMetadata: labelDescription
featureData
```

```
  featureNames: 7892501 7892502 ... 7892506 (6 total)
  fvarLabels: ID Gene title ... GO:Component ID (21 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 23836560
Annotation: GPL6244
```

Now we extract the samples that we need for our analysis. Samples with “normal” phenotype as healthy group and samples with “AML Patient” Source name as test group. By using GEO2R grouping tool and using its generated R script, we will have code below. 1 is for test , 0 is for healthy and x is for other groups. We will exclude other groups (marked as x) from:

```
gsms<-
paste0("111111111111XXXXXXXXXXXXXXXXXXXXXXXXXX0XXX0XXXXX",
       "XXXXXXXXXXXXXXXXXX0X0XXX0X0000X0XX00XX00X0X0X0X0X0",
       "XXX0XXX0XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",
       "00000000000000000000")
sml <- strsplit(gsms, split="")[[1]]

# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]

# making group names
gs <- sml[sel]
gs <- factor(gs)
groups <- make.names(c("healthy","test"))
levels(gs) <- groups
```

2-2- Data quality control

There are two main quality controls that we sould do. First we check if our data is log2-scaled and if not, we do it, then we check if our data is normalized and if not, we normalize it.

First we extract the expression matrix out of gset and save it in variable ex, we check the min and max of the matrix to see if it is scaled:

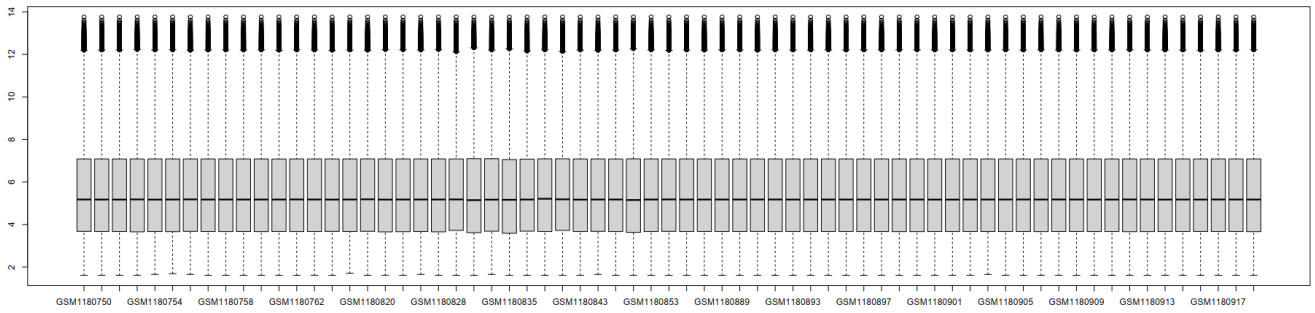
```
ex<-exprs(gset)

>max(ex)
[1] 13.76154
>min(ex)
[1] 1.611473
```

the values of min and max of ex show that data is log2-scaled, however if it was not, we could use code below to scale it:

```
ex <- log2(ex+1)  #adding 1 in order not to have negative values
exprs(gset)<- ex
```

Then we draw a boxplot of data to check if it is normalized and save it in file "boxplot.pdf"

```
pdf("output/boxplot.pdf",width = 15)
boxplot(ex)
dev.off()
```

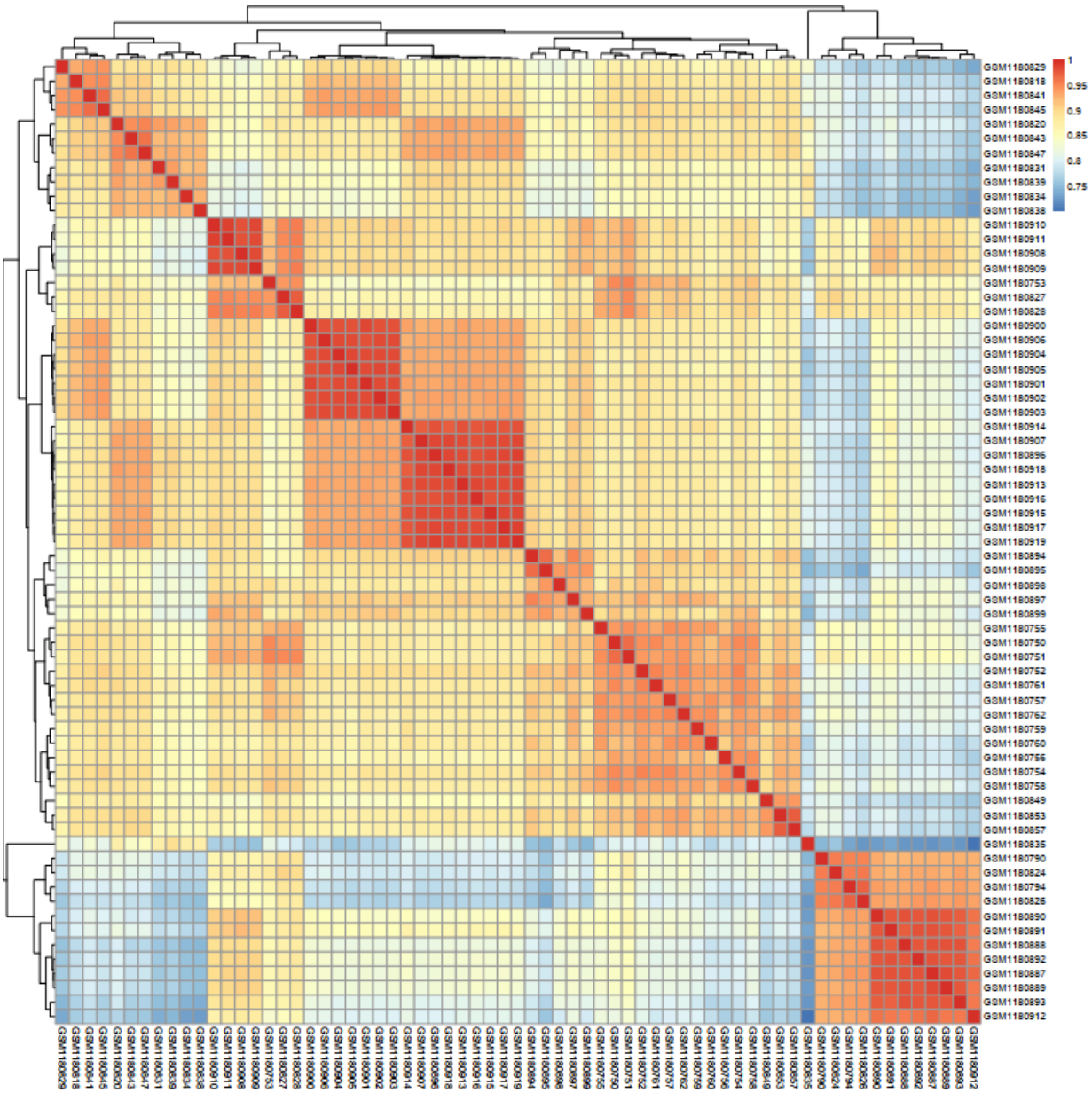


the boxplot shows that data is normalized but if it wasn't, we could mormalized it with code below:

```
ex <- normalizeQuantiles(ex)
exprs(gset)<- gset
```

We can also plot a heatmap to show the correlation between the genes of the expression matrix by code below and save it in file"heat.pdf".

