

به نام خدا

نام و نام خانوادگی : سارا رجب زاده

شماره دانشجویی :

گزارش پروژه درس مقدمه ای بر بیوانفورماتیک

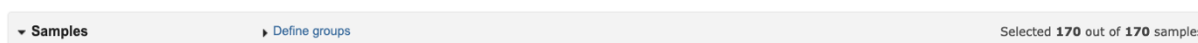
تیر ماه ۱۳۹۹

## مقدمه

ابتدا از سایت GEO داده های مربوط به سرطان خون که مشخص شده بود را می یابیم. سپس در پایین صفحه analyze with GEO2R را کلیک میکنیم ( چون داده microarray می باشد این گزینه فعال است اما اگر داده ای بود که با روش microarray نبود این گزینه وجود نداشت.):



حال باید نمونه ها را گروه بندی کنیم. بنابراین روی گزینه ی define groups کلیک می کنیم:



حال نمونه ها را به دو گروه healthy و AML تقسیم بندی می کنیم و داده های AML patient را در گروه AML و داده های normal را در گروه healthy می گذاریم و سایر آن ها را در گروهی قرار نمی دهیم.

A screenshot of the GEO2R interface. The top part shows a table of samples with columns for GSM ID, Sample Name, and Disease. The bottom part shows the "Quick start" instructions for using GEO2R.

GSM ID	Sample Name	Disease
GSM1180883	Primary T-ALL sample 10-spleen	T ALL Patient
GSM1180884	Primary T-ALL sample 11	T ALL Patient
GSM1180885	Primary T-ALL sample 12	T ALL Patient
GSM1180886	Primary T-ALL sample 13	T ALL Patient
healthy GSM1180887	Normal Granulocyte-5	Granulocytes
healthy GSM1180888	Normal Granulocyte-6	Granulocytes
healthy GSM1180889	Normal Granulocyte-7	Granulocytes
healthy GSM1180890	Normal Granulocyte-8	Granulocytes
healthy GSM1180891	Normal Granulocyte-9	Granulocytes
healthy GSM1180892	Normal Granulocyte-10	Granulocytes
healthy GSM1180893	Normal Granulocyte-11	Granulocytes
AML GSM1180894	Primary AML sample 201	AML Patient
AML GSM1180895	Primary AML sample 201-bone marrow	AML Patient
healthy GSM1180896	Normal T cell-9	T Cells
AML GSM1180897	Primary AML sample 9	AML Patient
AML GSM1180898	Primary AML sample 9-spleen	AML Patient
AML GSM1180899	Primary AML sample 14-bone marrow	AML Patient
healthy GSM1180900	Normal B cell-5	B Cells
healthy GSM1180901	Normal B cell-6	B Cells

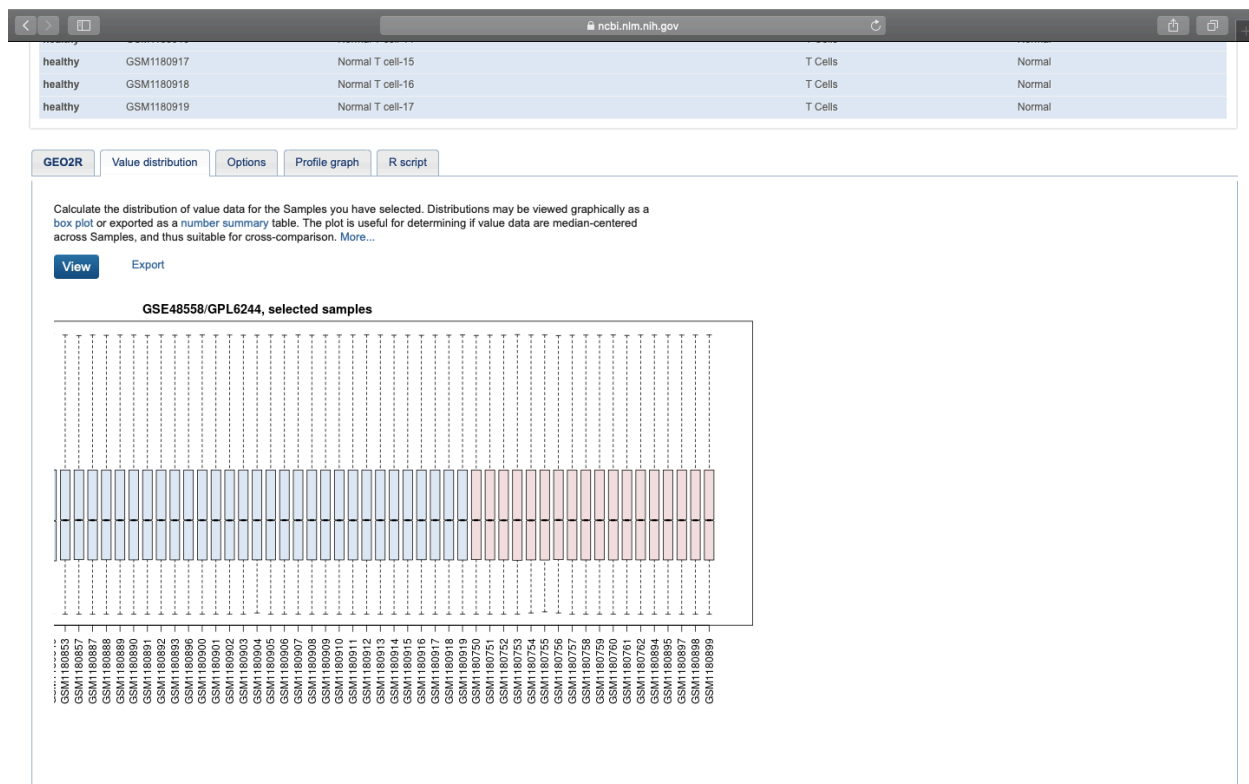
**Quick start**

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong to which group.
- Click 'Top 250' to perform the calculation with default settings.
- Results are presented as a table of genes ordered by significance. The top 250 genes are presented and may be viewed as profile graphs. Alternatively, the complete results table may be saved.
- You may change settings in Options tab.

How to use

Top 250 Save all results

در پایین صفحه روی value distribution کلیک می کنیم و تصویر زیر را مشاهده می کنیم:



همانطور که دیده می شود داده های healthy با رنگ آبی مشخص شده اند و داده های AML با رنگ قرمز. هر نمودار جعبه ای نشان دهنده ی یک sample می باشد و هر نمودار نشان دهنده ی توزیع آن sample است. مطابق این نمودارهای جعبه ای میتوان فهمید که آن ها با یکدیگر قابل مقایسه اند زیرا مقدار overlap آنها با یکدیگر زیاد است و میانه های آن ها نیز خیلی مشابه یکدیگر می باشد بنابراین normalize شده اند.

در پایین صفحه و در GEO2R tab روی Top 250 کلیک می کنیم و سپس صفحه ی زیر مشاهده می شود:

ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE48558

GEO2R Value distribution Options Profile graph R script

Quick start

Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
7989647	3.26e-66	1.01e-70	30.3	150.4	5.68	KIAA0101	KIAA0101
7909568	1.59e-65	9.84e-70	29.9	148.2	4.79	DTL	denticleless E3 ubiqui...
8019842	1.84e-65	1.70e-69	29.8	147.6	5.07	TYMS	thymidylate synthetase
7991406	7.63e-60	9.44e-64	27.1	134.6	3.98	PRC1	protein regulator of cy...
8122202	6.88e-59	1.06e-62	26.6	132.2	4.32	MYB///MYB	MYB proto-oncogene...
8061579	1.91e-57	3.54e-61	26	128.7	4.57	TPX2	TPX2, microtubule nu...
8014974	4.08e-56	8.83e-60	25.3	125.6	5.04	TOP2A	topoisomerase (DNA)...
7937020	5.46e-55	1.35e-58	24.8	122.9	4.39	MKI67	marker of proliferatio...
7982663	1.17e-54	3.25e-58	24.7	122	4.32	BUB1B	BUB1 mitotic checkp...
7966878	1.06e-53	3.37e-57	24.2	119.7	3.63	CIT	citron rho-interacting ...
7945014	1.06e-53	3.61e-57	24.2	119.6	3.98	CHEK1	checkpoint kinase 1
7913869	1.98e-53	7.37e-57	24.1	118.9	4.01	STMN1	stathmin 1
8146357	2.49e-53	1.00e-56	24	118.6	3.33	MCM4	minichromosome mai...
8071212	1.20e-52	5.18e-56	23.7	117	3.65	CDC45	cell division cycle 45
7909708	2.49e-52	1.16e-55	23.6	116.2	4.55	CENPF	centromere protein F
8094278	3.11e-52	1.54e-55	23.6	115.9	4.84	NCAPG	non-SMC condensin I...
8097356	3.73e-52	1.96e-55	23.5	115.7	3.66	PLK4	polo like kinase 4
8132318	2.51e-51	1.40e-54	23.2	113.7	3.98	ANLN	anillin actin binding pr...
8124388	2.81e-51	1.65e-54	23.1	113.5	4.76	HIST1H3B	histone cluster 1, H3b
7919614	4.69e-51	3.00e-54	23	113	3.11	HIST2H3D///HIST2H...	histone cluster 2, H3d...
7929258	4.69e-51	3.05e-54	23	112.9	4.08	KIF11	kinesin family membe...
8046488	2.74e-50	1.86e-53	22.7	111.1	3.88	CDCA7	cell division cycle ass...
		6.07e-53	22.5	110	2.38	IQGAP3	IQ motif containin G...

https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE48558#box\_plot

هر ID نشان دهنده ی یک platform می باشد چون هر probe در هر sample میزان بیان متفاوتی دارد بنابراین میان آن ها t test ای در نظر گرفته می شود تا بتوان فهمید که آیا فرض صفر ( که هر ژن آیا تفاوت بیان معنادار دارد یا خیر) رد می شود یا خیر. بنابراین چون هر t test یک مقدار p value می دهد یک ستون p value در نظر گرفته شده است.

