بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

درس: مقدمه بیوانفورماتیک

استاد درس: دکتر شریفی، دکتر کوهی

پروژه پایانی: تجربه اولیه در تحلیل داده‌های بیوانفورماتیک

(بررسی داده‌های سرطان خون)

دانشجو:

محمدحسین موثقی‌نیا

شماره دانشجویی:

۴۰۰۲۰۰۹۱۹

بهمن ۱۴۰۰

فهرست

۱. مقدمه

سرطان خون انواع مختلفی دارد، یکی از انواع آن لوسمی حاد میلوئیدی[1] است. در این نوع سرطان، سلول‌های مغز استخوان یا میلوسیت‌ها تحت تاثیر فاکتورهایی قرار می‌گیرند که باعث تولید میلوبلاست (نوعی از گلبول‌های سفید) و گلبول‌های قرمز و پلاک‌های غیر طبیعی (جهش یافته) می‌شود و تعداد زیادی سلول بیمار تولید می‌شود و در نتیجه آن فرآیند طبیعی عملکرد سلول‌های خونی از بین رفته و بدن دچار مشکلات جدی می‌شود. از جمله نتایج آن ضعف سیستم ایمنی بدن، کم خونی و اختلال انعقاد خون را می‌توان نام برد[2]. در این تحقیق به بررسی دادگان مربوط ریزآرایه‌های سلول‌های بیمار به این گونه سرطان در مقایسه با افراد سالم می‌پردازیم. در بخش‌های ابتدایی دادگان را از نظر کیفیت مورد بررسی و اصلاح قرار می‌دهیم و در ادامه به بررسی ژن‌های موثر در اینگونه بیماری و بررسی pathwayها و gene ontology مربوط به این ژن‌هایی که در بیماران افزایش یا کاهش بیان معنی‌داری داشته‌اند می‌پردازیم.

۱-۱ دادگان و ابزارها و روش‌های مورد استفاده

به منظور این بررسی از سایت GEO، دادگان شماره سری GSE48558 استفاده شده‌است. به عنوان ابزار بررسی و تحلیل داده‌ها از زبان برنامه نویسی R و محیط برنامه نویسی R-Studio استفاده شده‌است و پکیج‌های GEOquery ، ggplot2 ، pheatmap ، limma ، gplots مورد استفاده قرار گرفته‌است. از بین داده‌های سری مورد نظر، گروهی که source name آن‌ها AML Patient بوده به عنوان گروه test و گروهی که phenotype آن‌ها normal بوده است به عنوان گروه normal انتخاب شده‌اند و دیگر دسته‌ها مورد بررسی قرار نگرفته‌است. به منظور بررسی ژن‌های افزایش و کاهش یافته از دیتابیس‌های Enrichr استفاده شده‌است.

۱-۲ تنظیمات و موارد اولیه

به منظور تحلیل داده‌ها ابتدا می بایست، داده‌ها را دانلود و در دایرکتوری مناسب قرار داده و نمونه‌های نرمال و تست را مشخص نمود. برای این منظور از طریق قطعه کد زیر، محل فعلی دایرکتوری را به عنوان مرجع مشخص می کنیم و سپس براساس آن آدرس دهی‌های بعدی را انجام می دهیم:

```
curD <- dirname(rstudioapi::getActiveDocumentContext()$path)
setwd(sub(paste0("/",sub("(.+)/","",curD)),"",curD))
```

[1] AML: Acute myeloid leukemia
[2] https://en.wikipedia.org/wiki/Acute_myeloid_leukemia

سپس می توان شماره سری داده‌ها و پلتفرم آن‌ها را به صورت مجزا تعیین نمود، البته الزامی به این کار نیست و می توان داخل کد دستوری قرار داد، پس از آن می بایست داده‌ها را در دایرکتوری مورد دلخواه خود دانلود نمود:

```
series <- "GSE48558"
platform <- "GPL6244"
gset <- getGEO(series, GSEMatrix = TRUE, AnnotGPL = TRUE, destdir
= 'data/')
```

برای این منظور از تابعی در کتابخانه GEOquery به منظور دانلود استفاده می‌شود.

پس از آن درصورتی که داده چندین پلتفرم داشته باشد، صرفا پلتفرم خاص تعیین شده توسط سوال را می خواهیم. برای این منظور می بایست یک فیلتر روی داده صورت پذیرد:

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

البته درصورتی که از داده length گرفته شود و صرفا یک پلتفرم موجود باشد، کافیست همان جایگزین شود و دیگر به عنوان لیستی از پلتفرم‌ها نباشد. در نهایت می بایست گروه‌های نرمال و بیمار و غیره را مشخص کنیم و گروه‌های به جز بیمار و نرمال را از داده‌ها حذف کنیم:

```
gsms <-
paste0("111111111111XXXXXXXXXXXXXXXXXXXXXXXXXX0XXX0XXXXX",
            "XXXXXXXXXXXXXXXXXX0X0XXX0X0000X0XX00XX00X0X0X0X0X0",
            "XXX0XXX0XXXXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",
              "00000000000000000000")
sml <- strsplit(gsms, split="")[[1]]
### filter by X
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]
gs <- factor(sml)
groups <- make.names(c("normal", "test"))
levels(gs) <- groups
gset$group <- gs
```

از طریق قطعه کد بالا می توان داده‌ها را دسته بندی کرده و یک ستون جدید به داده‌ها افزود و گروه بندی مورد نظر را در آن گنجاند. (البته گروه‌های غیراز سالم و بیمار، از داده‌ها حذف می شوند.)

۲. کنترل کیفیت داده‌ها

این قسمت شامل موارد زیر می‌باشد که با توجه به تعریف صورت سوال پروژه در قسمت‌های مجزایی بررسی خواهند شد:

- نرمال‌سازی دادها در صورت نیاز، به منظور بررسی صحیح داده‌ها، چرا که در صورت عدم نرمال‌سازی ممکن است در صورت‌اندازه گیری نمونه‌ها با کیت‌های مختلف، تاثیرات محیطی و ... دو نمونه که در واقعیت شباهت زیادی به هم داشته‌اند به دلایل نام برده تفاوت فاحشی با هم پیدا کنند.
- کاهش ابعاد به منظور بررسی پیوستگی و شباهت دادهای مورد بررسی
- بررسی همبستگی بین نمونه‌ها به جهت اطمینان از مرتبط بودن نمونه‌های جمع آوری شده، می‌باشد.

البته بخش‌های بالا صرفا به منظور بررسی کیفیت داده نمی‌باشد و اطلاعات گسترده دیگری نیز در اختیار قرار می‌دهد که در هر بخش بیشتر بحث خواهد شد.