```
Pdf("output/heat.pdf",width = 15,height  = 15)
pheatmap(cor(ex)
dev.off()
```



3- Dimensionality reduction

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.[1]

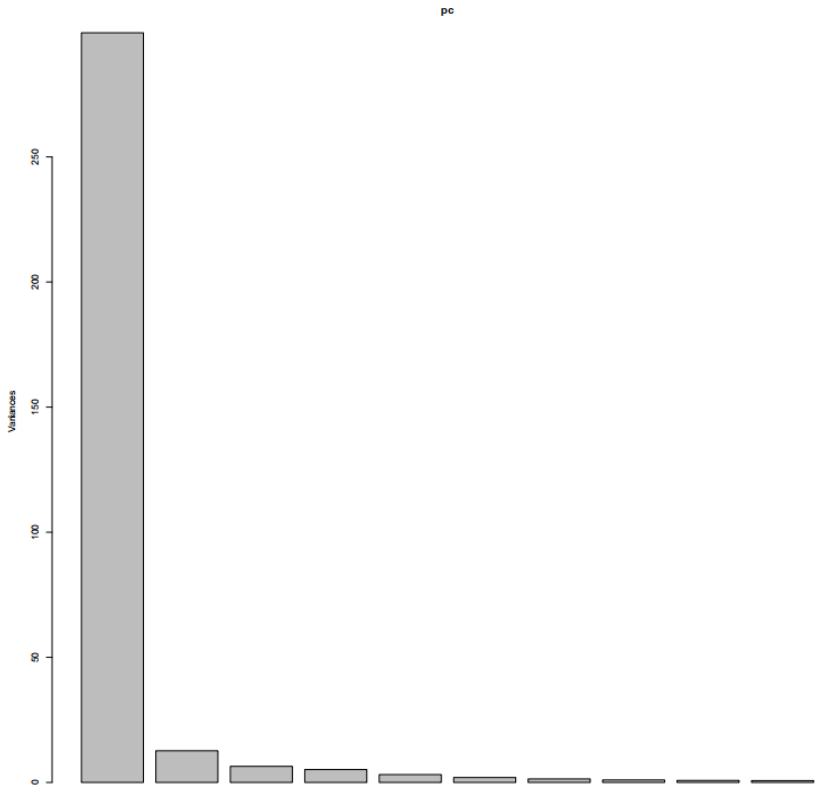
To do this reduction, we have to choose the dimentions of data to keep and dimentions to ignore, so we have to somehow sort the dimentions by their impoitance. The way to do this is to find principal components.

Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance, that is to say, the lines that capture most information of the data. [2]

3-1- Without scalling

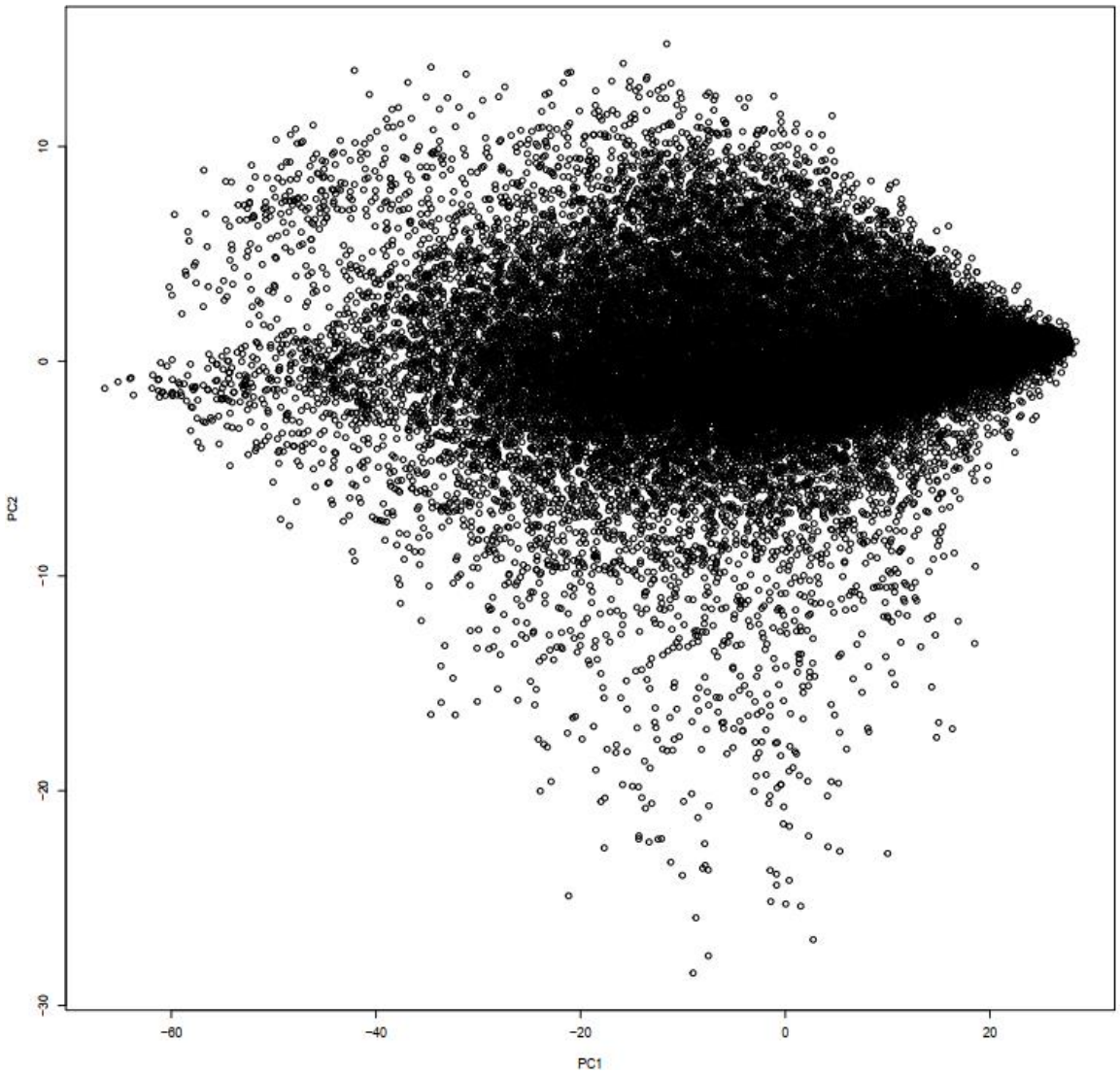
Using code below, we can find the principle components of data and plot them in file “pc_comp.pdf”.