اگر threshold مختص p value را 0.05 در نظر بگیریم هر سطر که p value آن کمتر از 0.05 باشد دارای تفاوت معنادار است و در غیر این صورت نیست. اما بهتر است به جای p value از adj p value استفاده کنیم تا احتمال خطا کاهش پیدا کند. بنابراین ستون adj p value برای این کار مناسبتر است.

اگر threshold را برای بررسی معنادار بودن دیتا ها 0.05 در نظر بگیریم و adj.P.Val را با آن مقایسه کنیم می بینیم که همه ی دیتا ها معنی دار هستند.

ستون t و B نمونه هایی از test statistic هستند اما زیاد اهمیتی ندارند زیرا t محاسبه میشود تا p value حساب شود و p value محاسبه میشود تا adj p value حساب شود.

ستون logFC یا Log fold change به این معناست که میزان بیان چند برابر شده است به این صورت که اگر میزان بیان ۴ برابر شده باشد logFC آن ۲ می باشد. معمولاً logFC و adj p val با هم معیاری برای انتخاب داده های مناسب هستند.

برای ستون logFC نیز باید threshold در نظر بگیریم که داده های مناسبی را انتخاب کنیم که یا بزرگتر از ۱ و یا کوچکتر از -۱ هستند.

از logFC نمیتوان به تنهایی برای بررسی داده ها استفاده کرد زیرا ممکن است واریانس داده ها زیاد باشد بنابراین adj p val نیز باید بررسی شود. ستون Gene.symbol نشان دهنده ی اسم ژن و ستون Gene.title نشان دهنده ی عنوان ژن می باشد. برای شروع زدن کد در محیط R ابتدا باید چند library را صدا بزنیم: ( در کد R با کامنت Libraries مشخص شده است).

```
7 library(Biobase)
8 library(GEOquery)
9 library(limma)
10 library(pheatmap)
11 library(ggplot2)
12 library(reshape2)
13 library(plyr)
```

حال در سایت GEO روی tab R script میزنیم:



کد R ای مشاهده میکنیم. ابتدا کد زیر مشاهده میشود:

```
gset <- getGEO("GSE48558", GSEMatrix=TRUE, AnnotGPL=TRUE)
```

عبارت GSE48558 accession number دیتاست ما می باشد.

حال برای آنکه در هر آنالیز داده کد R کپی و paste نشود accession number را در متغیری جدا قرار می دهیم تا در سری های بعد تنها آن متغیر عوض شود:

```
series <- "GSE48558"
```

عبارت getGEO ای است که در پکیج GEOquery تعریف شده است و دیتای خاص آن accession number را دانلود میکند.

در واقع TRUE بودن GSEMatrix به این معناست که ماتریس بیان ژن ها باید دانلود شود و TRUE بودن AnnotGPL به این معناست که باید Annotation GPL (تفسیر GPL) نیز دانلود شود.

مشکلی که در حال حاضر وجود دارد این است که پس از دانلود، موارد دانلود شده در یک پوشه ی موقت قرار می گیرند و با هر بار اجرای برنامه دیتا ها دوباره دانلود شده و در پوشه ی temp قرار می گیرند. برای جلوگیری از این کار ابتدا باید دستور set working directory اجرا شود:

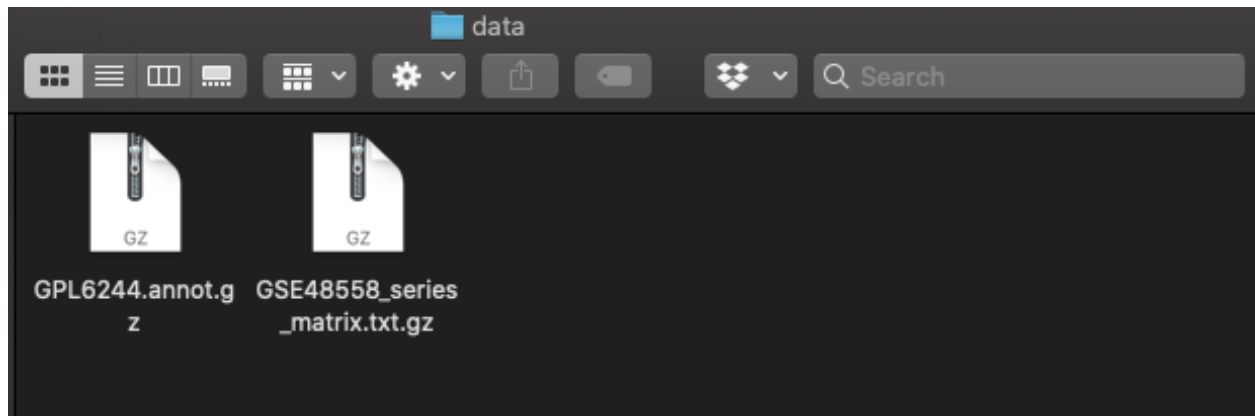
```
setwd("/Users/sara/Desktop/پروژه\بایو")
```

سپس به دستور getGEO اضافه میکنیم تا دیتا های دانلود شده را در پوشه ی data در آن working directory قرار دهد:

```
gset <- getGEO(series, GSEMatrix=TRUE, AnnotGPL=TRUE , destdir = "data/")
```

پس از اجرای این دستورات در محیط RStudio دیتا ها دانلود شده و در پوشه ی دیتا قرار میگیرند:

```
> gset <- getGEO(series, GSEMatrix =TRUE, AnnotGPL=TRUE , destdir = "data/")
Found 1 file(s)
GSE48558_series_matrix.txt.gz
Using locally cached version: data//GSE48558_series_matrix.txt.gz
Parsed with column specification:
cols(
  .default = col_double()
)
See spec(...) for full column specifications.
|=====| 100% 62 MB
Using locally cached version of GPL6244 found here:
data//GPL6244.annot.gz
Warning message:
closing unused connection 3 (https://ftp.ncbi.nlm.nih.gov/geo/series/GSE48nnn/GSE48558/matrix/)
> |
```



بر اساس نمونه هایی که گرفته میشود ممکن است از چند موجود نمونه گرفته شود و بنابراین چند platform داشته باشیم. اما در اینجا میدانیم نمونه های گرفته شده تنها مختص انسان هاست و بنابراین طول دیتا برابر با ۱ می باشد. در نتیجه فقط عضو اول (تنها عضو) را میخواهیم بنابراین داریم:

```
gset <- gset[[1]]
```

مرحله ی بعدی گروه بندی داده هاست. این گروه بندی را در سایت مشخص کردیم و بنابراین در کد R موجود در سایت داریم:

```
gsms <- paste0("1111111111111XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0XXX0XXXXX",  
  "XXXXXXXXXXXXXXXXXXXX0XXXX0X0000X0XX00XX0X0X0X0X0",  
  "XXX0XXX0XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",  
  "00000000000000000000")
```

اما برای بهتر نشان دادن گروه ها میتوان به این صورت عمل کرد که به سایت مراجعه کنیم و ببینیم از هر گروه به چه تعداد وجود دارد :

به طور کلی در میان داده های ما ۳ گروه وجود دارد:

- داده های Normal
- داده های AML
- داده های دیگر (که چون همگی Leukemia هستند با نام Leuk مشخص شده اند).

```
gr <- c(rep("AML" , 13),rep("Leuk" , 27),"Normal",rep("Leuk",3),  
  "Normal",rep("Leuk",23),"Normal","Leuk","Normal",rep("Leuk",3),  
  "Normal","Leuk",rep("Normal",4),"Leuk","Normal",rep("Leuk",2),  
  rep("Normal",2),rep("Leuk",2),rep("Normal",2),"Leuk","Normal",  
  ,"Leuk","Normal","Leuk","Normal","Leuk","Normal","Leuk","Normal",  
  ,rep("Leuk",3),"Normal",  
  rep("Leuk",3),"Normal",rep("Leuk",29),rep("Normal",7),  
  rep("AML",2),"Normal",rep("AML",3),rep("Normal", 20))
```

داده های لوکومیا را حذف می کنیم تا تنها normal ها با AML ها مقایسه شوند:

```
group <- which(gr!="Leuk")  
gr <- gr[group]  
gset <- gset[,group]
```

مرحله ی بعدی به دست آوردن ماتریس بیان از دیتا هاست. برای این کار از دستور `exprs` استفاده می کنیم:

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

با یافتن مقدار ماکسیمم و مینیمم `ex` میتوان به موضوعی پی برد:

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

از مقدار این دو میتوان فهمید که داده ها normalize شده هستند و در scale لگاریتمی هستند.  
حال به کنترل کیفیت می پردازیم:

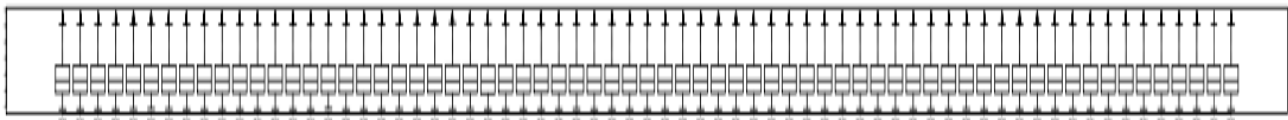
## کنترل کیفیت

برای این کار ابتدا نمودار جعبه ای و یا boxplot میکشیم تا مشاهده کنیم که آیا داده ها normalize شده هستند یا خیر.