۲-۱. نرمال‌سازی و نمودار جعبه‌ای

به منظور بررسی نرمال بودن داده‌ها، یکی از راه‌ها رسم نمودار جعبه‌ای نمونه‌هاست. درصورتی که نمودارهای جعبه‌ای حاصل به صورت منظمی در کنار هم و تقریبا مشابه یک دیگر از نظر چارک‌های اول، دوم و سوم باشند؛ می‌توان نتیجه گیری کرد که داده‌ها با هم هماهنگ و نرمال هستند. درغیر این صورت می‌بایست نرمال‌سازی با یکی از روش‌های مناسب، صورت پذیرد. یکی از روش‌های متداول و ساده نرمال‌سازی، نرمال‌سازی گسسته[1] می‌باشد، که به این صورت عمل می‌کند در هر دور اجرای آن ماکسیمم هر سری از داده‌ها را میانگین می‌گیرد و جایگزین تمامی آن‌ها می‌کند و در دور بعدی آن‌ها را در نظر نمی‌گیرد. به این‌ترتیب داده‌ها در یک طیف یکسان دسته‌بندی می‌شوند. البته روش‌های دیگری که تفاوت داده‌ها را بتواند بهتر پوشش دهد نیز وجود دارد.

[1] Quantile Normalization

داده‌های موجود در این پروژه در صورت رسم نمودار جعبه‌ای نرمال هستند. (شکل ۱) به منظور این کار می‌بایست، ابتدا ماتریس بیان ژن‌ها را از داده‌ها به دست آوریم، برای این منظور از طریق قطعه کد زیر این کار را انجام می‌دهیم:

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

درصورت نیاز به نرمال‌سازی، می‌توان از قطعه کد زیر استفاده نمود:

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

تابع exprs یک تابع دو طرفه بوده و می‌تواند مقادیر را علاوه بر خروجی گرفتن، مقداردهی نیز انجام دهد. تابع normalizeQuantiles نیز همان نرمال‌سازی گسسته را برای داده‌های بیان ژن انجام می‌دهد.

که البته در داده‌های موجود، با توجه به شکل ۱ نیازی به نرمال‌سازی نیست، درصورت اجرای نرمال‌سازی بالا و رسم دوباره نمودار جعبه ای، نمودار حاصل، شکل ۲ خواهد بود که کاملا مشخص است که تغییر قابل توجه‌ای رخ نداده است. بنابراین داده‌ها به صورت پیشفرض نرمال بوده‌اند.

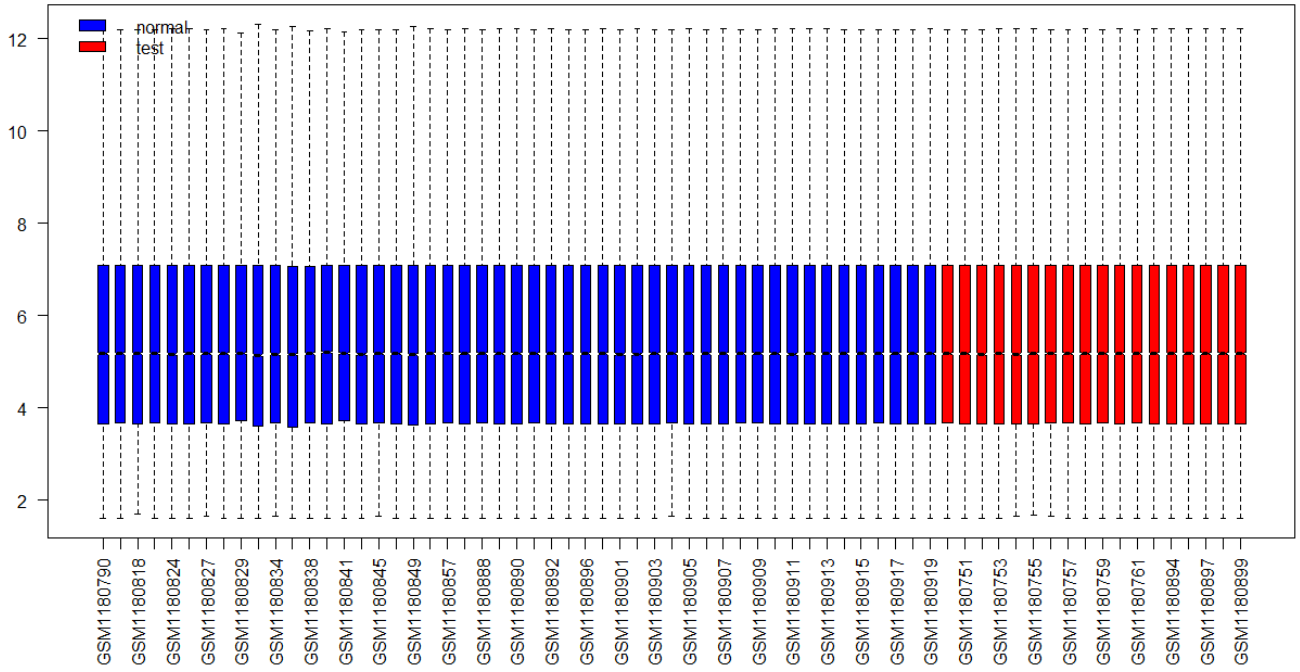
سپس به منظور رسم نمودار جعبه‌ای بر اساس ژن‌های بیان شده از طریق قطعه کد زیر اقدام می‌کنیم:

```
boxplot(ex)
```