```
Pc <- prcomp(ex)
Pdf(“output/pc_comp.pdf” ,width=15, height=15)
Plot(pc)
dev.off()
```



Now we plot data only by the first two principle components with code below, in file"ex_pc_1_2.pdf".

```
Pdf("output/ex_pc_1_2.pdf" ,width=15, height=15)
Plot(pc$x[,1:2])
dev.off()
```



We can see that there are significant problems in what we did. The first principle component is carrying much more weight than others and when we plot our data with respect to 1st and 2nd component, we are actually sorting the genes with respect to their amount of expression in 1st and 2nd PC. What we really want is finding genes

that have different expression in diffrent samples to find the relation between this difrrence in expression and AML.

3-2- With scalling

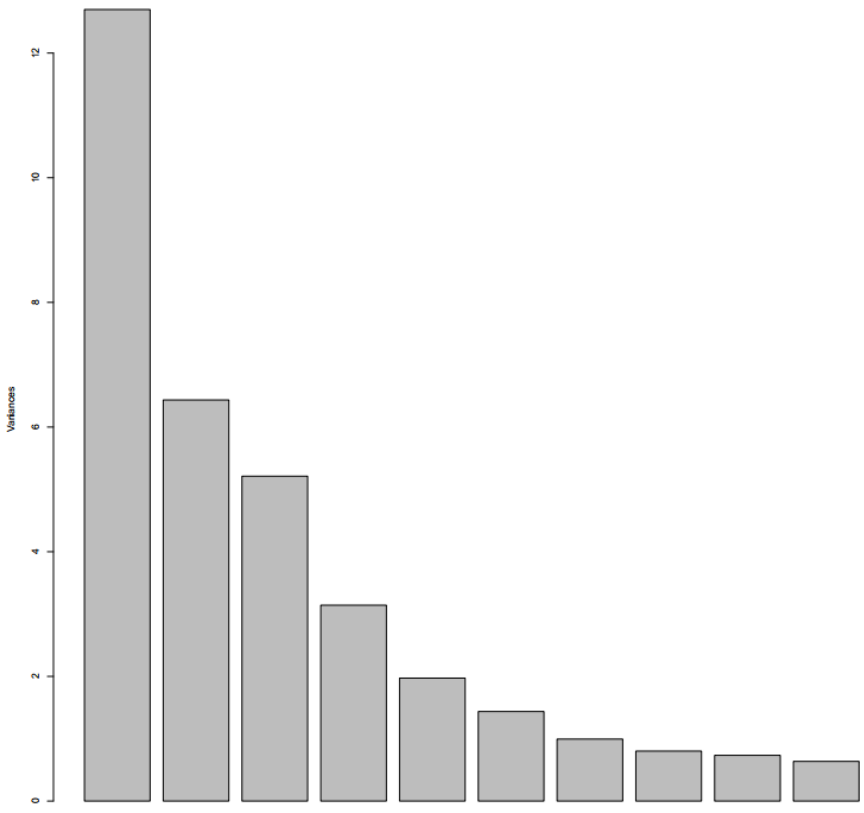
In order to solve those issues, we should remove the effect of genes wich have almost same expression in all samples, to do this, we should substract the average of a gene expression in all samples, from its expression in each sample. This is called scalling and a function does this for us.

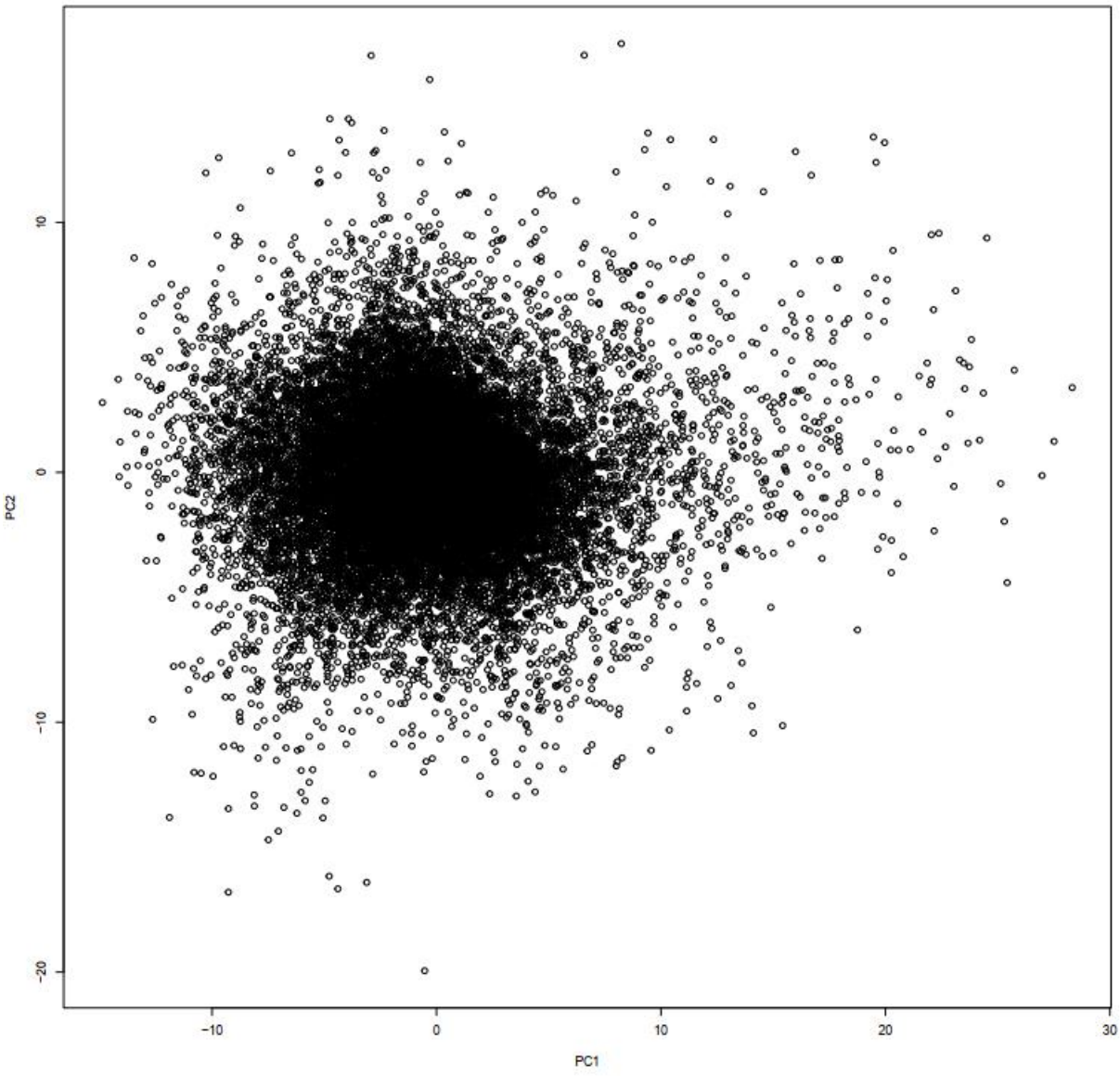
Using code below, we can do scalling and then find principle components of this scaled expression matrix and plot the principle components in file" output/pc_comp_scale.pdf"

