به نام خدا

# تحلیل دادگان ریزآرایه‌ی لوکمی حاد مغز استخوان

مقدمه‌ای بر بیوانفورماتیک

پروژه ی نهایی


دکتر علی شریفی زارچی

دکتر سمیه کوهی

گردآورنده: مهدی کافی

# فهرست مطالب

## مقدمه

سرطان‌های لوکمی در سلول‌هایی به وجود می‌آیند که در شرایط طبیعی به سلول‌های خون تبدیل می‌شوند. یکی از انواع این سرطان خون، لوکمی حاد مغز استخوان (AML)[1] است. این بیماری به سرعت پیشرفت می‌کند، به سختی درمان می‌شود و متاسفانه در طی مدت کوتاهی معمولا چند ماه، بیمار را از بین می‌برد. AML از مغز استخوان (قسمت‌های نرم داخل بعضی استخوان‌های مشخص که سلول‌های جدید خون در آنجا ساخته می‌شوند) آغاز می‌شود و در اغلب قریب به اتفاق موارد، سریعا از مغز استخوان خارج شده و با جریان خون به سایر نقاط بدن مانند غدد لنفاوی، کبد، سیستم اعصاب مرکزی و ... گسترش می‌یابد. در این پروژه قصد داریم که با استفاده از تحلیل داده‌های ریزآرایه[2]ی لوکمی حاد مغز استخوان، ژن‌هایی که در این سرطان نقش مؤثری دارند را بیابیم. [1]

# مواد و روش‌ها

## چکیده

برای تحلیل از داده‌های GSE48558 استفاده شده‌است. برای تحلیل این داده‌ها از زبان R و همینطور کتابخانه‌های GEOquery، limma، umap، pheatmap، ggplot2، reshape2 و plyr استفاده شده‌است. در ابتدا داده‌ها بر اساس phenotype و source name دسته بندی شدند. دسته بندی به این صورت انجام شده است که داده‌هایی که phenotype آن‌ها normal بوده است، در گروه نرمال و داده‌هایی که source name آن‌ها AML patient بوده‌است، در گروه تست قرار گرفته‌اند. سپس ماتریس بیان ژن‌ها برای این نمونه‌ها ایجاد شده‌است. برای کنترل کیفیت داده‌ها از نمودارهای جعبه‌ای استفاده شده‌است. سپس همبستگی دو به دوی نمونه‌های نرمال و تست و همینطور خوشه بندی نمونه‌ها صورت گرفت. سپس داده‌های ژن‌ها و همینطور نمونه‌ها کاهش ابعاد یافتند و مشخص شد که داده‌ها برای تحلیل‌های آینده مناسب هستند. سپس دسته‌های نرمال و تست، برای یافتن ژن‌های با میزان بیان متفاوت معنی‌دار مقایسه شدند و با استفاده از این ژن‌ها و پایگاه داده Enrichr، سعی شد که ژن‌هایی که در این سرطان نقش مهمی بازی می‌کنند، پیدا شوند و بررسی‌های gene ontology و pathway انجام شوند. درنهایت برای بررسی صحت نتایج به دست آمده، این نتایج با مقاله‌های موجود مقایسه شدند.

## واکشی داده‌ها

در زبان R با قطعه کد زیر داده‌ها را دانلود و دسته بندی و سپس ماتریس بیان ژن را از آن‌ها استخراج کردیم.