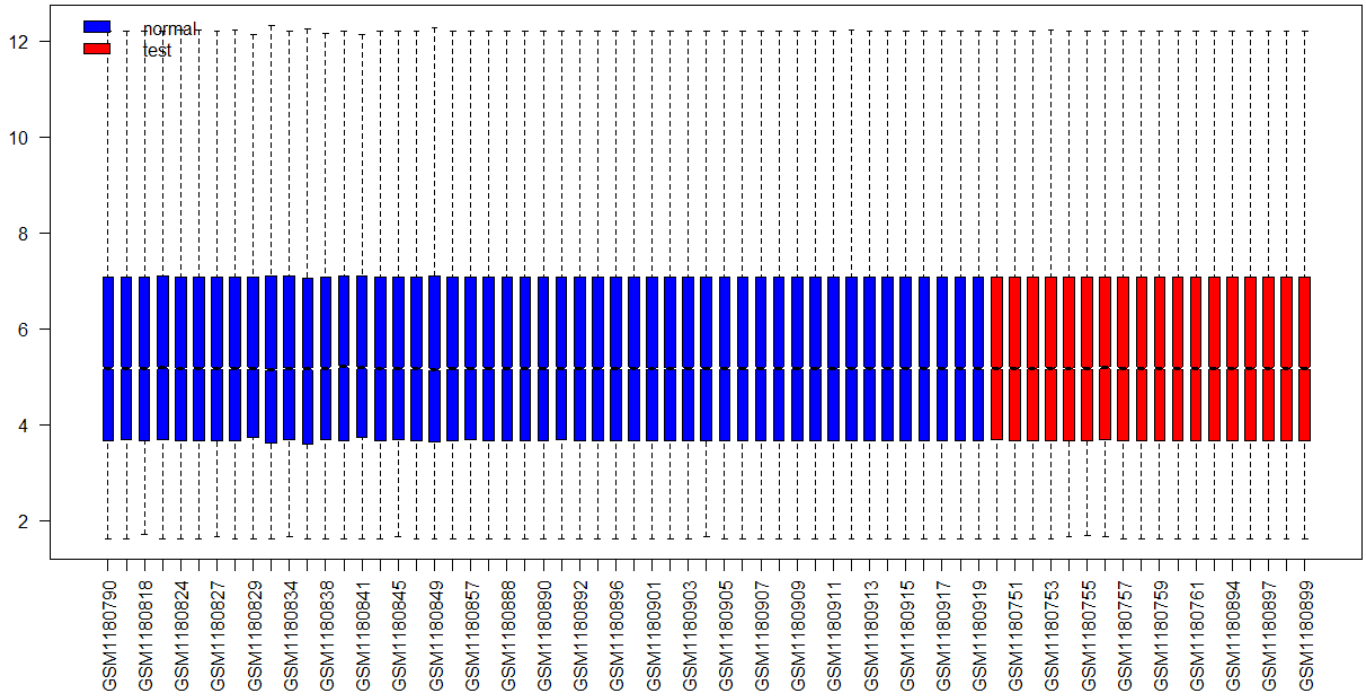
به منظور ساخت بهتر نمودار جعبه‌ای و ایجاد قابلیت تفکیک بین نمونه‌های سالم و بیمار، با استفاده از قطعه کد زیر می توان این تفاوت را ایجاد نمود:

```
dev.new(width=3+ncol(gset)/6, height=5)
ord <- order(gs)
palette(c("blue", "red"))
par(mar=c(7,4,2,1))
boxplot(ex[,ord], boxwex=0.6, notch=T, outline=FALSE, las=2,
col=gs[ord])
legend("topleft", groups, fill=palette(), bty="n")
```

در کد بالا، ابتدا ابعاد نمودار خروجی مبتنی بر نمونه‌ها مشخص می‌شود، سپس ترتیب نمایش نمونه‌ها بر اساس نرمال به بیمار مرتب می‌شود و پس از آن رنگ‌بندی نمودار برای جداسازی نرمال از بیار و در نهایت رسم نمودار و تعیین راهنما برای آن (مشخص کردن این که هر رنگ مرتبط به کدام نمونه است.)



شکل ۱- نمودار جعبه‌ای نمونه‌های سالم و بیمار. نمونه‌های بیمار با رنگ قرمز و نمونه‌های سالم با رنگ آبی مشخص شده‌اند. نمودار جعبه‌ای نشان دهنده نرمال بودن داده‌ها و عدم نیاز آن‌ها به نرمال‌سازی است. (به دلیل هماهنگی کامل داده‌ها با هم از نظر چارک اول، میانه و چارک سوم و همچنین ماکسیمم و مینیمم‌ها.



شکل ۲- نمودار جعبه‌ای نمونه‌های سالم و بیمار. نمونه‌های بیمار با رنگ قرمز و نمونه‌های سالم با رنگ آبی مشخص شده‌اند. نمودار جعبه‌ای پس از اعمال نرمال‌سازی گسسته نشان دهنده این است که داده‌ها قبل از آن نیز نرمال بودنه‌اند چرا که در مقایسه با شکل۱ تغییر محسوسی مشاهده نمی‌شود. از طرف دیگر شکل۱ خود کاملا مشخص است که نرمال می‌باشد.

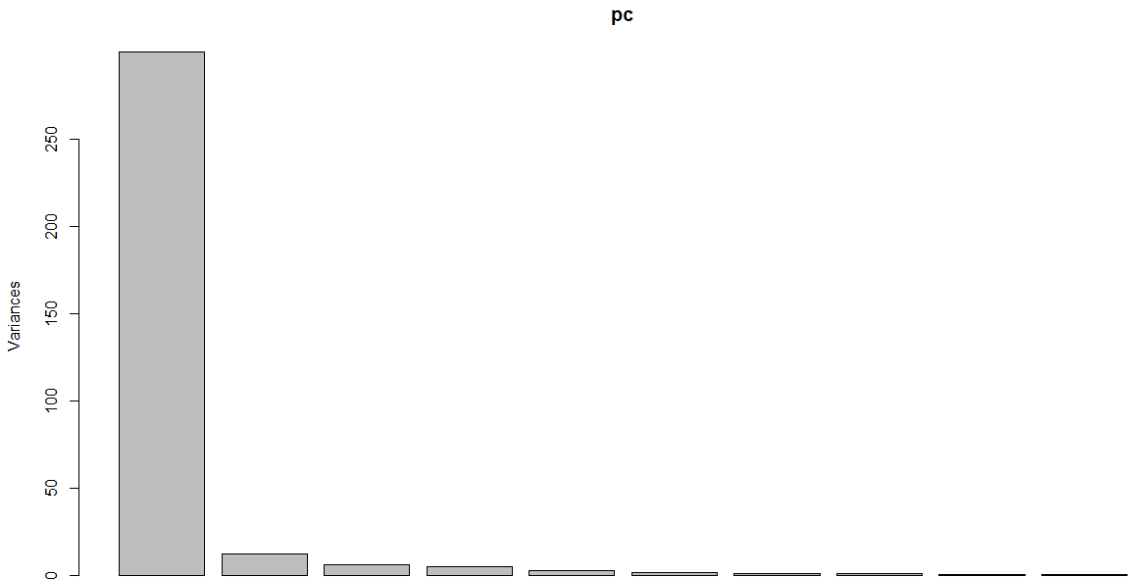
۲-۲. کاهش ابعاد داده

به منظور کاهش ابعاد روش‌های مختلفی وجود دارد، یکی از این روش‌ها PCA[1] می‌باشد. در این روش، چندین جهت مختلف برای محورهای بررسی داده در نظر گرفته می‌شود و در هر کدام از این روش‌ها مقدار تفکیک بین داده‌ها مورد بررسی قرار می‌گیرد. برای این منظور می‌توان از قطعه کد زیر برای بیان ژن‌ها استفاده کرد:

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

داخل متغیر pc تمامی جهت‌هایی که می‌تواند داده را تفکیک کند به‌ترتیب میزان تمایز مرتب شده‌اند و در صورت رسم نمودار آن (شکل۳) می‌توان تفاوت‌های میزان تفکیک در هر کدام از این مولفه‌ها را مشاهده کرد. همانگونه که در شکل۳ مشخص است میزان تمایز در مولفه اول از همه بیشتر و به‌ترتیب کاهش می‌یابد.