```
ex.scale<-t(scale(t(ex),scale=F)) #scale data
pc <- prcomp(ex.scale)
pdf("output/pc_comp_scale.pdf",width = 15,height  = 15)
plot(pc)
dev.off()
```

Now we plot data only by the first two principle components (of scaled matrix) with code below, in file" scale_ex_pc1_2.pdf".

```
pdf("output/scale_ex_pc1_2.pdf",width = 15,height  = 15)
plot(pc$x[,1:2])
dev.off()
```





We can see that after scaling, our pronciple components are more comparabale and the plot of genes have more accaptable distribution (most of the genes are placed around (0,0) meaning that they have same expression in most of samples but as we go right, there are genes that have higher expression in PC1 and as we go up, there are genes that have higher expression in PC2 (than average expression of the gene).

3-3- Dimensionality reduction on genes

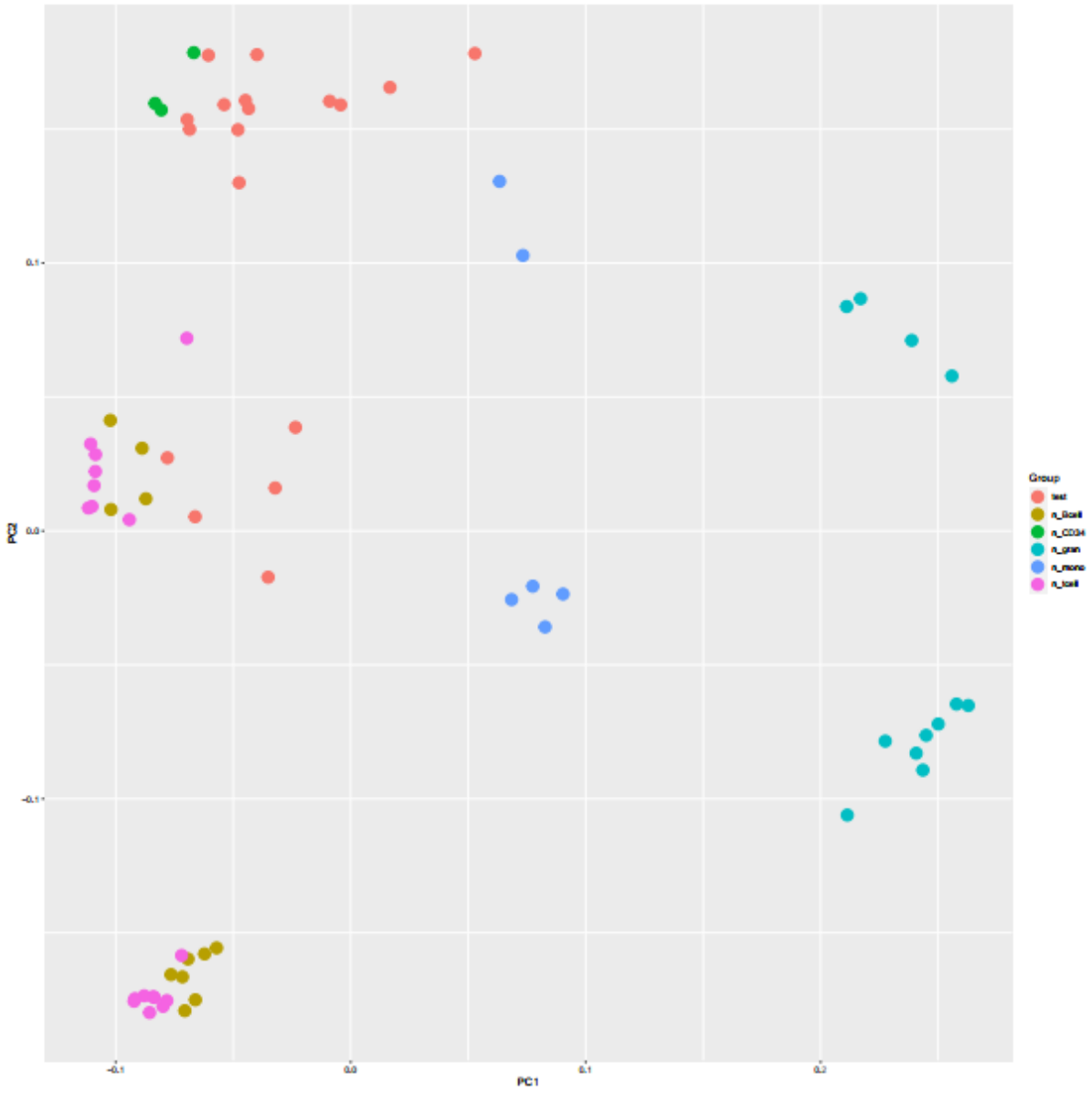
What we did in last parts was reducing some of samples and keep all genes in our dimention reduction just to show the concept of dimensionality reduction and show the nececessity of scalling . For further analaysis we should perform the dimention reduction on genes and keep all samples.

Normal samples have different Source Names, so first we devide the healthy group into 5 groups with respect to their source names, so that we can recognize every group on plot seperatly. We will also use this grouping in next section to check correlations between samples. By using GEO2R grouping tool and using its generated R script, we will have code below.