نمودار جعبه ای در پوشه ی result موجود می باشد. (boxplot.pdf)  
ابتدا فایل pdf ای میسازیم و سپس boxplot را در آن قرار می دهیم و برای مشخص شدن اسم sample ها width آن را برابر با تعداد sample ها یعنی ۱۷۰ قرار می دهیم.

```
pdf("result/boxplot.pdf",width = 170)
boxplot(ex)
dev.off()
```

نتیجه مطابق شکل زیر می باشد:



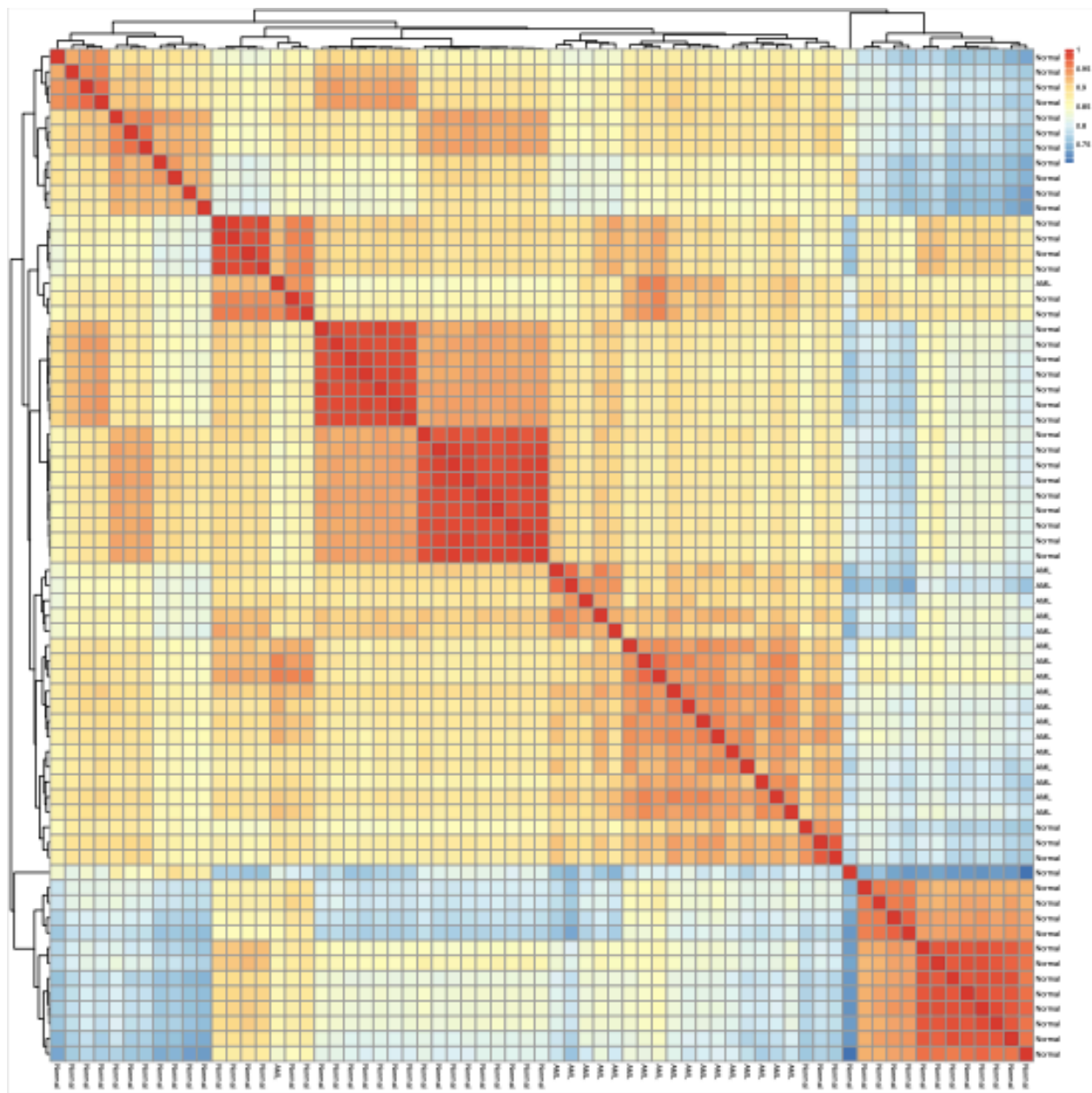
از این نمودار می توان مشاهده کرد که داده ها normalize شده هستند.  
کار دیگری که میتوان در راستای کنترل کیفیت داده ها انجام داد رسم نمودار heatmap بر اساس correlation دو به دوی آن هاست.

این نمودار در پوشه ی result موجود می باشد. (CorHeatmap.pdf)  
برای این کار ابتدا pdf ای می سازیم و سپس نمودار heatmap برای correlation ها را در آن قرار می دهیم و label های موجود در سطر ها و ستون های نمودار را برابر با نام گروه آن می گذاریم:

```
pdf("result/CorHeatmap.pdf",width = 20 , height = 20)
pheatmap(cor(ex), labels_row = gr , labels_col = gr)
dev.off()
```

نتیجه ی این دستورات به شکل زیر است:





رنگ آبی نشان دهنده ی correlation کم و متفاوت بودن است و هر چه به سمت قرمز می رویم correlation افزایش می یابد.

از این نمودار می توان میزان correlation بین هر گروه از داده ها را فهمید.  
دلیل correlation کمتر بین داده های AML با خود آن ها variation بالای سلول سرطانی می باشد.

## PCA

قدم بعدی برای کنترل کیفیت داده ها کشیدن نمودار principal component analysis می باشد.  
برای این کار از دستور prcomp استفاده می کنیم:

```
pc <- prcomp(ex)
```

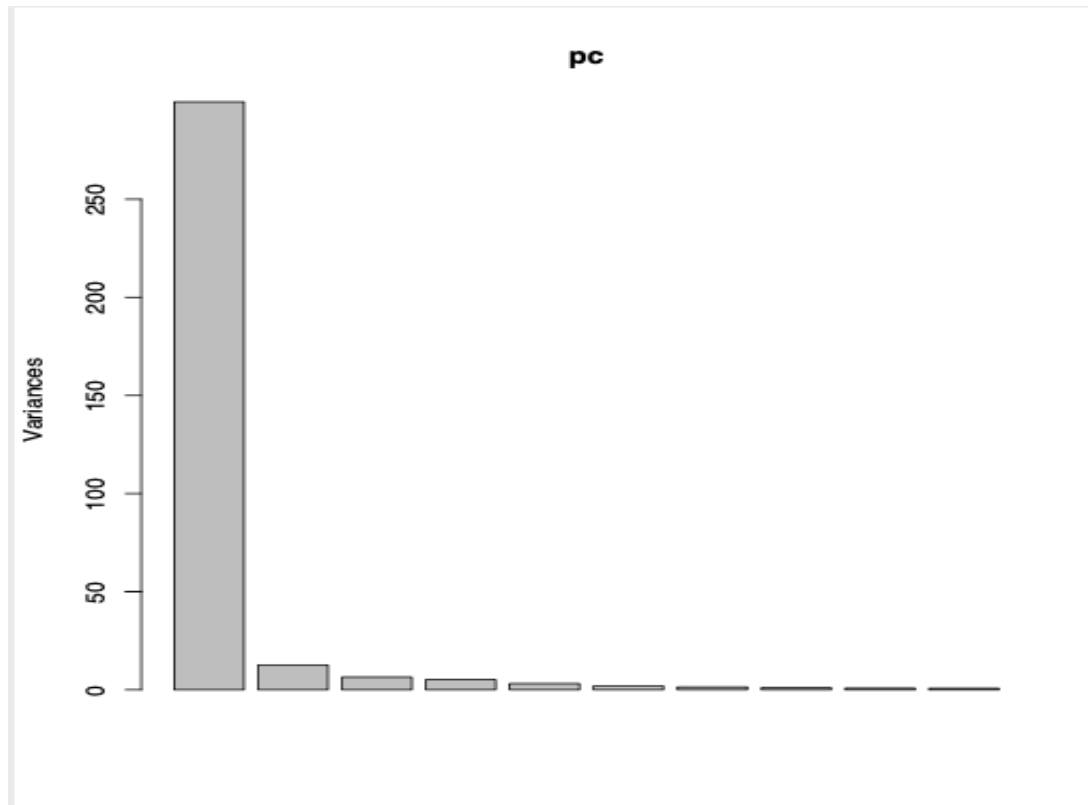
سپس برای کشیدن نمودار آن ابتدا pdf ای می سازیم و سپس نمودار PC را در آن می گذاریم.(نمودار آن در پوشه ی result موجود است.(PC.pdf)

```
pdf("result/PC.pdf")
```

```
plot(pc)
```

```
dev.off()
```

نتیجه ی اجرای این دستورات به شکل زیر می باشد:



این نمودار نشان دهنده ی این است که درجه ی اهمیت PC1 از دیگر PC ها بیشتر است و پس از آن PC2 و ....

درون pc بخش های مختلفی قرار دارد که با اجرای دستور زیر می توان آن را فهمید:

```
names(pc)
```

با اجرای این دستور به خروجی درون شکل می رسیم:

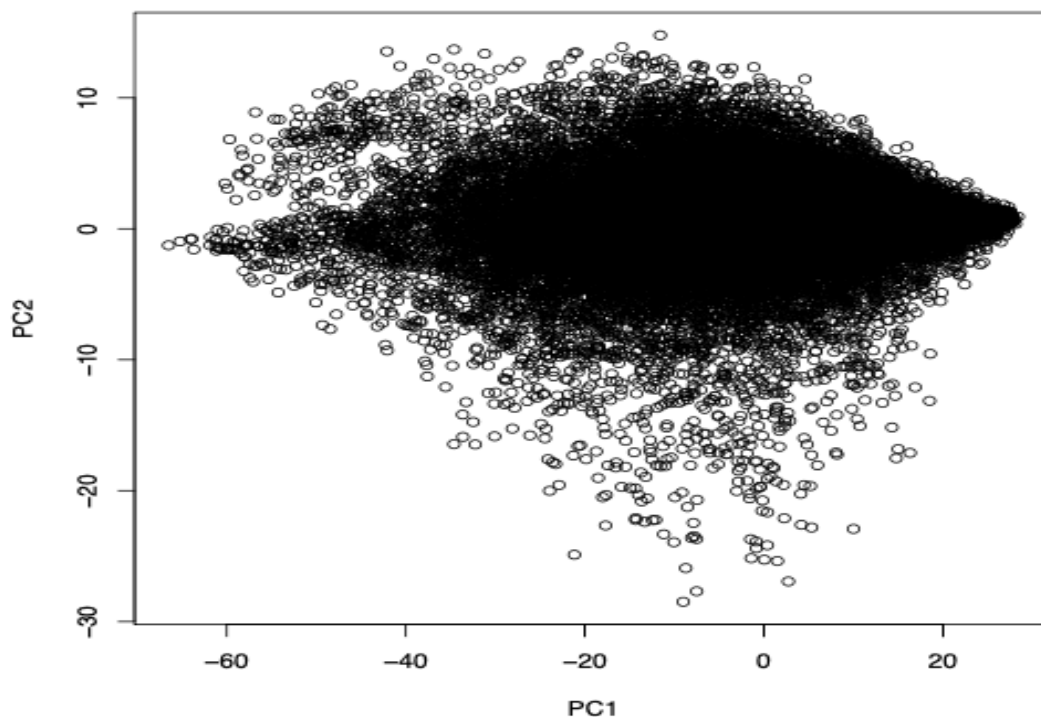
```
> names(pc)
```

```
[1] "sdev"      "rotation" "center"    "scale"     "x"
```