[1] PCA: Principal Component Analysis



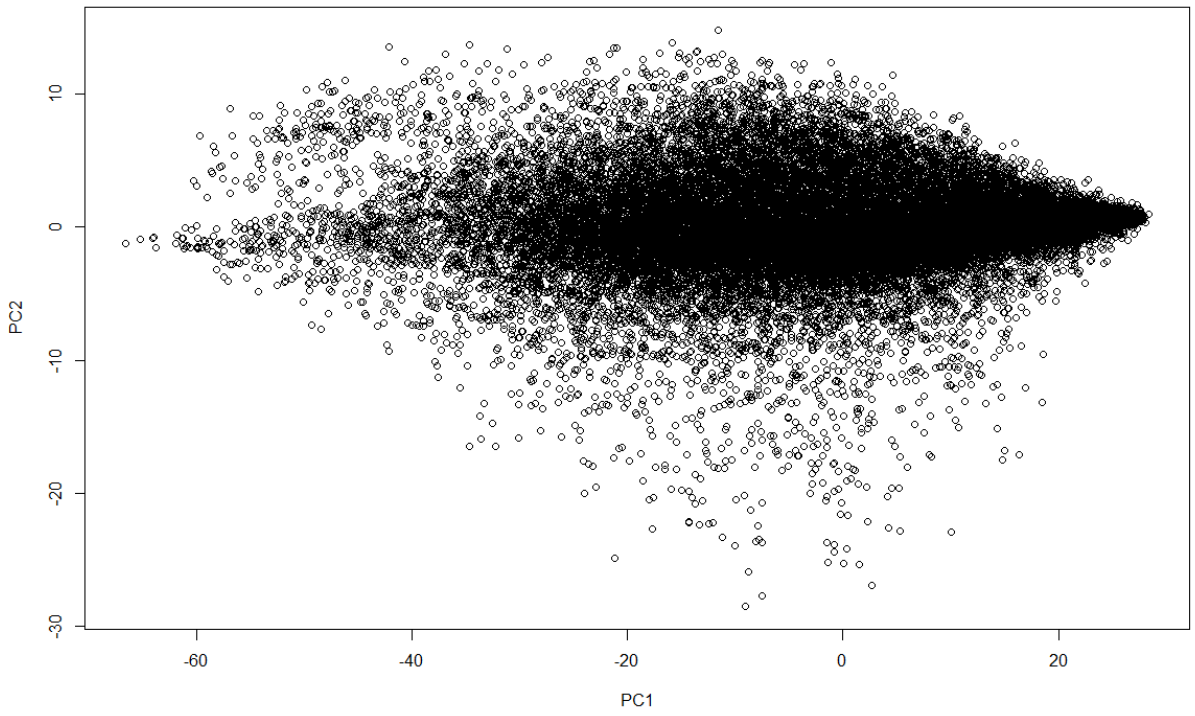
شکل ۳- بررسی هرکدام از مولفه‌های خروجی تابع prcomp نشان دهنده میزان واریانس (تمایز) ایجاد شده میان ژن‌ها در هرکدام از این مولفه‌هاست. به‌ترتیب میزان تمایز از اولین به آخرین مولفه مرتب شده‌است. بنابراین بهترین گزینه برای رسم نمودار کاهش ابعاد مولفه اول و دوم می‌باشد. البته درصورت نیاز به رسم نمودار سه بعدی می‌توان از مولفه سوم نیز کمک گرفت.

درصورت رسم نمودار دو بعدی براساس مولفه اول و دوم که بیشترین تفکیک را ایجاد می‌کنند، مشاهده می‌شود که یک توزیع افقی گسترده برای ژن‌ها وجود دارد.(شکل٤) مولفه اول در شکل٤ نشان دهده میزان بیان ژن‌های مختلف در نمونه می‌باشد. همانگونه که درصورت رسم PCA برای نمونه‌ها مطابق شکل۵، در این نمودار نیز این نمونه‌ها قابل جداسازی از یکدیگر نیستند و این نشان دهنده وجود مشکل در روش تحیلی می‌باشد که باید اصلاح شود. به نوعی این خروجی ارزش زیادی ندارد چرا که هدف یافتن تفاوت بیان‌هاست، برای رفع این مشکل با تغییر ابعاد داده‌ها تاثیر عدم بیان یا بیان بالای ژن‌ها را از بین برد. برای این منظور بیان ژن‌ها از میانگین بیان خودشان کسر می‌شوند (به عبارت دیگر میانگین بیان تمامی ژن‌ها صفر شوند.) تا اثر بیان دائمی یک ژن یا عدم بیان آن از بین برود و فقط تفاوت‌ها ارزشمند شود و مولفه‌های PCA موارد بهتری را نمایش دهند. برای این منظور از طریق کد زیر این تغییر ابعاد انجام می‌شود:

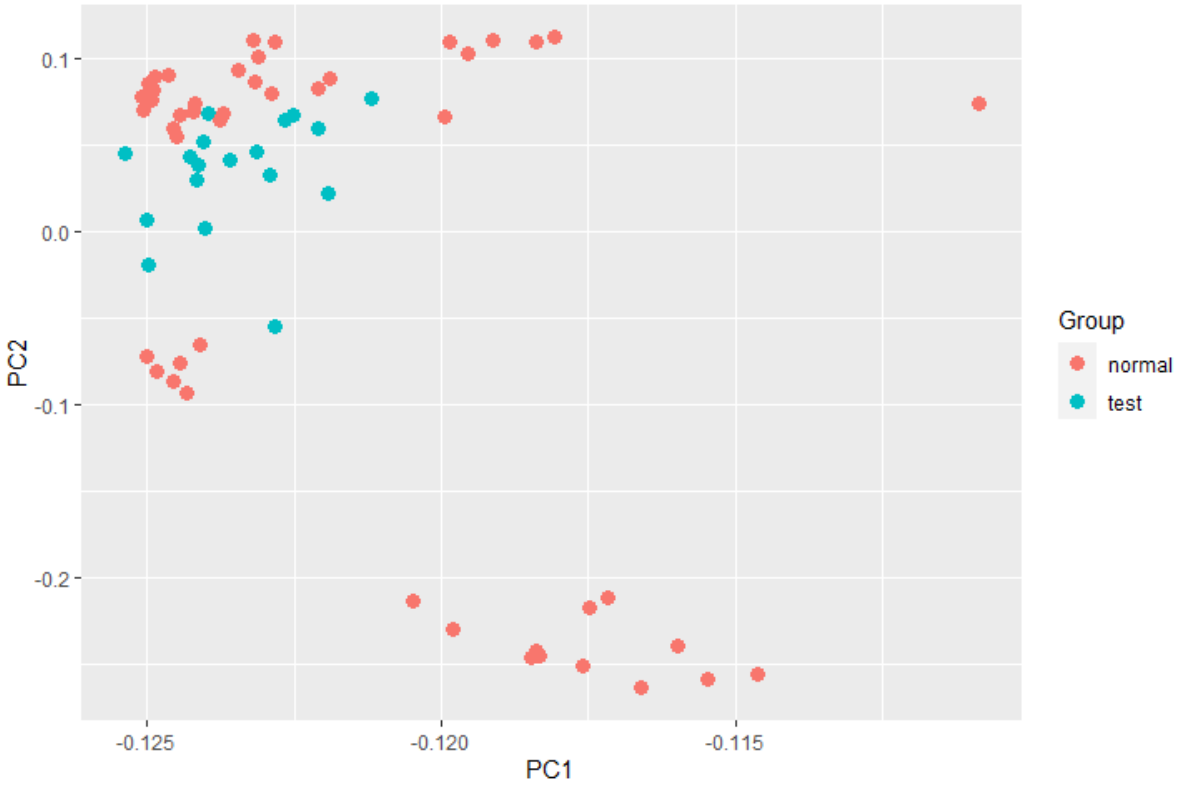
```
ex.scale <- t(scale(t(ex), scale = F))
```

به دلیل این که تابع scale ستون‌ها را تغییر ابعاد می‌دهد، می‌بایست ماتریس بیان راترانهاده کرده و سپس تغییر ابعاد اعمال شود و سپس دوباره به حالت‌ترانهاده اولیه برگردد. به منظور این که تابع scale به صورت پیشفرض داده‌ها را تقسیم بر انحراف معیار نیز می‌کند، می‌بایست این عملیات را از طریق False کردن متغیرscale در تابع، لغو نمود.

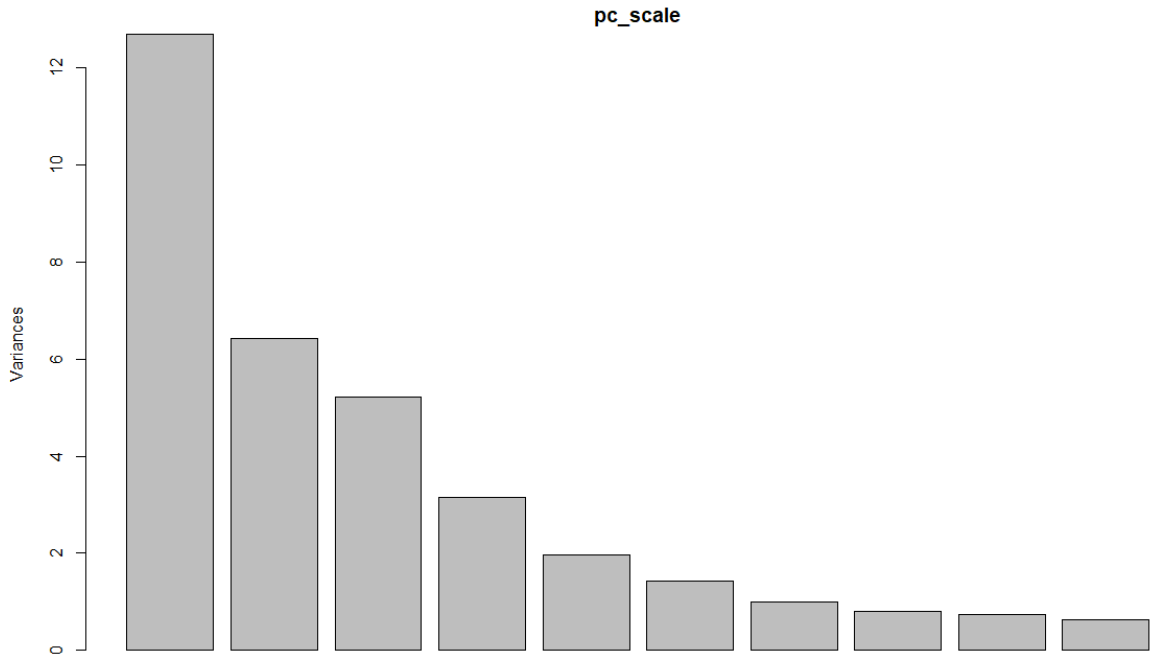
درصورتی که براساس مقادیر جدید، مولفه‌های مختلف روش PCA مورد بررسی قرار داده شود، شکل٦ حاصل می‌شود که تمایزها در مولفه‌های مختلف معقول‌تر شده‌اند و از طرفی در صورت رسم نمونه‌ها مطابق شکل۷ ، می‌توان نتایج را مشاهده نمود که دیگر مولفه اول به سمت بیان کامل یا عدم بیان کامل گرایش ندارد و داده‌ها به صورت متفاوتی قرار گرفته‌اند. به منظور بررسی نمونه‌ها نیز می‌توان نمودار را مطابق روش قبلی برای نمونه‌ها رسم کرد (شکل۸) به این‌ترتیب می‌توان مشاهده کرد که نمونه‌های مختلف تقریبا در کلاسترهای مختلفی قابل جداسازی از یکدیگر هستند و مولفه اول و دوم روش PCA پس از اعمال تغییر ابعاد داده‌ها، به خوبی توانسته است تمایز میان آن‌ها را نمایش دهد. البته بعضی از نمونه‌های سرطانی شباهت زیادی به نمونه‌های سالم دارند ولی با این حال اکثر آن‌ها قابل تمایز هستند.