[1] Acute Myeloid Leukemia
[2] Microarray

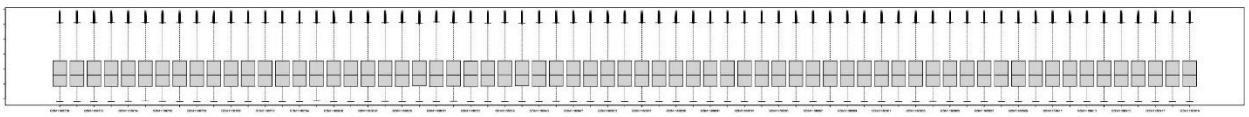
```
###Fetching Data
series = "GSE48558"
platform = "GPL6244"
gset = getGEO(series, GSEMatrix = TRUE, AnnotGPL = TRUE, destdir = "data/")
if (length(gset) > 1) idx <- grep(platform, attr(gset, "names")) else idx <-
1
gset <- gset[[idx]]
groups = c(rep("test", 13), rep("X", 27), "normal", rep("X", 3), "normal"
, rep("X", 23), "normal", "X", "normal", rep("X", 3), "normal", "X"
, rep("normal", 4), "X", "normal", rep("X", 2), rep("normal", 2)
, rep("X", 2), rep("normal", 2), "X", "normal", "X", "normal", "X"
, "normal", "X", "normal", "X", "normal", rep("X", 3), "normal"
,rep("X", 3), "normal", rep("X", 29), rep("normal", 7), rep("test", 2)
, "normal", rep("test", 3), rep("normal", 20))
sel <- which(groups != "X")
groups = groups[sel]
gset_groups = gset[, sel]
ex = exprs(gset_groups)
```

## کنترل کیفیت

با قطعه کد زیر نمودارهای جعبه‌ای را برای بیان ژن‌ها برای ۶۷ نمونه انتخابی رسم کردیم.

```
###Box plot
pdf("results/boxplot.pdf", width = 67)
boxplot(ex)
dev.off()
```

نمودار خروجی در زیر آورده شده‌است.



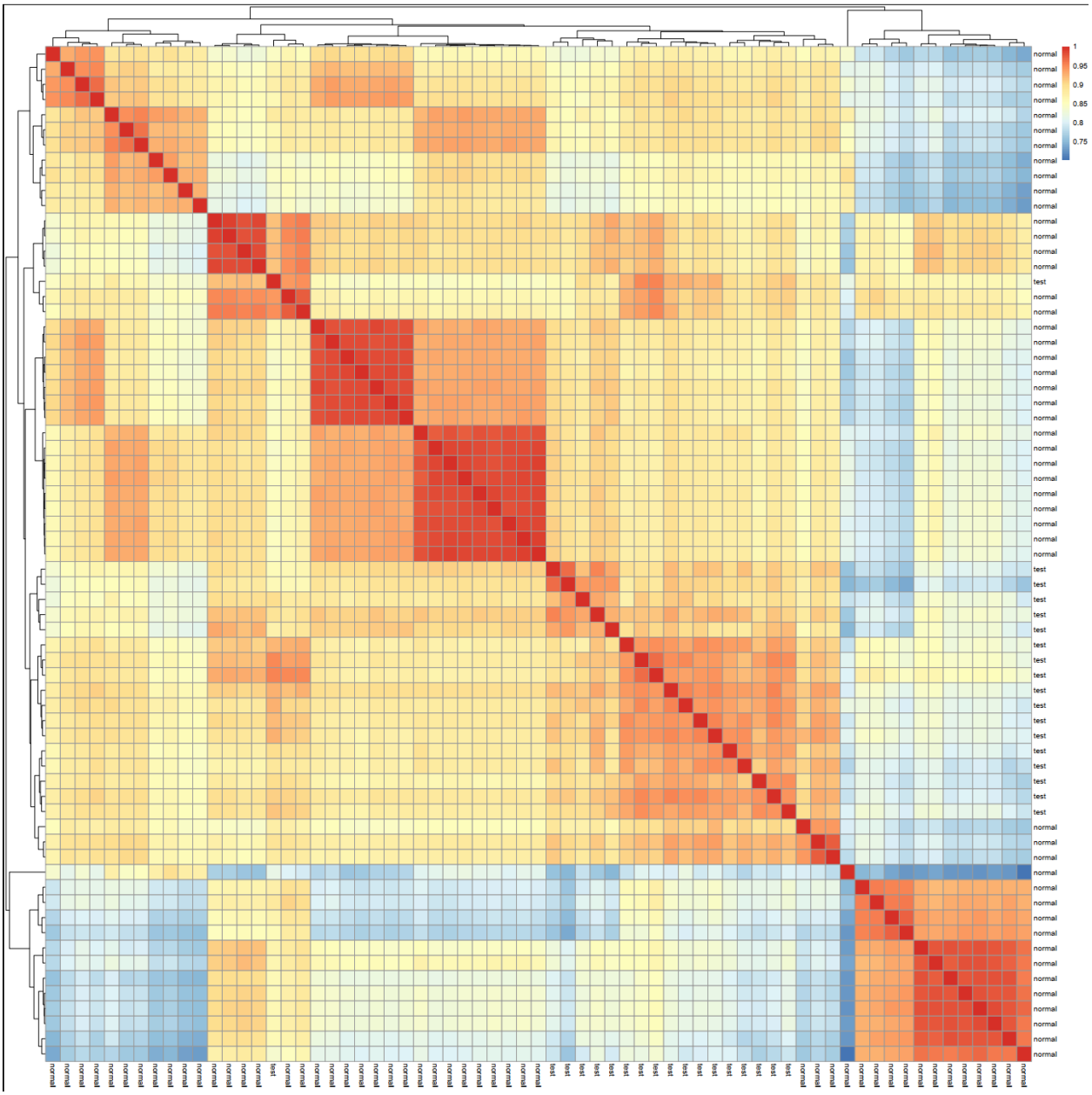
در نمودار دیده می‌شود که مقدار کمینه، بیشینه، چارک اول، میانه و چارک سوم برای تمام ۶۷ نمونه در یک بازه نزدیک به هم قرار دارد و همینطور با بررسی مقدار کمینه و بیشینه متوجه می‌شویم که داده‌ها لگاریتمی هستند در نتیجه داده‌ها مناسب تحلیل هستند و نیازی به تغییر ندارند.