```
gsms<-
paste0("0000000000000XXXXXXXXXXXXXXXXXXXXXXXXXXX3XXX3XXXXX",
       "XXXXXXXXXXXXXXXXXX1X5XXX3X3441X5XX55XX55X1X5X1X5X2",
       "XXX2XXX2XXXXXXXXXXXXXXXXXXXXXXXXXXX3333333005000",
       "11111115444435555555")
sml <- strsplit(gsms, split="")[[1]]
# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
# making group names
gr <- sml[sel]
gs <- factor(gr)
groups <
make.names(c("test","n_Bcell","n_CD34","n_gran","n_mono","n_tcel
l"))
  levels(gs) <- groups
```

Now we find principle components of genes and then plot samples by 1st and 2nd PC by code below. We save the output in file" PCA_samples.pdf":

```
pcr <- data.frame(pc$rotation[,1:2], Group=gs)
pdf("output/PCA_samples.pdf",width = 15,height  = 15)
ggplot(pcr,aes(PC1,PC2,color=Group))+geom_point(size=5)
dev.off()
```



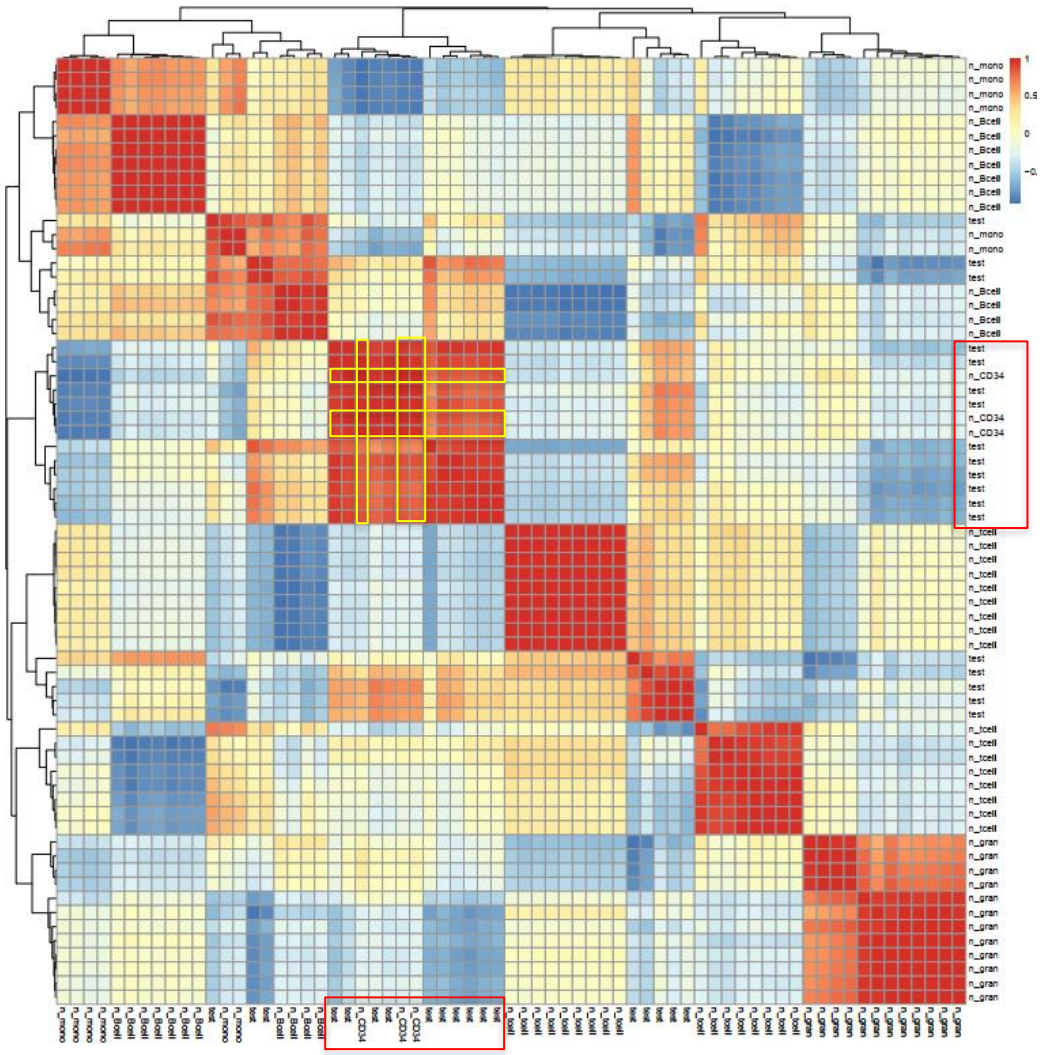
This plot containes 67 samples of 2 groups, AML and healthy(with 5 subgroups) that are ploted with respect to first to principle components(genes). Since almost all of the samples of each group are close together and separate from other groups, we can understand that our data has a good quality and we can identify responsible genes for AML, effectively.

4- Finding the most correlated group to AML

Now that we have proven the quality of our data ,found its principle components and grouped samples by their source name, we can find the most correlated group to AML so that in next sections we analyze gene expression differentiation of AML and those sampels.

In order to find correlation between each group(with respect to 5 first principle components), we use corr function and then draw its heatmap so that we find most correlated group on the plot. For this, we use code below and save output in file ”sample_corr.pdf”.

```
pdf("output/sample_corr.pdf",width = 15,height  = 15)
pheatmap(cor(t(pc$rotation[,1:5])),labels_row = gs, labels_col = gs)  # heatmap of pc$rotation(samples) with 5 first principle components
dev.off()
```



The most correlated samples to AML samples are CD34 and are marked in the heatmap. In next section we will use CD34 samples for gene expression differentiation analysis.

5- Gene expression differentiation analysis

In last section we saw that CD34 samples have the maximum correlation with AML samples, so now we want to analyze their gene expression differentiation. To do this, first we make a design matrix wich tells wich sample belongs to wich group:

```
gs<-factor(gs)
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)
```

```
> head(design)
           test n_Bcell n_CD34 n_gran n_mono n_tcell
GSM1180750    1       0      0      0      0       0
GSM1180751    1       0      0      0      0       0
GSM1180752    1       0      0      0      0       0
GSM1180753    1       0      0      0      0       0
GSM1180754    1       0      0      0      0       0
GSM1180755    1       0      0      0      0       0
```

We needed this matrix for lmfit function wich fits a linear model to data. The use of this linear model is to check how different our samples are(different in gene expression in this case).