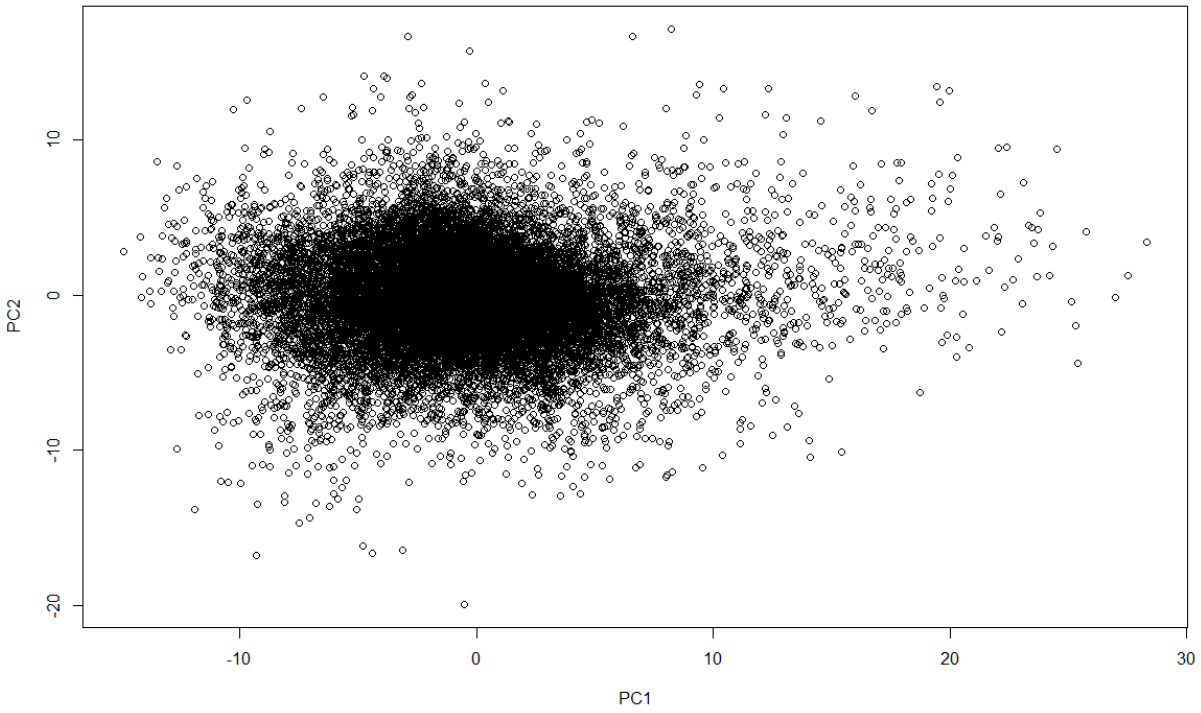
شکل ٤- رسم داده‌ها مبتنی بر مولفه اول و دوم روش PCA. همانطور که در نمودار مشخص است تفاوت‌های موجود در مولفه اول به نوعی صرفا نشان می‌دهد که بعضی از ژن‌ها خیلی کم بیان شده‌اند یا اصلا بیان نشده‌اند و در مقابل بعضی دیگر بسیار زیاد بیان شده‌اند. این به نوعی داده‌ای ناکارآمد محسوب می‌شود.



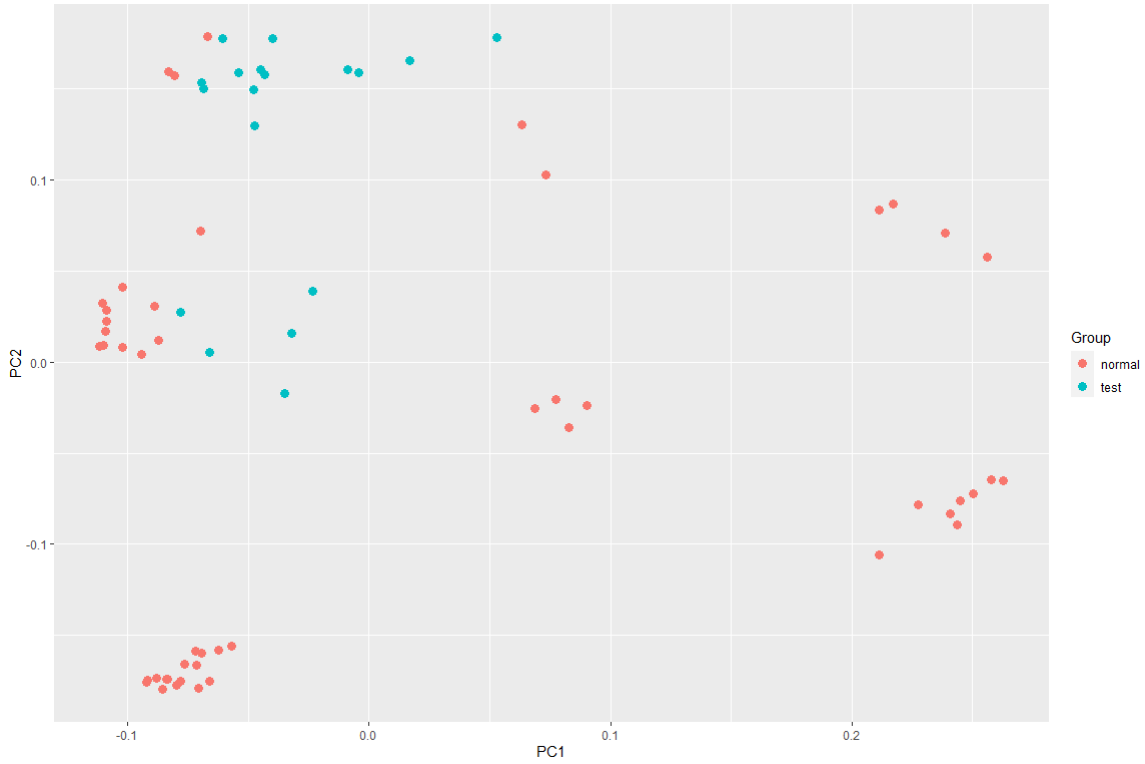
شکل ۵ – تفاوت نمونه‌ها مبتنی بر مولفه‌های اول و دوم به دست آمده از روش PCA، پیش از اعمال تغییر ابعاد.