## همبستگی بین نمونه‌ها

با قطعه کد زیر نمودار همبستگی بین نمونه‌ها براساس گروه‌های انتخابی را رسم کردیم.

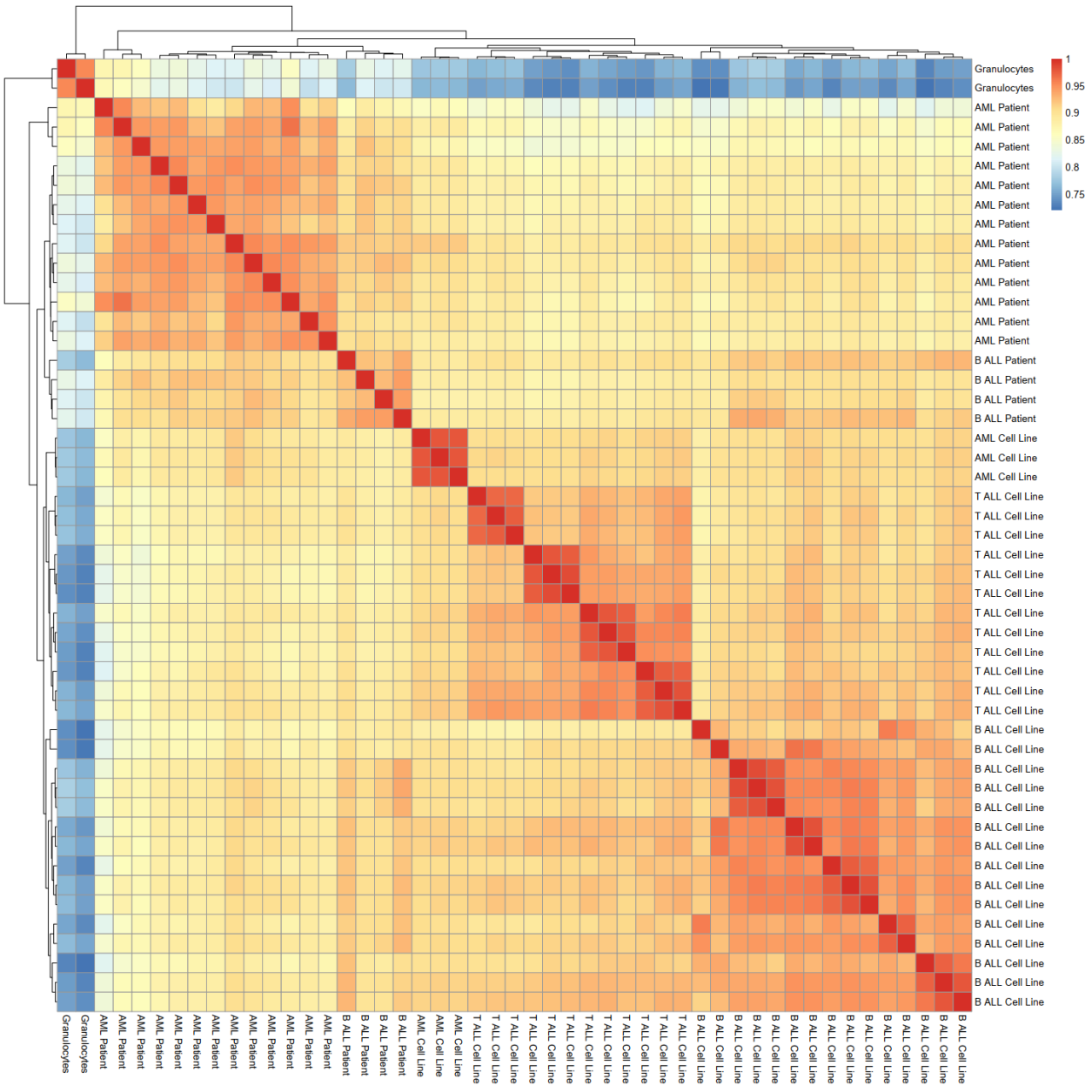
```
###Correlation Heat map
pdf("results/cor_heatmap.pdf", width = 20, height = 20)
pheatmap(cor(ex), labels_row = groups, labels_col = groups)
dev.off()
```