```
fit <- lmFit(gset, design)  # fit linear model
```

However we don't want gene expression differentiation in all samples but in AML and CD34 samples so we need to clarify that wich two groups of samples we want to analyze. So in code below, first we clarify that we want the difference between test(AML) and CD34 group (AML – CD34) and we fit a linear model just to analyze the difference between these two groups.

```
# set up contrasts of interest and recalculate model coefficients
cts <- paste("test", "n_CD34", sep="-")
cont.matrix <- makeContrasts(contrasts=cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
```

Now we use the fited model and a bayesian model to calculate p_values. In next step we return the most significant diffrences by fdr(false discovery rate) adjustment and

sort it by B statistic and save them in tT table. From all the info in table, we keep gene symbol, gene Id, adjusted p_value and LogFC.

```
# compute statistics and table of top significant genes
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
tT<-subset(tT,select=c("Gene.symbol", "ID","adj.P.Val","logFC"))
```

This is how tT looks like now:

```
> head(tT)
        Gene.symbol      ID    adj.P.Val     logFC
8160088        MPDZ 8160088 2.565924e-36 -3.670064
8101284       PRKG2 8101284 5.887234e-30 -5.874765
7910915       CHRM3 7910915 3.200429e-28 -4.378570
8160168       FREM1 8160168 1.119567e-23 -3.409825
8121814      NKAIN2 8121814 5.166675e-23 -4.673490
8008588         HLF 8008588 2.786807e-20 -4.140364
```

Then we write the table in file “diff.txt”.

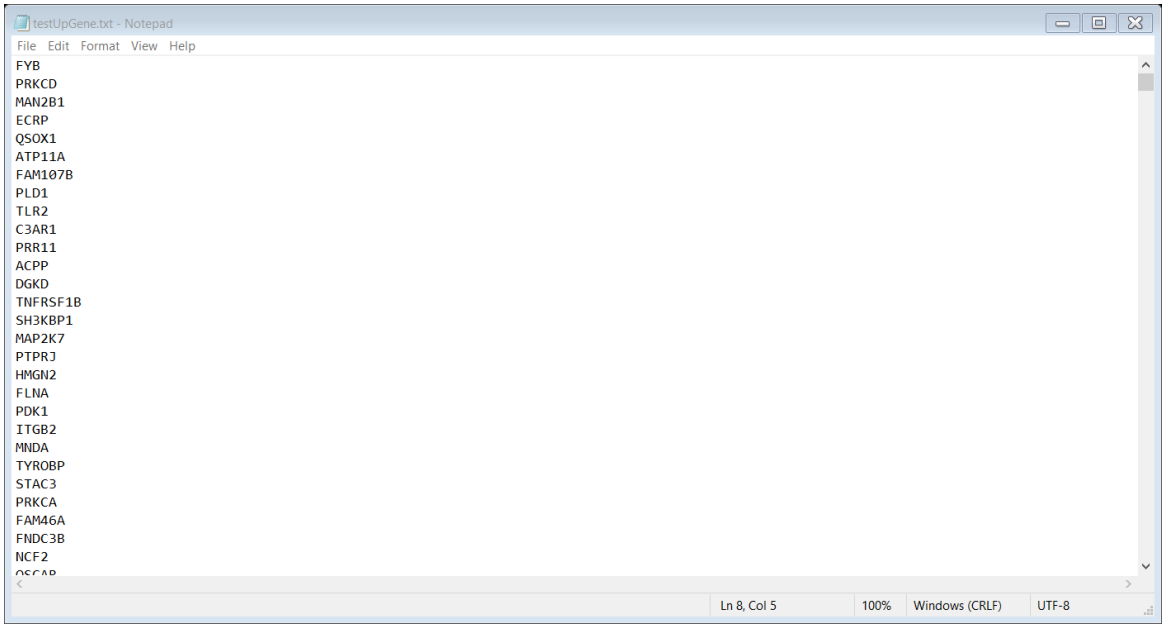
```
write.table(tT,file="output/diff.txt",row.names=F,sep="\t",quote
=F)
```

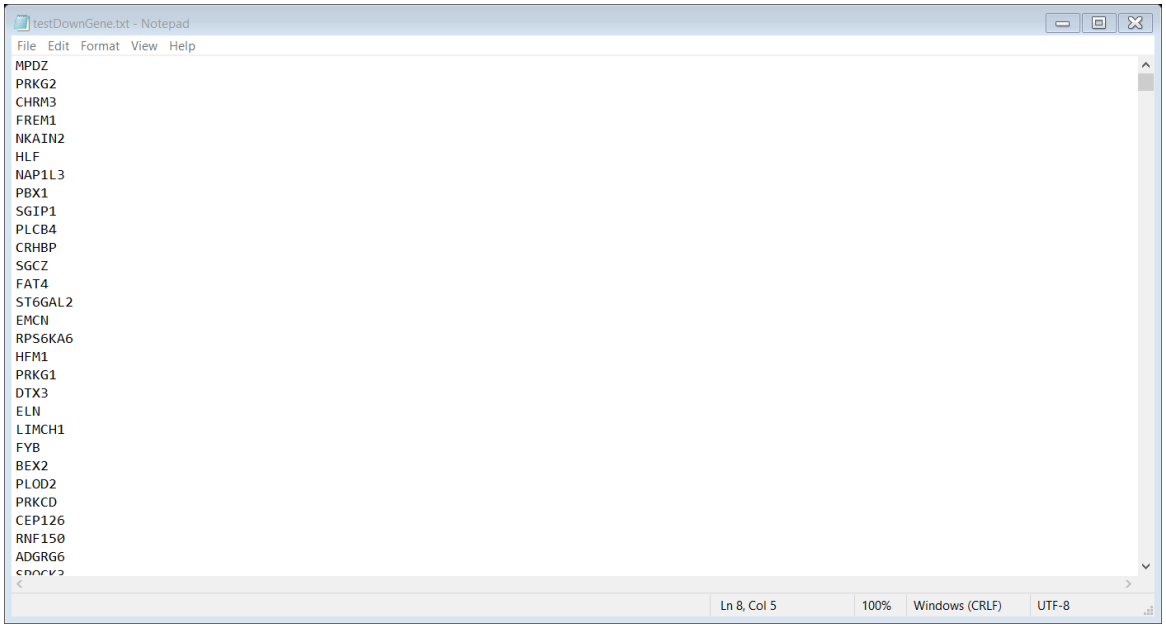
Now we have found genes that are differently expressed between AML and CD34 but we want to be more specific. By code below we will find genes that were significantly higher in AML (LogFC>1) and had a meaningfull difference (adj.p.value>0.05) and save it in file “test.up.gene.txt”