برای نشان دادن ژن ها در فضای PC که فضایی است که ابعاد آن کاهش یافته است باید نمودار x را تحت PC بکشیم بنابراین در همان pdf قبل این کار را می کنیم:

```
pc <- prcomp(ex)
pdf("result/PC.pdf")
plot(pc)
plot(pc$x[,1:2])
dev.off()
```

در واقع  $x$  ماتریس ژن ها در فضایی است که کاهش ابعاد یافته است.  
به جای آن که همه ی ۱۷۰ تا را نمایش دهیم تنها دو ستون اول که از اهمیت بیشتری برخوردارند را می کشیم.  
پس از اجرای آن شاهد نمودار زیر خواهیم بود:



هر کدام از نقاط نشان دهنده ی یک ژن است و این نمودار به شکلی است که بیشترین variation را در راستای PC1 می بینیم.  
مشکلی که در این نمودار وجود دارد این است که گاهی ژن هایی وجود دارد که همیشه میزان بیان آنها صفر می باشد و گاهی ژن هایی وجود دارد که همیشه میزان بیان بالایی دارند بنابراین این نمودار نشان دهنده ی اطلاعات جدید نمی باشد.

برای آن که نمودار را طوری رسم کنیم که تنها تفاوت های ژن ها مشخص باشد باید میانگین همه ی ژن ها را صفر کنیم. به این صورت که میزان بیان هر ژن را منهای میانگین بیان آن ژن کنیم. به بیانی دیگر می خواهیم داده ها را scale کنیم.

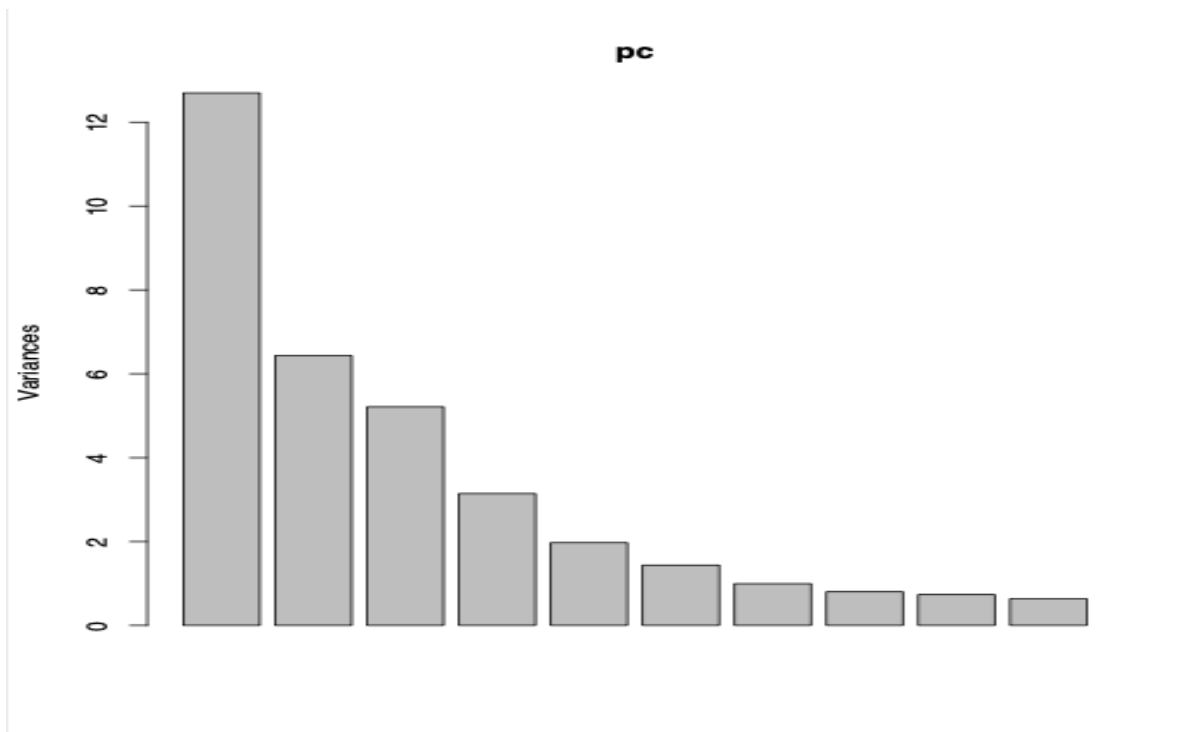
برای این کار از دستور scale استفاده می کنیم. دستور scale در طول همه ی ستون ها داده ها را scale می کند اما ژن های ما در سطر موجود است. برای همین ابتدا آن را transpose میکنیم (با دستور t()) تا در ستون قرار بگیرند. سپس آن را scale کرده و برای برگرداندن آن به حالت اول دوباره آن را transpose می کنیم و با False کردن scale مطمئن می شویم که تنها کاری که تابع scale انجام می دهد center کردن مقادیر است و کاری به بزرگی مقادیر ندارد. دستور آن در زیر مشاهده می شود:

```
ex.scale <- t(scale(t(ex), scale = F))
```

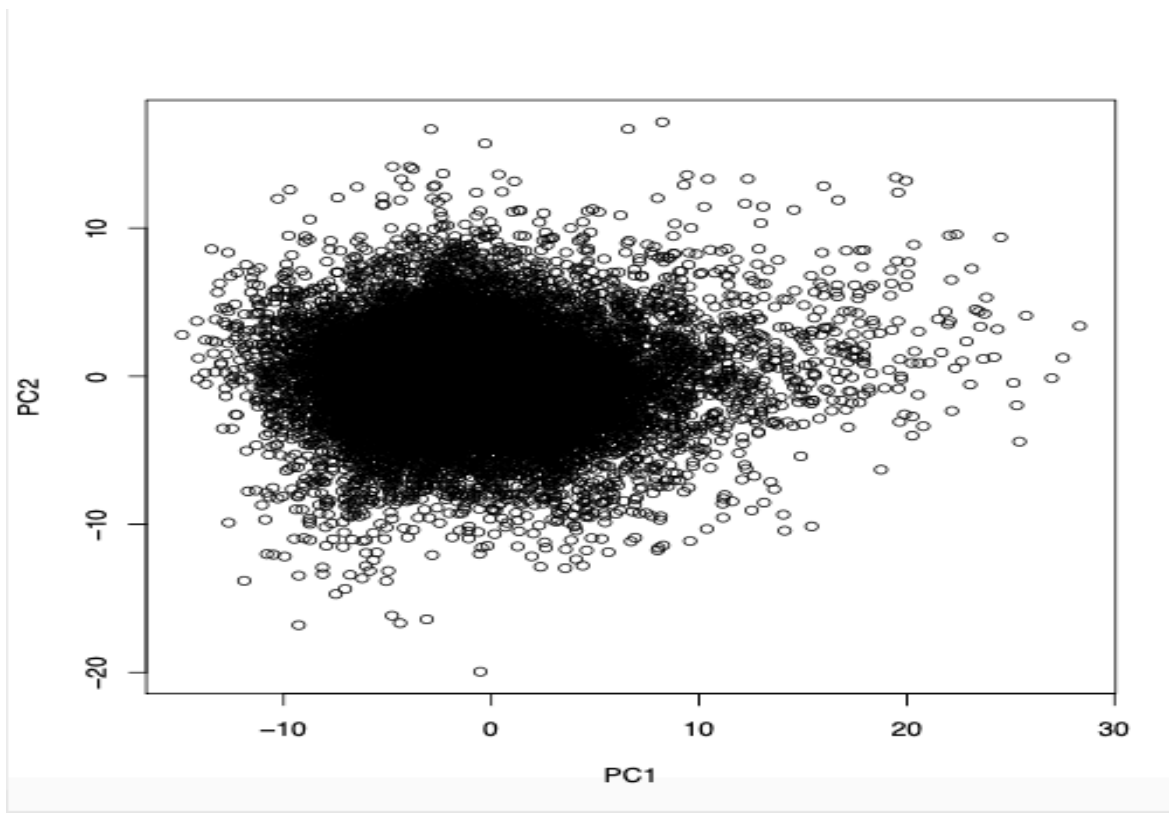
حال دستورات موجود در بالا برای محاسبه ی principal component را این بار برای داده های scale شده اجرا می کنیم:

```
pc <- prcomp(ex.scale)
pdf("result/PC_scaled.pdf")
plot(pc)
plot(pc$x[,1:2])
dev.off()
```

نتیجه ی این دستورات دو نمودار موجود در پوشه ی result و در PC\_scaled.pdf می باشد که به شکل زیر است:



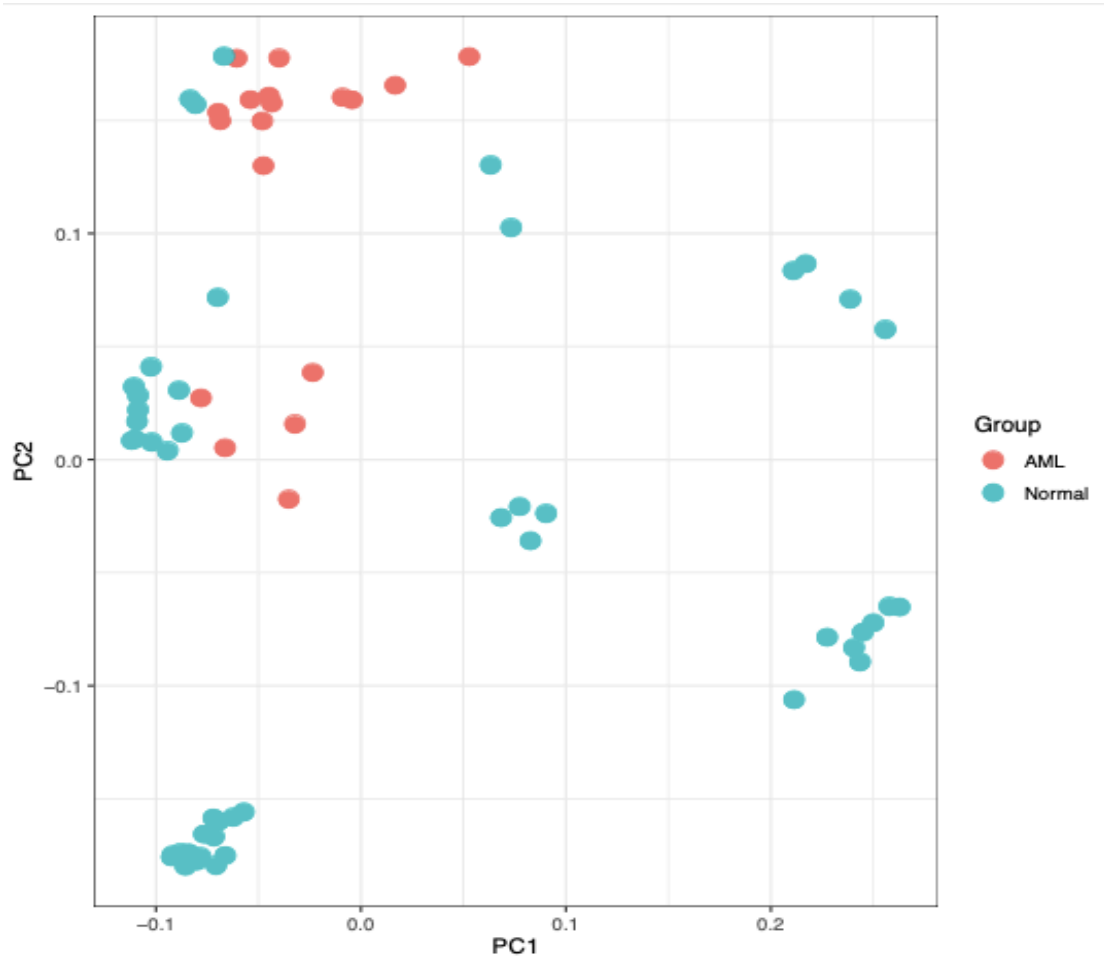
مطابق نمودار بالا متوجه می شویم که این طور نیست که PC1 همه کاره باشد و بقیه هیچ کاره. آن ها با یکدیگر قابل مقایسه تر شده اند.



مطابق نمودار بالا می فهمیم که ژن ها توزیع منطقی تری پیدا کرده اند. حال PC هر نمونه را رسم می کنیم. هر sample در pc\$rotation قرار دارد. برای این کار ابتدا sample ها را به صورت دیتا فریم ذخیره می کنیم. (زیرا در دیتا فریم می توان کلاس های مختلفی داشت) در این کار به همه ی principal component ها احتیاج نداریم و نهایتاً نموداری ۳ بعدی می کشیم. (گاهی ابعاد بدین صورت اتفاق می افتد). کشیدن نمودار ها را بر اساس گروه بندی ای که داریم مشخص می کنیم. برای رسم نمودار آن از کتابخانه ی ggplot2 استفاده می کنیم. ابتدا pdf ای می سازیم و درون آن نمودار را قرار می دهیم. (این نمودار در پوشه ی result به نام PCA\_samples.pdf موجود است.) رنگ بندی نمودار را بر اساس گروه آن مشخص می کنیم. سایز نقطه ها را با دستور size بزرگ می کنیم و تم آن را bw می گذاریم. دستورات R به صورت زیر می باشد:

```
pcr <- data.frame(pc$rotation[,1:3] , Group = gr)
pdf("result/PCA_Samples.pdf")
ggplot(pcr,aes(PC1,PC2 , color = Group))+geom_point(size = 4) + theme_bw()
dev.off()
```

با اجرای این دستورات به شکل زیر می رسم:



از این نمودار نتیجه می گیریم که داده های AML با normal در راستای PC1 خوب جدا شده اند.

## کاهش ابعاد

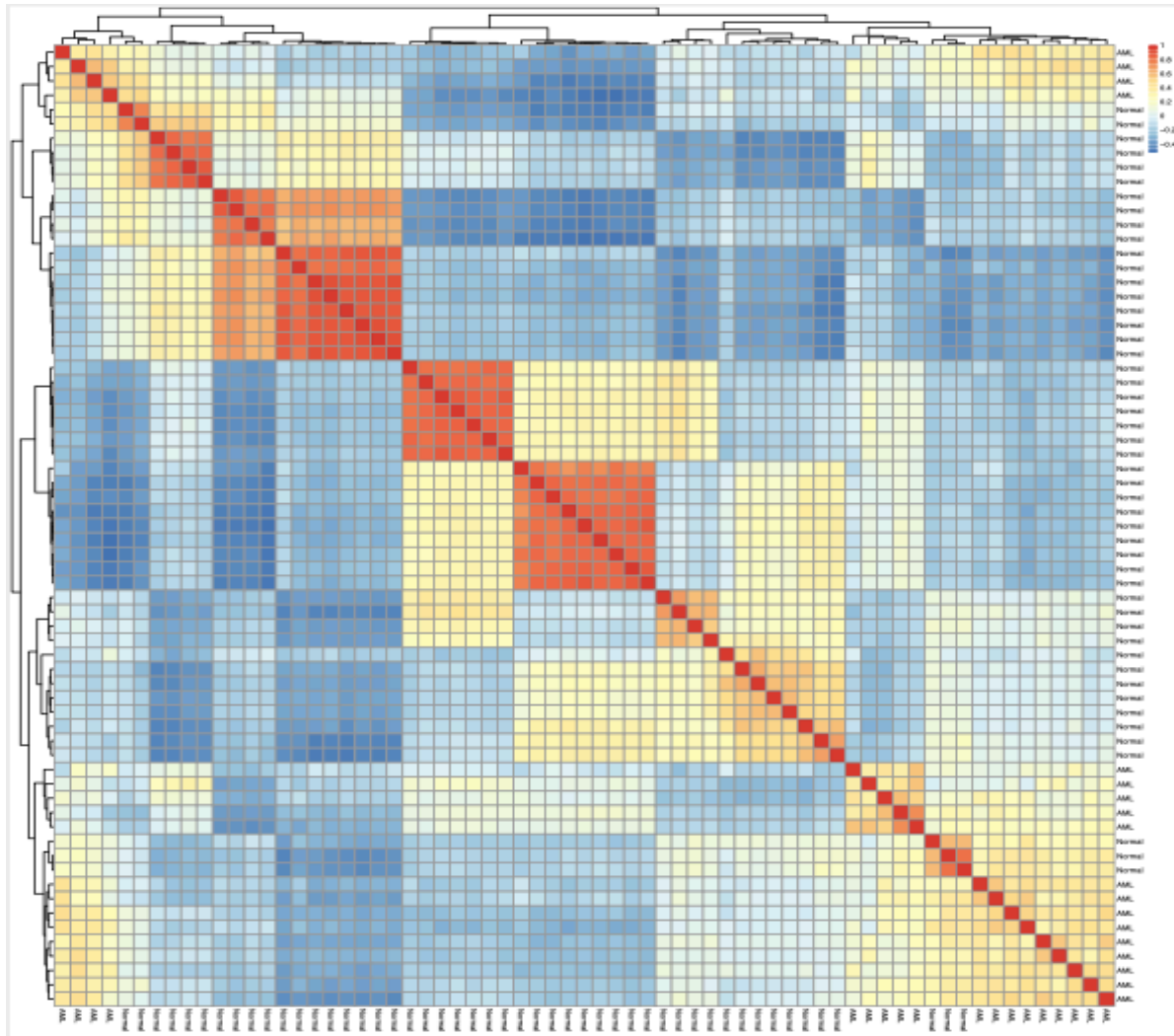
برای این بخش دوباره از PCA استفاده می کنیم که در بخش کنترل کیفیت درباره ی آن مفصل بحث شد. بیشتر مراحل کاهش ابعاد در قسمت scale کردن داده ها پیش می رود ولی به طور کلی تمام موضوعات PCA بحث شده در کنترل کیفیت در کاهش ابعاد مؤثر است.

## بررسی همبستگی

برای بررسی همبستگی داده ها از correlation استفاده می کنیم و نمودار مربوط به آن را به صورت heatmap رسم می کنیم:

```
pdf("result/heatmap_cor.pdf", width = 20 , height = 20)
pheatmap(cor(ex.scale), labels_row = gr , labels_col = gr)
dev.off()
```

نتیجه ی این دستورات به شکل یک pdf در پوشه ی result موجود است. (heatmap\_cor.pdf)  
نتیجه به صورت نمودار زیر می باشد:



از این نمودار می توان نتیجه گرفت که همبستگی داده های نرمال با نرمال زیاد است اما همبستگی داده های AML با AML چندان زیاد نیست به این دلیل که همان طور که در بخش کنترل کیفیت توضیح داده شد میزان variation در درون یک سلول سرطانی بسیار بالاست. همان طور که مشاهده می شود میزان همبستگی داده های AML با نرمال کم است که نشان می دهد این داده ها با یکدیگر متفاوت شده اند.