نمودار همبستگی در ادامه آمده‌است.




با توجه به این نمودار دیده می‌شود که نمونه‌هایی که از یک گروه هستند همبستگی بیشتری با یکدیگر دارند. اما نکته دیگری که باید به آن توجه کنیم این است که برخی نمونه‌ها از گروه نرمال همبستگی منفی با یکدیگر دارند که احتمالا به دلیل تفاوت در source name آن‌ها است؛ برای بررسی این موضوع نمودار همبستگی بین نمونه‌های گروه نرمال را بر اساس source name، رسم کردیم. این نمودار با استفاده از قطعه کد زیر رسم شده‌است.

```
###Correlation Heat Map for Normal Samples
normal_groups <- c(rep("test", 13), rep("X", 27), "normal", rep
("X", 3), "normal", rep("X", 23), "normal", "X", "normal", rep(
"X", 3), "normal", "X", rep("normal", 4), "X", "normal", rep("X
", 2), rep("normal", 2), rep("X", 2), rep("normal", 2), "X", "n
ormal", "X", "normal", "X", "normal", "X", "normal", "X", "norm
al", rep("X", 3), "normal",rep("X", 3), "normal", rep("X", 29),
 rep("normal", 7), rep("test", 2), "normal", rep("test", 3), re
p("normal", 20))
normal_sel <- which(normal_groups == "normal")
normal_groups <- normal_groups[normal_sel]
normal_gset <- getGEO(series, GSEMatrix = TRUE, AnnotGPL = TRUE
, destdir = "data/")
if (length(gset) > 1) idx <- grep(platform, attr(gset, "names")
) else idx <- 1
normal_gset <- normal_gset[[idx]]
normal_gset <- normal_gset[, normal_sel]
normal_ex <- exprs(normal_gset)
normal_sourcenames <- normal_gset$source_name_ch1
pdf("results/normal_cor_heatmap.pdf", width = 15, height = 15)
pheatmap(cor(normal_ex), labels_row = normal_sourcenames, label
s_col = normal_sourcenames)
dev.off()
```




دیده می‌شود که نمونه‌هایی که source name یکسانی دارند به یکدیگر نزدیکتر هستند و بعضی گروه‌ها از سایر گروه‌ها نیز دور هستند.