```
test.up <- subset(tT, logFC>1 & adj.P.Val<0.05)
# splits gene names by /// and keeps only unique ones
test.up.gen<-
unique(as.character(strsplit2(test.up$Gene.symbol,"///")))
# writes gene named in file
write.table(test.up.gen, file="output/testUpGene.txt",quote=F,
row.names = F,col.names = F)
```

we can do the same to find significantly lower genes in AML that are meaningfull (LogFC<-1 and adj.p.value>0.05) and save it in file “test.down.gene.txt”

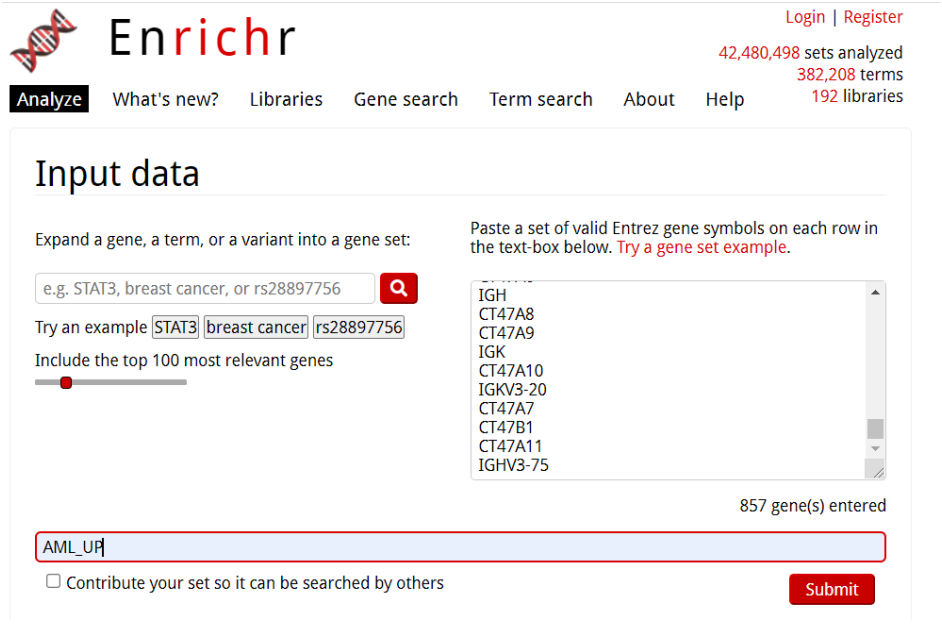
```
test.down <- subset(tT, logFC<-1 & adj.P.Val<0.05)
test.down.gen<-
unique(as.character(strsplit2(test.down$Gene.symbol,"///")))
write.table(test.down.gen,file="output/testDownGene.txt",quote=F
, row.names = F,col.names = F)
```

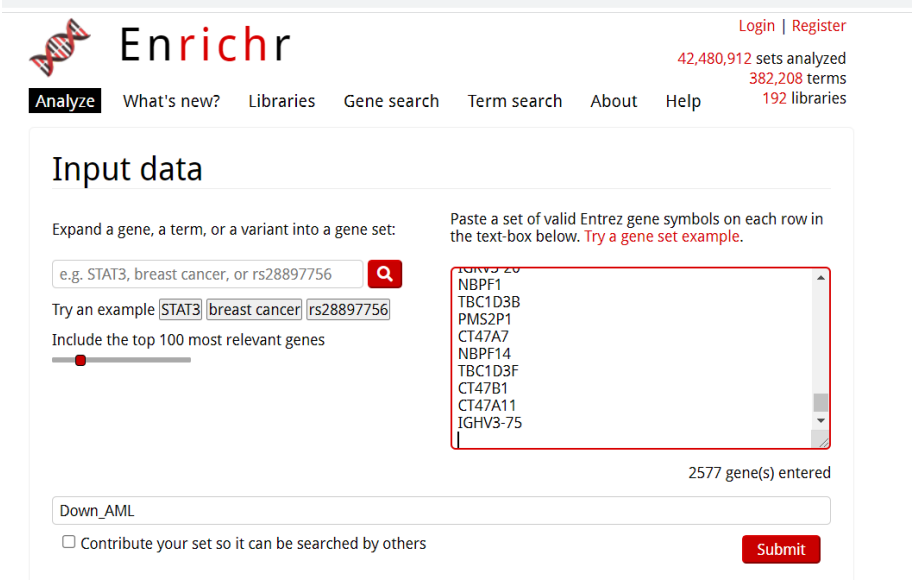




6- Transcription factors, gene antology and pathway analysis

Now that we have found genes that are significantly more or less in AML, we should focous on the effects of these genes and how they cause AML. For this means I used Enrichr.





As you can see there are 857 genes that were higher in AML sampls and 2577 genes that were lower. All transcription factors, pathways and ontologies found by Enrichr are reported in attachments and those who seemed significant or more interesting are going to be expliand more in next section.

6-1- Transcription Factors

UP AML
hsa-miR-215-5p
FOXM1 23109430 ChIP-Seq U2OS Human
E2F4 ENCODE
SPI1 human
JDP2 human tf ARCHS4 coexpression
ZNF750 KD HUMAN GSE38039 CREEDSID GENE 2704 DOWN
SGO1-AS1
RNASE2
SPI1 GM12878 hg19
foxa1 21151129 mcfdash7 lof human gpl10558 gse25315 up
DOWN AML
SPI1 23127762 ChIP-Seq K562 Human
FOXM1 ENCODE
JDP2 human tf ARCHS4 coexpression
MAF SIRNA MACROPHAGE HUMAN GSE98368 RNASEQ DOWN
LINC02705
SPI1
mmu-miR-380-3p
SPI1 GM12878 hg19
ets 00000000 2008 ovarian cancer cells gof human gpl6244 gse21129 up
BRCA1

6-2- Pathways

UP AML
Phagosome
Leptin influence on immune response
Microglia Pathogen Phagocytosis Pathway WP3937
MKNK1 human kinase ARCHS4 coexpression
Neutrophil Degranulation via FPR1/IL8
g2-m checkpoint
Immune System Homo sapiens R-HSA-168256
IL8- and CXCR1-mediated signaling events Homo sapiens f6a58ef3-6193-11e5-8ac5-06603eb7f303
DOWN AML
Phagosome
TYROBP causal network in microglia WP3945
Lysosome
CSF1R human kinase ARCHS4 coexpression
MKNK1 human kinase ARCHS4 coexpression
IRAK3 human kinase ARCHS4 coexpression
Neutrophil Degranulation via FPR1 Signaling
G2-M Checkpoint
Cell Cycle: G1/S Check Point Homo sapiens h g1Pathway
Immune System Homo sapiens R-HSA-168256
Adaptive Immune System Homo sapiens R-HSA-1280218
superpathway of conversion of glucose to acetyl CoA and entry into the TCA cycle Homo sapiens PWY66-407
RAC1 signaling pathway Homo sapiens faffa4fc-6194-11e5-8ac5-06603eb7f303

6-3- Ontology

UP AML
Blood
Immune system
Monocyte
neutrophil activation involved in immune response (GO:0002283)
azurophil granule (GO:0042582)
abnormal neutrophil physiology MP:0002463
Centralspindlin complex
DOWN AML
Blood
Immune system
microtubule motor activity (GO:0003777)
azurophil granule (GO:0042582)
decreased CD4-positive, alpha beta T cell number MP:0008075
Chronic obstructive pulmonary disease (HP:0006510)
Centralspindlin complex

7- Comparing results and biological facts

So far, we have reported genes that are higher or lower in AML paitients and their corresponding transcription factors, pathways and ontologies. Now we will compare some of them with biological facts about AML:

a) Transcription factors:

- **E2F4 ENCODE:** According to [3] *"human clinical data from the Gene Expression Profiling Interactive Analysis (GEPIA) revealed that increased E2F4 expression was associated with poor prognosis in AML patients. Moreover, the experimental results showed that E2F4 was aberrantly overexpressed in human AML patients and cell lines."*

- **FOXM1:** According to [4] *"experiments demonstrated that the viability of primary AML cells, but not normal CD34+ cells, depended on FOXM1 expression. we also provide further support for the role of FOXM1 in chemotherapy resistance, proliferation and stem cell function in AML."*

- **BRCA :** According to [5] *"The homologous DNA repair pathway genes BRCA1 and BRCA2 are classically associated with increased susceptibility to hereditary breast and ovarian cancer due to increased vulnerability to double stranded DNA breaks (mutator phenotype). In addition to their role in breast/ovarian cancer, defects in these genes may predispose to myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML). "*