## بررسی تمایز بین ژن ها

برای یافتن تمایز میان ژن ها ابتدا باید تکه ای از کد R script موجود در سایت GEO را به R studio منتقل کنیم و بر روی آن تغییراتی اعمال کنیم:

```

1. gr <- factor(gr)
2. gset$description <- gr
3. design <- model.matrix(~ description + 0, gset)
4. colnames(design) <- levels(gr)
5. fit <- lmFit(gset, design)
6. cont.matrix <- makeContrasts(AML-Normal, levels=design)
7. fit2 <- contrasts.fit(fit, cont.matrix)
8. fit2 <- eBayes(fit2, 0.01)
9. tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
10. tT <- subset(tT, select=c("Gene.symbol", "Gene.ID", "adj.P.Val", "logFC"))
11. write.table(tT, "result/AML_Normal.txt", row.names=F, sep="\t", quote = F)

```

خط ۲ توصیف هر sample می باشد که در واقع گروهی است که تعریف کرده ایم که این گروه ها به صورت text است و description به صورت factor می باشد.

ابتدا گروه را به صورت factor ذخیره می کنیم (خط ۱) سپس توصیف یا description را همان گروه قرار می دهیم (خط ۲). تا خط چهارم مختص ساختن design می باشد. ماتریس design به این صورت است که به ازای هر نمونه یک سطر و به ازای هر گروه یک ستون دارد و با صفر و یک مشخص می کند که هر sample عضو کدام گروه است:

	AML	Normal
GSM1180750	1	0
GSM1180751	1	0
GSM1180752	1	0
GSM1180753	1	0
GSM1180754	1	0
GSM1180755	1	0

خط ۵ به این صورت است که بر اساس design ساخته شده linear model ای به دیتا fit می کند (یعنی خطی به آن fit می شود که نقاط به بهترین وجه قرار بگیرند و سپس می توان از شیب آن خط متوجه شد که تفاوت بیان وجود داشته است یا خیر. هر چه شیب خط بیشتر باشد تفاوت بیان بیشتر است و هر چه نقاط به خط fit تر باشند معنادارتر است.)

در خط ۶ باید تعریف کنیم که چه چیزهایی را می خواهیم با یکدیگر مقایسه کنیم و چون می خواهیم AML را با normal ها مقایسه کنیم AML-Normal گذاشته ایم.

در خط ۷ شیبی محاسبه می کنیم که بر اساس تفاوتی که در خط ۶ بیان شد باشد.



در خط ۸ مدل بیز را برای محاسبه ی p value ها بر آن اعمال می کنیم.  
 در خط ۹ از fit2 به دست آمده از خط ۸ significant ترین تفاوت ها را بر می گردانند.(در adjust هر  
 روشی می تواند باشد. بر اساس آماره ی B مرتب شده اند و چون همه ی آن ها را می خواهیم تعداد یا  
 number آن را بی نهایت قرار داده ایم).  
 در خط ۱۰ از جدول tT یک subset میگیریم و تنها آنهایی که لازممان می شوند را نگه می داریم.  
 و در نهایت در خط ۱۱ نتیجه در فایل text ای در پوشه ی result قرار گرفته  
 است.(AML\_Normal.txt)  
 بخشی از این فایل text در ادامه آورده شده است:

Gene.symbol	Gene.ID	adj.P.Val	logFC
MPO	4353	3.61781344674199e-19	5.56350115651021
FLT3	2322	4.83571557426423e-19	5.2500645271644
KIAA0101	9768	6.30816005381361e-19	4.55913523978118
BUB1B	701	1.66404320184616e-18	2.75655355297846
SUCNR1	56670	1.93857268145515e-18	2.99681551599093
MCM10	55388	3.71213666952313e-18	2.31884765424603
TPX2	22974	4.69552919298598e-18	3.15641491614172
CIT	11113	1.14794610995488e-17	2.37075070203741
CDC45	8318	1.65866463412725e-17	2.28750067397506
IQGAP3	128239	1.77553965334511e-17	1.66969735012925
POLQ	10721	1.77553965334511e-17	2.09197758109184
CPXM1	56265	6.59236540681646e-17	3.77695351746145
STK38	11329	7.22846118132058e-17	-1.8804326870068
ANLN	54443	7.50468496299332e-17	2.64104618221088
PRC1	9055	7.50468496299332e-17	3.0800967854263
MELK	9833	1.35263969928417e-16	2.29731792385714
TOP2A	7153	1.7319446378169e-16	3.29810352836282
NCAPG	64151	2.37600560889478e-16	3.22796342043764
CBX7	23492	4.52689592358881e-16	-2.24007964966893
KIF23	9493	4.59347924168233e-16	2.82172753519388
TYMS	7298	5.46160241575207e-16	3.67035228143311
KIT	3815	7.79374037619323e-16	4.86400851021202
ECRP///ECRP	643332///643332	7.88557587748143e-16	4.55318219877211
DTL	51514	7.98952126104647e-16	3.67921774278912
PLCL2	23228	7.98952126104647e-16	-1.89971809574036
CDT1	81620	8.59049950225188e-16	1.71491252610091
KIF14	9928	1.19628115231684e-15	2.19593366886508
PECR	55825	1.65214459104176e-15	-2.16251306721542
DIAPH3	81624	2.07208874427226e-15	1.91092967643311
NRG4	145957	2.33222629471214e-15	3.4068875179161
CDC25A	993	2.95877390100045e-15	1.79713843244785
CENPI	2491	3.39242148650163e-15	2.24023758153968
MRC1	4360	3.39242148650163e-15	3.04859703215193

حال از tT به دست آمده می توان ژن هایی را که بالاترین تفاوت بیان را داشته اند پیدا کرد به این صورت که  
 آنهایی را که logFC آن ها بیشتر از یک می باشد(به این معنا که حداقل دو برابر شده اند) و آن هایی که  
 adj p val آن ها بیشتر از میزان threshold ای است که در نظر گرفته شده (0.05) را انتخاب کنیم:

```
aml.up<- subset(tT, logFC >1 & adj.P.Val<0.05)
```

پس از آن باید ژن های تکراری حذف شوند و نتایج آن را در فایل text ای در پوشه ی result ذخیره می کنیم و همین کار ها را برای ژن های down شده هم انجام می دهیم. علاوه بر آن برای حذف نام سطر و ستون rownames و colnames را false می کنیم. علاوه بر آن ژن هایی موجود هستند که چند اسم دارند و آن اسم ها با "///" از یکدیگر جدا شده اند بنابراین از دستوراتی در R استفاده کرده ایم که تمام اسمی ژن ها پشت سر هم نمایش داده شوند:

```
aml.up<- subset(tT, logFC >1 & adj.P.Val<0.05)
```

```
aml.up.Gene <- unique(as.character(strsplit2(aml.up$Gene.symbol , "///")))
```

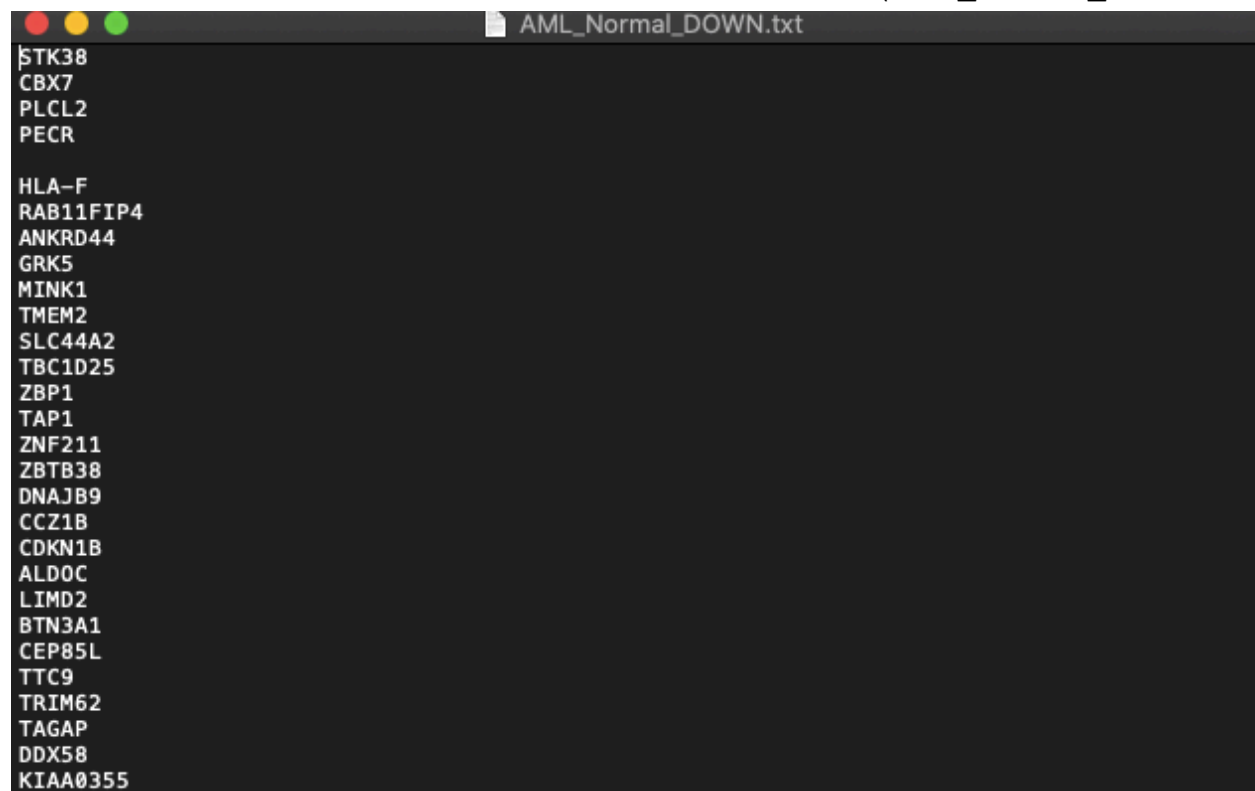
```
write.table(aml.up.Gene , file = "result/AML_Normal_UP.txt" , quote = F,  
row.names = F, col.names = F)
```

```
aml.down<- subset(tT, logFC < -1 & adj.P.Val<0.05)
```

```
aml.down.Gene <- unique(as.character(strsplit2(aml.down$Gene.symbol,"///")))
```

```
write.table(aml.down.Gene , file = "result/AML_Normal_DOWN.txt" , quote = F,  
row.names = F, col.names = F)
```