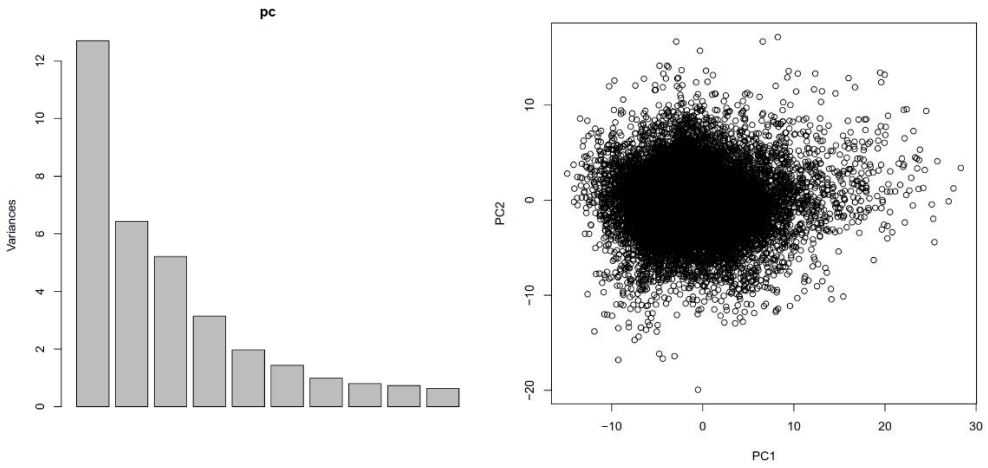
## کاهش ابعاد

در ابتدا کاهش ابعاد را روی ژن‌ها انجام دادیم و سپس مشاهده کردیم که میزان بیان برخی ژن‌ها در تمام نمونه‌ها تقریبا یکسان بود و با توجه به اینکه در کاهش ابعاد هدف این است که داده‌ها با بیشترین تفاوت را بیابیم، مورد مطرح شده، باعث می‌شود که کاهش ابعاد به خوبی عمل نکند بنابراین باید تأثیر این ژن‌ها را از بین ببریم. به این منظور میانگین بیان هر ژن را از میزان بیان آن ژن‌ها کم می‌کنیم در اینصورت ژن‌هایی که در همه نمونه‌ها بیان یکسانی داشتند، برابر با صفر می‌شوند. این کار را با قطعه کد زیر به این صورت انجام دادیم که در ابتدا ماتریس بیان ژن را، ترانهاده کرده و سپس مقادیر آن را scale می‌کنیم و مقدار scale را در تابع برابر با False قرار می‌دهیم تا میزان بیان ژن که میانگین از آن کم شده‌است را بر انحراف معیار تقسیم نکند و سپس روی داده به دست آمده، کاهش ابعاد را اعمال می‌کنیم.

```r
###PC on genes
pc = prcomp(ex)
pdf("results/PC.pdf")
plot(pc)
plot(pc$x[,1:2])
dev.off()

ex_scaled = t(scale(t(ex), scale = FALSE))
pc = prcomp(ex_scaled)
pdf("results/pc_scaled.pdf")
plot(pc)
plot(pc$x[, 1:2])
dev.off()
```