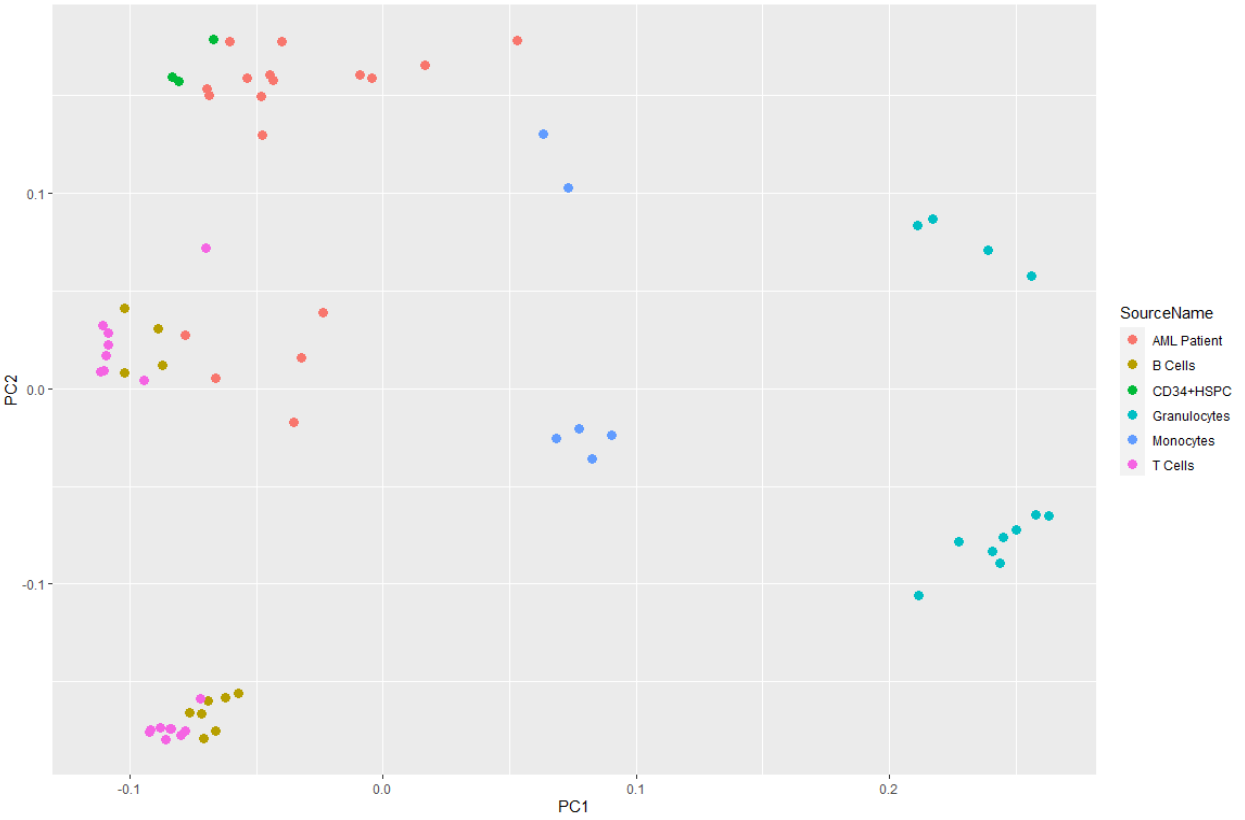
شکل ٦– مولفه‌های مختلف به دست آمده مبتنی بر روش PCA پس از اعمال تغییر ابعاد. همانگونه که مشخص است دیگر تفاوت فاحش میان مولفه اول و دیگر مولفه‌ها نیست و با اعمال تغییر ابعاد، امکان این فراهم شد تا تاثیر بسیار بزرگی که ژن‌هایی که در تمامی نمونه‌ها بیان شده بودند و همچنین ژن‌هایی که اصلا بیان نشده بودند از بین برود و این روش بتواند تمایز بهتری متناسب با تفاوت بیان ژن‌ها ارائه بکند.



شکل ۷- تفاوت بیان ژن‌های مختلف نمونه‌ها مبتنی بر مولفه اول و دوم روش PCA پس از تغییر ابعاد داده. همانگونه که در شکل مشهود است، با تغییر ابعاد داده و از بین بردن تاثیر بیان ژن‌ها و یا عدم بیان آن‌ها در اکثر نمونه‌ها، می‌توان به خروجی بهتری مبتنی بر تفاوت بیان ژن‌ها دست یافت.



شکل ۸- تفاوت نمونه‌ها مبتنی بر مولفه‌های اول و دوم به دست آمده از روش PCA، پس از اعمال تغییر ابعاد نمایش داده شده بر اساس دسته‌های سالم(normal) و بیمار(test).



شکل ۹- تفاوت نمونه‌ها مبتنی بر مولفه‌های اول و دوم به دست آمده از روش PCA، پس از اعمال تغییر ابعاد نمایش داده شده بر اساس دسته‌های source name.

همانگونه که در شکل۹ مشخص است نمونه‌های هر دسته از سلول‌ها تقریبا پراکندگی نزدیک به هم دارند و این نشان دهنده شباهت بین هرکدام از گونه‌هاست و تاییدی بر کیفیت مناسب نمونه‌ها می‌باشد، چرا که درصورتی که این پراکندگی زیاد می‌بود، نمی‌توانست به شکل مناسبی در ادامه به منظور بررسی تمایز ژن‌های بیان شده در نمونه‌های سالم و سرطانی مورد استفاده قرار گیرد.

به منظور رسم نمودارهای ۳ و ٦ قطعه کد زیر مورد استفاده قرار می‌گیرد:

```
plot(pc)
```

به منظور رسم نمودارهای ٤ و ۷ قطعه کد زیر مورد استفاده قرار می‌گیرد:

```
plot(pc$x[,1:2])
```

در قطعه کد بالا، X نشان دهنده ژن‌ها می‌باشد که بر اساس ۲ مولفه اول برای رسم ارسال شده‌است.

به منظور رسم نمودارهای ۵ و ۸ و ۹ از کتاب‌خانه ggplot2 استفاده شده‌است و قطعه کد زیر مورد استفاده قرار گرفته‌است (البته برای نمودار ۹ به منظور نمایش نام هر نمونه، به جای گروه از source name نمونه‌ها استفاده شده‌است.):

```
pcr <- data.frame(pc$r[,1:3], Group = gset$group)
ggplot(pcr, aes(PC1, PC2, color=Group)) + geom_point(size=3)
```

در قطعه کد بالا، مبتنی بر rotation در pc که همان نمونه‌های مختلف هستند، یک دیتافریم به همراه گروه‌های آن‌ها ساخته شده‌است و سپس به ggplot داده شده تا مبتنی بر مولفه اول و دوم و همچنین رنگ‌بندی براساس گروه‌ها، نمودار را رسم نماید.