نتایج این دستورات در دو فایل text در پوشه ی result موجود است.(AML\_Normal\_UP.txt و  
(AML\_Normal\_DOWN.txt



The screenshot shows a text editor window titled "AML\_Normal\_DOWN.txt". The content is a list of gene symbols, one per line. The symbols are: STK38, CBX7, PLCL2, PECR, HLA-F, RAB11FIP4, ANKRD44, GRK5, MINK1, TMEM2, SLC44A2, TBC1D25, ZBP1, TAP1, ZNF211, ZBTB38, DNAJB9, CCZ1B, CDKN1B, ALDOC, LIMD2, BTN3A1, CEP85L, TTC9, TRIM62, TAGAP, DDX58, and KIAA0355.

```
AML_Normal_UP.txt
MPO
FLT3
KIAA0101
BUB1B
SUCNR1
MCM10
TPX2
CIT
CDC45
IQGAP3
POLQ
CPXM1
ANLN
PRC1
MELK
TOP2A
NCAPG
KIF23
TYMS
KIT
ECRP
DTL
CDT1
KIF14
DIAPH3
NRG4
CDC25A
CENPI
MRC1
```

حالا از این ها برای gene ontology استفاده می کنیم:  
برای مثال از aml.UP.Genes استفاده می کنیم. وارد سایت enrichr می شویم و درون box موجود نام  
ژن ها را قرار می دهیم و در قسمت description می نویسیم AML Up Genes و سابمیت می کنیم:

**Enrichr**

26,819,833 lists analyzed  
335,639 terms  
167 libraries

Login | Register

Analyze | What's new? | Libraries | Gene search | Term search | About | Help

### Input data

Choose an input file to upload. Either in BED format or a list of genes.  
Try an example [BED file](#).  

Choose File

No file chosen

Paste a list of valid Entrez gene symbols on each row in the text-box below. Try a [gene set example](#).

WDR36  
CTSW  
MAFB  
PYGL  
RPL27A  
RIMKLB  
MCM8  
MRPS27  
CRHBP  
GNL3

1376 gene(s) entered


AML Up Genes


☐ Contribute your list so it can be searched by others


Submit


Please acknowledge Enrichr in your publications by citing the following references:  
Chen LY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;12(14).


Kuleshov MY, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016; [gkw377](#).


 modEnrichr  
A suite of gene list enrichment analysis tools

 FlyEnrichr

 YeastEnrichr

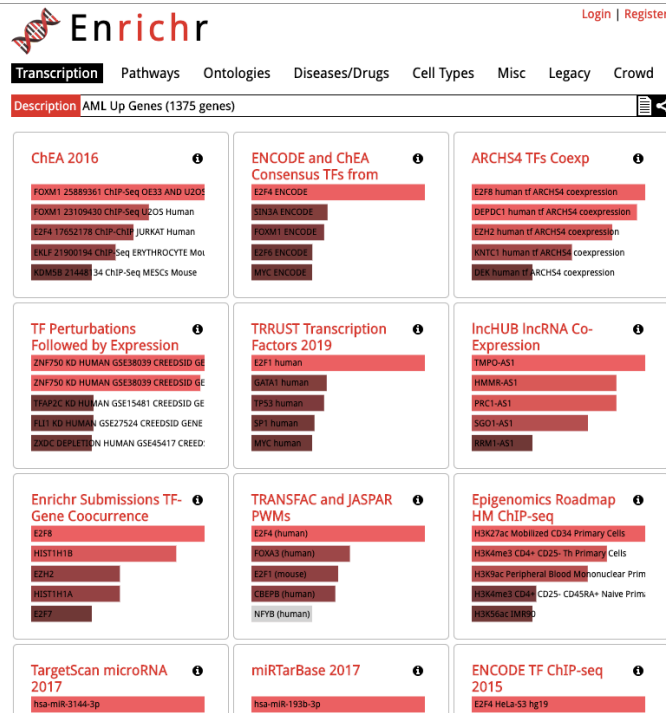
 WormEnrichr

 FishEnrichr

 Click here to raise an issue on GitHub

To help us ensure continued support for Enrichr please [upload a letter of support](#)

پس از سابمیت صفحه ی زیر باز می شود:



اگر وارد پایگاه داده ی TRANSFAC and JASPAR PWMs شویم صفحه ی زیر دیده می شود:

## TRANSFAC and JASPAR PWMs

Bar Graph **Table** Grid Network Clustergram

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 (human)	0.0002224	0.07251	1.89	15.86
2	FOXA3 (human)	0.008371	1.000	1.69	8.07
3	NFYB (human)	0.05171	1.000	2.42	7.18
4	E2F1 (mouse)	0.01471	1.000	1.14	4.80
5	CBEPB (human)	0.01692	1.000	1.15	4.68
6	MYCN (human)	0.1594	1.000	1.73	3.18
7	STAT1 (human)	0.06790	1.000	1.13	3.05
8	HNRNPK (human)	0.1709	1.000	1.69	2.99
9	BCL6 (human)	0.07420	1.000	1.14	2.97
10	NFE2L1 (human)	0.1576	1.000	1.57	2.89



Showing 1 to 10 of 314 entries | [Export entries to table](#)

[Previous](#) [Next](#)

Terms marked with an \* have an overlap of less than 5

همان طور که مشاهده می شود تمام adj p value ها بالاتر از threshold ما که 0.05 است می باشد. بنابراین نمی توانیم از این پایگاه داده استفاده کنیم.

در پایگاه داده ی TRRUST Transcription factors 2019 تصویر زیر مشاهده می شود:

**TRRUST Transcription Factors 2019** Bar Graph **Table** Clustergram  

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	RBL2 mouse	1.944e-8	0.000001850	11.64	206.62
2	TFDP1 human	1.944e-8	0.000001586	11.64	206.62
3	E2F1 human	1.175e-16	6.712e-14	4.45	163.24
4	HOXA9 mouse	0.000008630	0.0004106	12.12	141.34
5	GATA1 human	6.394e-11	1.825e-8	5.36	125.79
6	E2F4 mouse	0.00001724	0.0007032	8.73	95.72
7	E2F3 human	0.000008487	0.0004405	7.83	91.46
8	E2F4 human	0.000001141	0.00007238	6.32	86.54
9	MYC human	4.525e-10	5.168e-8	3.93	84.50
10	RUNX1 human	2.097e-7	0.00001497	5.09	78.29

Showing 1 to 10 of 435 entries | [Export entries to table](#)  
Terms marked with an \* have an overlap of less than 5

Previous Next

چون adj p val خوبی دارند قابل اطمینان هستند و مطابق تصویر های زیر میتوان transcription factor های خوب را شناسایی کرد:

**TRRUST Transcription Factors 2019** Bar Graph **Table** Clustergram  



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	RBL2 mouse	1.944e-8	0.000001850	11.64	206.62
2	TFDP1 human	1.944e-8	0.000001586	11.64	206.62
3	E2F1 human	1.175e-16	6.712e-14	4.45	163.24
4	HOXA9 mouse	0.000008630	0.0004106	12.12	141.34
5	GATA1 human	6.394e-11	1.825e-8	5.36	125.79
6	E2F4 mouse	0.00001724	0.0007032	8.73	95.72
7	E2F3 human	0.000008487	0.0004405	7.83	91.46
8	E2F4 human	0.000001141	0.00007238	6.32	86.54
9	MYC human	4.525e-10	5.168e-8	3.93	84.50
10	RUNX1 human	2.097e-7	0.00001497	5.09	78.29

Showing 1 to 10 of 435 entries | [Export entries to table](#)  
Terms marked with an \* have an overlap of less than 5

Previous Next

**TRRUST Transcription Factors 2019** Bar Graph **Table** Clustergram  

Hover each row to see the overlapping genes.

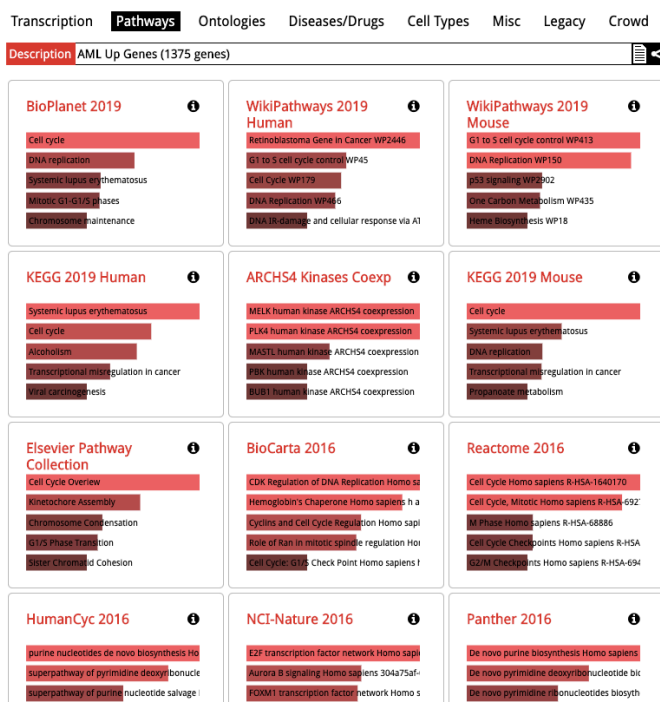
10 entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	RBL2 mouse	1.944e-8	0.000001850	11.64	206.62
2	TFDP1 human	1.944e-8	0.000001586	11.64	206.62
3	E2F1 human	1.175e-16	6.712e-14	4.45	163.24
4	HOXA9 mouse	0.000008630	0.0004106	12.12	141.34
5	GATA1 human	6.394e-11	1.825e-8	5.36	125.79
6	E2F4 mouse	0.00001724	0.0007032	8.73	95.72
7	E2F3 human	0.000008487	0.0004405	7.83	91.46
8	E2F4 human	0.000001141	0.00007238	6.32	86.54
9	MYC human	4.525e-10	5.168e-8	3.93	84.50
10	RUNX1 human	2.097e-7	0.00001497	5.09	78.29