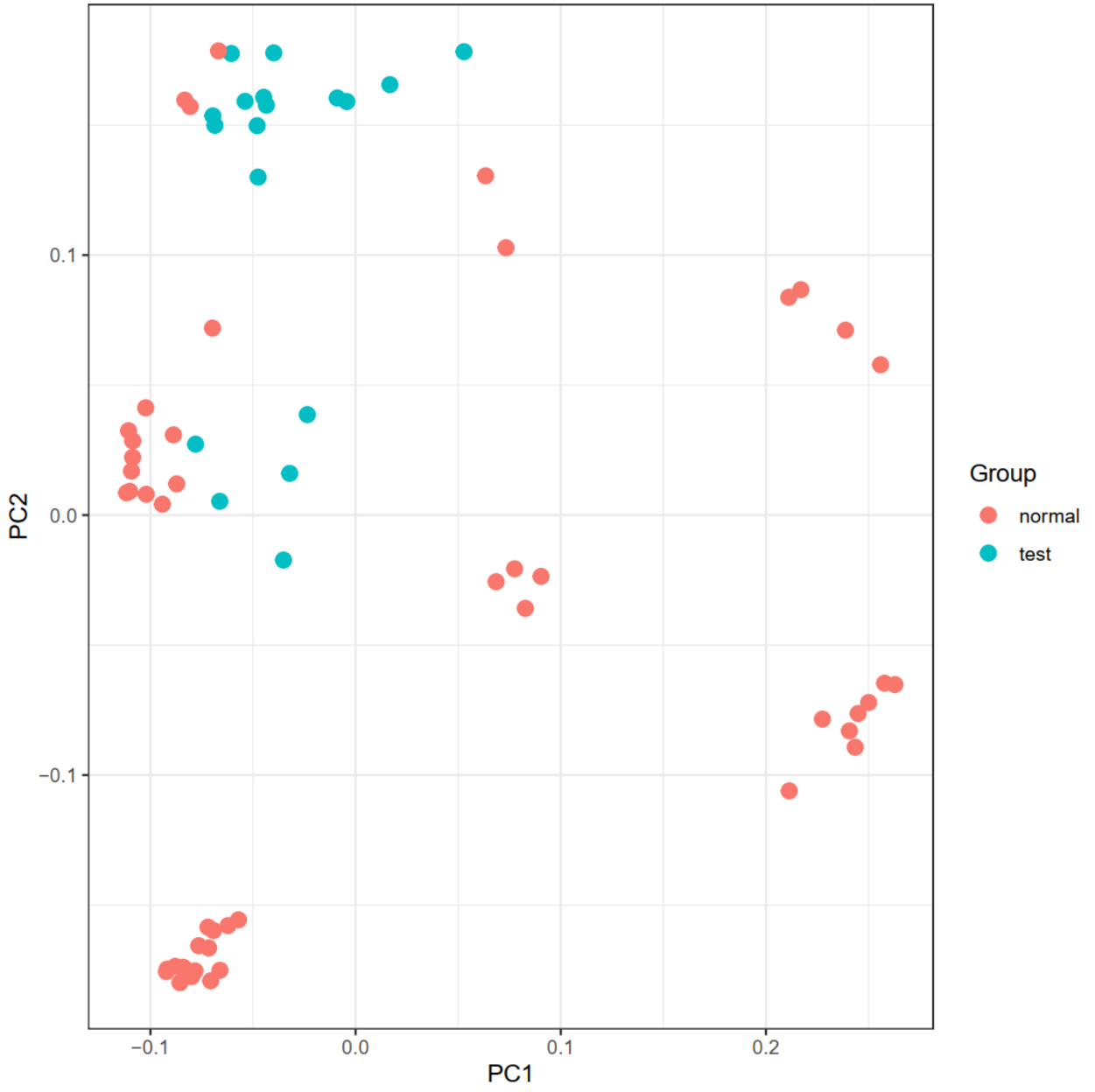
نمودار خروجی قطعه کد بالا که داده‌ها در ابتدا scale شده‌است به صورت زیر است؛ همچنین نمودار کاهش ابعاد بدون scale شدن نیز در فایل‌های اضافی پروژه آورده شده‌است.




در ادامه کاهش ابعاد را روی نمونه‌ها انجام دادیم. به این منظور از قطعه کد زیر استفاده کردیم.

```
###PC on samples
pcr = data.frame(pc$rotation[, 1:3], Group = groups)
pdf("results/pca_samples.pdf")
ggplot(pcr, aes(x=PC1, y=PC2, color = Group)) + geom_point(size
= 3) + theme_bw()
dev.off()
```

نمودار کاهش ابعاد یافته حاصل از این قطعه کد روی نمونه‌ها به صورت زیر است.



دیده می‌شود که اگر نمونه‌ها را بر روی principal component 1، تصویر کنیم، نمونه‌های تست بسیار به یکدیگر نزدیک هستند و همینطور نمونه‌های نرمال نیز چند گروه را تشکیل می‌دهند که اعضای گروه‌ها به یکدیگر نزدیک هستند.

## بررسی تمایز در بیان ژن‌ها

در این مرحله برای یافتن ژن‌هایی که تفاوت معناداری در بیان دارند، ابتدا لازم است تفاوت بیان ژن‌ها بین نمونه‌های تست و نرمال را به دست بیاوریم، برای این منظور از قطعه کد زیر استفاده کردیم.

```
###Differential Expression Analysis
gs <- factor(groups)
gset_groups$group <- gs
design <- model.matrix(~group + 0, gset_groups)
colnames(design) <- levels(gs)

fit <- lmFit(gset_groups, design)  # fit linear model
cts <- "test-normal"
cont.matrix <- makeContrasts(contrasts= cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="bonferroni", sort.by="logFC", numb
er=Inf)
tT <- subset(tT, select=c("Gene.symbol","Gene.ID", "adj.P.Val",
"logFC"))
```

در این قطعه کد در ابتدا ماتریسی ساختیم که مشخص می‌کند هر نمونه از چه گروهی است. سپس یک مدل خطی بر داده‌ها fit می‌کند و سپس ماتریس تفاوت بیان ژن‌ها بین نمونه‌های تست و نرمال را ایجاد می‌کند و در نهایت یک جدول top table ایجاد می‌کند این جدول شامل مقادیر زیادی از جمله p-value، adjusted p-value، آماره B، logFC و... است که تنها مقادیری که از آن را نیاز داریم برای مرحله بعد انتخاب می‌کنیم.