- **SPI1 :** According to [6] *"The marked heterogeneity of acute myeloid leukemia (AML) renders precisely predicting patient prognosis extremely difficult. Genetic alterations, fusions and mutations, may result in misexpression of key genes in AML. We aimed to investigate the expression patterns of 4 novel genes;* FIS1, SPI1, PDCD7 *and* Ang2 *to determine their potential prognostic role in AML patients."*

b) Pathways:

- **Lysosome:** According to [7] *"Lysosomes promote cancer invasion and metastasis. Invasion and metastasis are the most prominent biological characteristics of malignant cancers and are also the leading causes of death among patients."*

- **G2-M Checkpoint :** According to [8] *"Acute myeloid leukemia (AML) cells exposed to genotoxic agents arrest their cell cycle at the G2/M checkpoint and are inherently chemoresistant."*

- **Phagosome:** According to [9] *"The leukemias belonged to the FAB-M1/M7 (n = 5), M5b (n = 2), and contained phagocytic blasts in various percentages (< 0.2-36.7%). The blasts contained either single or multiple cytoplasmic vacuoles, in some of which the phagosomes were visible. The engulfed hemopoietic cells (red cells, erythroblasts, lymphocytes, and thrombocytes) were also noted in their cytoplasm. These observations confirmed that hemophagocytosis by leukemic blasts is a common and characteristic feature of this type of leukemia."*

- **Cell Cycle: G1/S Check Point Homo sapiens h g1Pathway:** According to [10] *"A recent approach in AML cells enforced this particular G1-checkpoint by stabilizing p53 and the Cdk inhibitor p21 to counter polyploidy through induction of a stable G1-arrest."*

c) Ontology:

We can see that ontologies found by Enrichr (such as **immune system, blood, bone marrow cells**,..) are very related to AML:

According to [11] *"Most often, AML starts in early forms of **white blood cells**.*

*AML is fast growing. The leukemia cells enter the **blood** quickly and sometimes can spread to the liver, spleen, central nervous system (brain and spinal cord), and testicles."*

According to [12] *"Acute myeloid leukemia (AML) is a **blood** cancer that starts in your **bone marrow**. As part of your skeletal system, this spongy tissue inside your bones produces blood cells. The bone marrow produces three types of cell lines:*

White blood cells.

Red blood cells.

Platelets.

White blood cells help your body fight infections. In people with AML, the bone marrow makes abnormal white blood cells. These cancer cells are called myeloid blasts (myeloblasts).

*AML quickly moves from the bone marrow into your **bloodstream** and can even involve other parts of your body. AML may spread to your lymph nodes, brain, liver, cerebral spinal fluid, skin, spleen or testicles as examples."*

References

[1] "Dimensionality reduction," wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Dimensionality_reduction.

[2] "A Step-by-Step Explanation of Principal Component Analysis (PCA)," builtin, [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

[3] "BRCA1 & BRCA2 Germline Variants Are Enriched in MDS/AML and Portend Higher Average Mutational Burden," *Blood,* 2018.

[4] "E2F4 functions as a tumour suppressor in acute myeloid leukaemia via inhibition of the MAPK signalling pathway by binding to EZH2," *Epub,* 2020.

[5] "FOXM1 contributes to treatment failure in acute myeloid leukemia," *DOI,* 2018.

[6] "The clinical and prognostic significance of FIS1, SPI1, PDCD7 and Ang2 expression levels in acute myeloid leukemia," *sciencedirect,* 2019.

[7] "The role of lysosomes in cancer development and progression," *BMC,* 2020.

[8] "G2/M checkpoint stringency is a key parameter in the sensitivity of AML cells to genotoxic stress," *Oncogene,* 2008.

[9] "Hemophagocytosis by leukemic blasts in 7 acute myeloid leukemia cases with t(16;21)(p11;q22): common morphologic characteristics for this type of leukemia," *NIH,* 2000.

[10] "Cell cycle control in acute myeloid leukemia," *American journal of cancer research,* 2012.

[11] "If You Have Acute Myeloid Leukemia (AML)," 2022. [Online]. Available: https://www.cancer.org/cancer/acute-myeloid-leukemia/if-you-have-acute-myeloid-leukemia.html#:~:text=Most%20often%2C%20AML%20starts%20in,spinal%20cord)%2C%20and%20testicles..

[12] "Acute Myeloid Leukemia (AML)," 2022. [Online]. Available: https://my.clevelandclinic.org/health/diseases/6212-acute-myeloid-leukemia-aml#:~:text=AML%20quickly%20moves%20from%20the,spleen%20or%20testicles%20as%20examples..

Attachments

All returned results from Enrichr are available in file “attachments.pdf”