Showing 1 to 10 of 435 entries | [Export entries to table](#)  
Terms marked with an \* have an overlap of less than 5

Previous Next


## در قسمت pathways داریم:



## از پایگاه داده ی KEGG داریم:

**KEGG 2019 Human** Bar Graph **Table** Clustergram  

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	Systemic lupus erythematosus	4.268e-20	1.315e-17	4.92	219.50
2	Cell cycle	8.853e-15	1.363e-12	4.34	140.44
3	Alcoholism	3.472e-13	3.564e-11	3.47	99.69
4	DNA replication	3.756e-7	0.00001928	5.25	77.71
5	Transcriptional misregulation in cancer	2.911e-10	2.241e-8	3.05	66.97
6	Propanoate metabolism	0.000005224	0.0002299	5.00	60.81
7	Valine, leucine and isoleucine degradation	0.00001442	0.0004936	3.94	43.91
8	Base excision repair	0.00004894	0.001370	4.41	43.75
9	One carbon pool by folate	0.0002499	0.005131	5.09	42.23
10	Viral carcinogenesis	1.048e-7	0.000006456	2.61	41.87

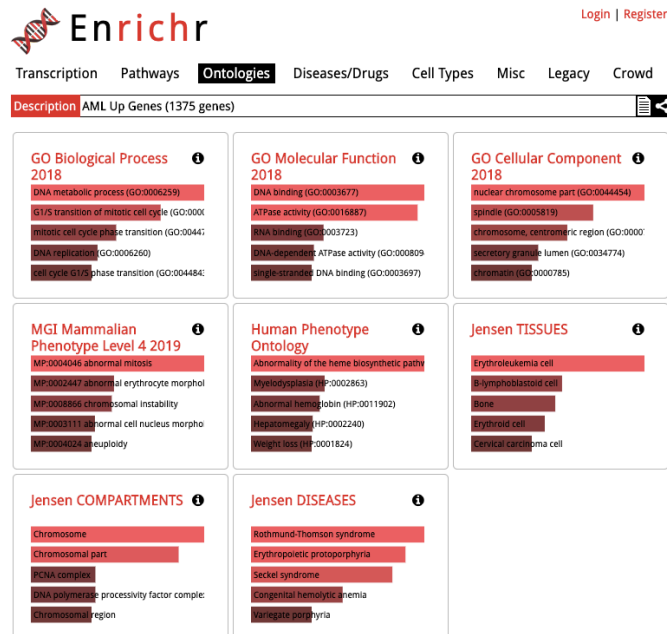
Showing 1 to 10 of 293 entries | [Export entries to table](#)

Terms marked with an \* have an overlap of less than 5

[Previous](#) [Next](#)

همان طور که مشاهده می شود این ژن ها روی افزایش چرخه ی سلولی تأثیر میگذارند و سرطان هم باعث افزایش چرخه ی سلولی می شود.

## در قسمت ontologies تصویر زیر مشاهده می شود:



## در قسمت Jensen COMPARTMENTS داریم:

**Jensen COMPARTMENTS** Bar Graph **Table** Clustergram

Hover each row to see the overlapping genes.

10 entries per page Search:

Index	Name	Combined score
1	DNA packaging complex	400.80
2	Nucleosome	323.38
3	Chromosome	285.11
4	Kinetochore microtubule	284.83
5	Chromosomal region	277.05
6	Chromosomal part	274.80
7	Ndc80 complex	274.28
8	Condensed chromosome outer kinetochore	249.57
9	protein-DNA complex	241.84
10	Cyclin A2-CDK2 complex	206.36

Showing 1 to 10 of 1,845 entries | [Export entries to table](#) Previous Next

Terms marked with an \* have an overlap of less than 5

می توان فهمید که این ژن ها باعث افزایش نوکلئوزوم ها می شوند.  
همچنین از داده های Jensen DISEASES داریم:



## Jensen DISEASES

Bar Graph

**Table**

Clustergram



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	Erythropoietic protoporphyria	0.000008630	0.007814	12.12	141.34
2	Variegate porphyria	0.00002848	0.01032	10.39	108.74
3	Cutaneous porphyria	0.0001051	0.02380	11.64	106.59
4	Porphyria	0.00007163	0.01853	9.09	86.76
5	Rothmund-Thomson syndrome	0.000006782	0.01228	6.84	81.46
6	Hereditary coproporphyria	0.0006578	0.07007	8.31	60.90
7	Meier-Gorlin syndrome	CSF1R, CEBPA, MPL, GATA2, RUNX1	0.02367	6.71	60.74
8	Hematologic cancer	0.0002868	0.04328	7.27	59.32
9	Seckel syndrome	0.00001019	0.006154	4.71	54.09
10	Congenital hemolytic anemia	0.00001921	0.008697	4.85	52.66

Showing 1 to 10 of 1,042 entries | [Export entries to table](#)

[Previous](#) [Next](#)

Terms marked with an \* have an overlap of less than 5

در این پایگاه داده سرطان خون به عنوان بیماری ناشی از این ژن ها آورده شده است.  
در قسمت DISEASES /DRUGS داریم:



Enrichr

[Login](#) | [Register](#)

[Transcription](#) [Pathways](#) [Ontologies](#) **[Diseases/Drugs](#)** [Cell Types](#) [Misc](#) [Legacy](#) [Crowd](#)

Description AML Up Genes (1375 genes)

<b>COVID-19 Related Gene Sets</b> <ul style="list-style-type: none"> <li>COVID-19 patients PBMC up</li> <li>SARS Perturbation Up Genes Mouse Lung In</li> <li>SARS Perturbation Up Genes Mouse Lung In</li> <li>Up-regulated by caffeine and theanine in IAV</li> <li>SARS-CoV perturbation Down Genes bronch</li> </ul>	<b>PheWeb 2019</b> <ul style="list-style-type: none"> <li>tachycardia NOS</li> <li>leukemia and pneumonia</li> <li>disorders of iron metabolism</li> <li>congenital anomalies of genital organs</li> <li>disorders of menstruation and other abnor</li> </ul>	<b>ClinVar 2019</b> <ul style="list-style-type: none"> <li>primary autosomal recessive microcephaly</li> <li>disorder of organic acid metabolism</li> <li>acute myeloid leukemia</li> <li>hereditary anemia</li> <li>neoplasm of the breast</li> </ul>
<b>DepMap WG CRISPR Screens Sanger CellLines</b> <ul style="list-style-type: none"> <li>HARA</li> <li>PC1-15A</li> <li>ES5</li> <li>OPM-2</li> <li>PC1-30</li> </ul>	<b>DepMap WG CRISPR Screens Broad CellLines</b> <ul style="list-style-type: none"> <li>HB11:19</li> <li>U-937</li> <li>P21-PUG</li> <li>TF-1</li> <li>U510-211H</li> </ul>	<b>GWAS Catalog 2019</b> <ul style="list-style-type: none"> <li>immature fraction of reticulocytes</li> <li>high light scatter reticulocyte percentage of</li> <li>plateletcrit</li> <li>hemoglobin levels</li> <li>beta thalassemia/hemoglobin E disease</li> </ul>
<b>UK Biobank GWAS v1</b> <ul style="list-style-type: none"> <li>Platelet count 30080 raw</li> <li>Platelet crit 30090 raw</li> <li>Mean corpuscular haemoglobin 30050 raw</li> <li>Mean spherised cell volume 30270 raw</li> <li>Immature reticulocyte fraction 30280 raw</li> </ul>	<b>DisGeNET</b> <ul style="list-style-type: none"> <li>Leukemia, Myelocytic, Acute</li> <li>leukemia</li> <li>Myeloid Leukemia, Chronic</li> <li>Carcinogenesis</li> <li>Mammary Neoplasms</li> </ul>	<b>DSigDB</b> <ul style="list-style-type: none"> <li>LUCANTHONE CTD 00006227</li> <li>Enterolactone CTD 00001393</li> <li>Dasatinib CTD 00004330</li> <li>ECUMESTROL CTD 00000117</li> <li>Testosterone CTD 00006844</li> </ul>
<b>ARCHS4 IDG Coexp</b> <ul style="list-style-type: none"> <li>UICK2 IDG kinase ARCHS4 coexpression</li> </ul>	<b>LINCS L1000 Chem Pert up</b> <ul style="list-style-type: none"> <li>LP005 HEPG2 24H-WZ-3105-0.37</li> </ul>	<b>LINCS L1000 Chem Pert down</b> <ul style="list-style-type: none"> <li>LP006 HT29 24H-palbociclib-0.37</li> </ul>

در پایگاه داده ی DisGeNET داریم:



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	Beta thalassemia intermedia	1.469e-8	0.000001552	8.00	144.29
2	Absent ear	0.000007324	0.0004067	9.70	114.66
3	Congenital absence of external ear	0.000007324	0.0004044	9.70	114.66
4	Chromosome Breakage	1.794e-9	2.205e-7	5.62	111.71
5	Thalassemia Intermedia	1.168e-8	0.000001262	6.17	112.71
6	Myeloid Leukemia, Chronic	4.738e-21	1.552e-17	2.31	108.01
7	delta-Thalassemia	0.0001051	0.003677	11.64	106.59
8	Acute Undifferentiated Leukemia	1.191e-10	2.208e-8	4.59	104.96
9	Undifferentiated leukemia	4.166e-10	6.203e-8	4.71	101.64
10	Leukemia, Myelocytic, Acute	3.697e-23	3.633e-19	1.93	99.70

Showing 1 to 10 of 7,167 entries | [Export entries to table](#)

[Previous](#) [Next](#)

Terms marked with an \* have an overlap of less than 5

همان طور که مشاهده می شود با adj p value خوبی می توان گفت که این ژن ها مربوط به بیماری Leukemia بوده اند.  
پس میتوان با اطمینان خاطر بیشتری بیان کرد که آنالیز درستی را پشت سر گذاشته ایم.