سپس با توجه به اینکه در محاسبه جدول tT، مقادیر بیان نمونه‌های نرمال رو از گروه تست کم کردیم بنابراین مقادیر مثبت logFC، نشانگر آن است که ژن در نمونه‌های تست، بیان بیشتری داشته و همینطور مقادیر منفی بیانگر آن است که ژن در نمونه‌های تست بیان کمتری داشته‌است. برای پیدا کردن ژن‌هایی که تفاوت بیان معناداری بین نمونه‌های تست و نرمال دارند از قطعه کد زیر استفاده کردیم.

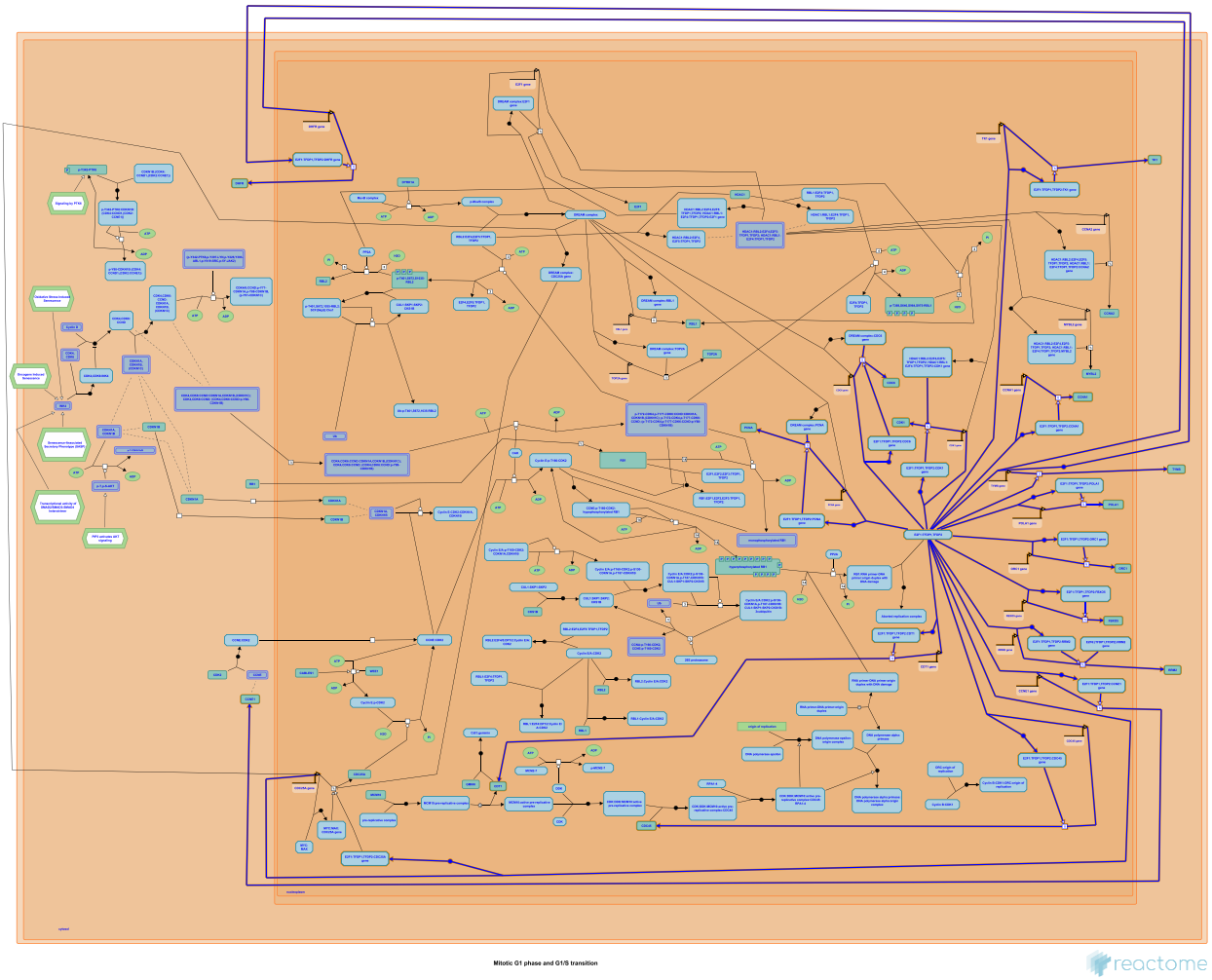
```
aml.up <- subset(tT, logFC > 1 & adj.P.Val < 0.05)
aml.up.genes <- unique(as.character(strsplit2(aml.up$Gene.symbo
l, "///")))

aml.down <- subset(tT, logFC < -1 & adj.P.Val < 0.05)
aml.down.genes <- unique(as.character(strsplit2(aml.down$Gene.s
ymbol, "///")))
```

در این قطعه کد، در ابتدا ژن‌هایی که تفاوت بیان بیش از دو برابری در نمونه‌های تست داشتند و همینطور با مقدار adjusted p-value کمتر از ۰/۰۵ بودند را به عنوان ژن‌های با بیان بیشتر معنادار در نمونه‌های AML معرفی کردیم و سپس این نمونه‌ها را مرتب‌تر کردیم؛ سپس کاری مشابه را برای بیان معنادار کمتر در نمونه‌های تست انجام دادیم. لیست این ژن‌ها در فایل‌های اضافی پروژه آورده شده‌است.

## بررسی gene ontology و pathway

برای بررسی gene ontology و pathway، به این صورت عمل کردیم که در پایگاه داده Enrichr، لیست ژن‌هایی که با بیان معنادار بیشتر در نمونه‌های تست به دست آورده‌بودیم را وارد کردیم و ژن‌هایی که باعث این می‌شدند که ژن‌های مورد نظر ما، بیان بیشتری داشته‌باشند را پیدا کردیم به طور مثال ژن‌های E2F4 و FOXM1 باعث می‌شوند که ژن‌هایی که ما پیدا کردیم، بیان بیشتری داشته‌باشند و سپس این ژن‌ها را با مقالات مقایسه کردیم و دیدیم که هر دوی این ژن‌ها باعث سرطان AML می‌شوند.[2] و همینطور می‌دانیم که ژن E2F4 نقش مهمی در فرآیند cell cycle دارد و به طور خاص با بررسی pathwayها متوجه شدیم که این ژن با نقشی که در pathway، G1/S-specific transcription بازی می‌کند می‌تواند باعث AML شود. این pathway در سایت reactome نیز آمده‌است و در آنجا هم به ژن‌های خانواده E2F اشاره کرده‌است. تصویر این pathway در زیر آمده‌است.




و روند مشابهی را برای ژن‌هایی که کاهش بیان معناداری در نمونه‌های تست داشتند اعمال کردیم و در اینجا نیز ژن HIVEP2 را به دست آوردیم که در یک مقاله که در فایل‌های اضافی آورده شده‌است، به عنوان ژنی مؤثر در AML معرفی شده‌است.[3]

# منابع


[1] ق. نادر و ک. سید سعید, "سرطان نیازمند ایده‌های نوین," [درون خطی]. Available: http://cnin.ir/Cancer-Types.aspx?10554.

[2] Y. Feng, L. Li, Y. Du and F. Chen, "E2F4 functions as a tumour suppressor in acute myeloid leukaemia via inhibition of the MAPK signalling pathway by binding to EZH2," *Journal of Cellular and Molecular Medicine,* pp. 2157-2168, 2020.

[3] A. S. L. N. L. R. H. M. W. N. H. M. E. M. N. H. N. Y. C. Y. C. T. S. G. M. F. A. C. D. I. L. B. N. F. P. D. D. Noa Novershtern, "Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis," *Cell,* vol. 144, no. 2, pp. 296-309, 2011.