۲-۳. بررسی همبستگی بین نمونه‌ها

بررسی همبستگی دو به دو نمونه‌ها با یکدیگر با هدف یافتن میزان ارتباط بین نمونه‌های مختلف با یکدیگر می‌باشد. برای این منظور می‌توان از Heatmap استفاده نمود. با توجه به تغییر ابعاد انجام شده در بخش قبل، مبتنی بر خروجی‌های آن اقدام به رسم Heatmap می‌شود. ابتدا ماتریس همبستگی دو به دو برای نمونه‌ها مشخص می‌شود. سپس Heatmap براساس این همبستگی رسم می‌شود.

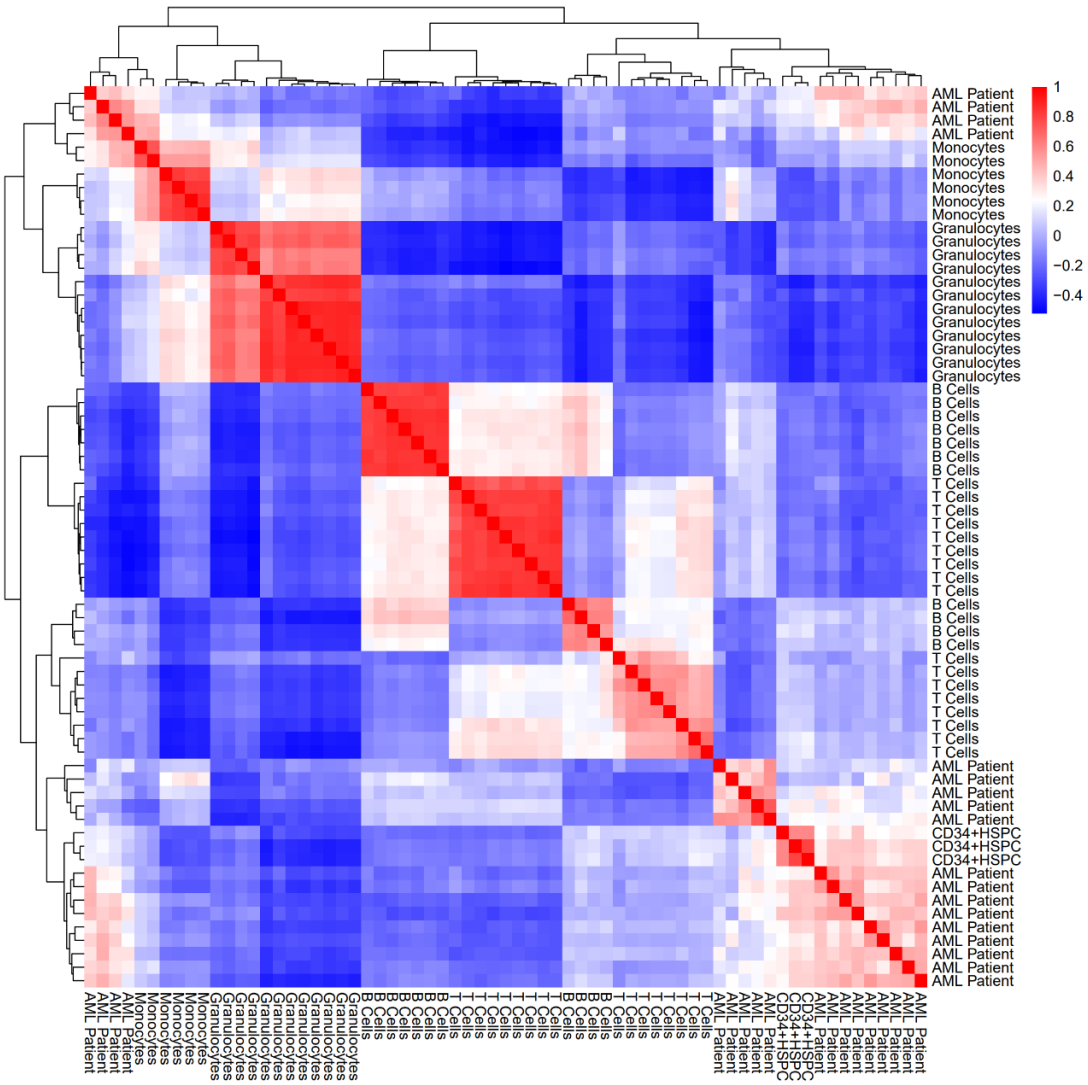
۲-۳-۱. بررسی همبستگی بین تمامی نمونه‌ها

مطابق کد پایین، ابتدا همبستگی نمونه‌ها محاسبه شده و سپس با استفاده از کتاب‌خانه pheatmap اقدام به رسم Heatmap مبتنی بر همبستگی نمونه‌ها می‌شود. رنگ‌بندی نمودار به صورت آبی و قرمز بوده و نامگذاری براساس source_name نمونه‌هاست. نتیجه این نمودار در شکل۹ قابل مشاهده است.

```
ex.scale.cor <- cor(ex.scale)
pheatmap(ex.scale.cor,
         labels_row = gset$source_name_ch1,
         labels_col = gset$source_name_ch1,
         color = bluered(255), border_color = NA)
```

به منظور نمایش بهتر خروجی، Heatmap خروجی در فایل heatmap-all.pdf در دایرکتوری results ذخیره شده‌است. همانگونه که در شکل۹ مشخص است میزان همبستگی در نمونه‌های AML

Patient و نمونه‌های سالم، به صورت کلی چهار نمونه AML Patient و ٦ نمونه Monocytes و ١٢ نمونه Granulocytes از دیگر نمونه‌ها مجزا هستند و همبستگی کمتری به دیگر اعضا دارد، از طرفی در بین این گروه ارتباط AML Patient با یک دیگر نسبتا بالاست و همچنین ارتباط بعضی از Monocytesها با یکدیگر بسیار بالاست و همچنین از بین Granulocytesها بعضی ارتباط بالایی با یکدیگر دارند. در مقابل در دسته پایین، ارتباط T Cellها با خودشان و B Cellها با خودشان بسیار زیاد است و همچنین ارتباط AML Patientها با خودشان و همچنین CD34+HSPCها. به صورت کلی این ارتباطات بالا و در نقطه مقابل عدم ارتباط با دیگر سلول‌ها، نشان دهنده نمونه برداری مناسب و هماهنگ بودن نمونه‌ها با یکدیگر است که این موضوع به نوعی جزئی از کنترل کیفیت داده‌ها محسوب می‌شود.



شکل ١٠- بررسی میزان همبستگی تمامی نمونه‌ها با یکدیگر از طریق Heatmap.

۲-۳-۲. بررسی همبستگی بین نمونه‌های سالم

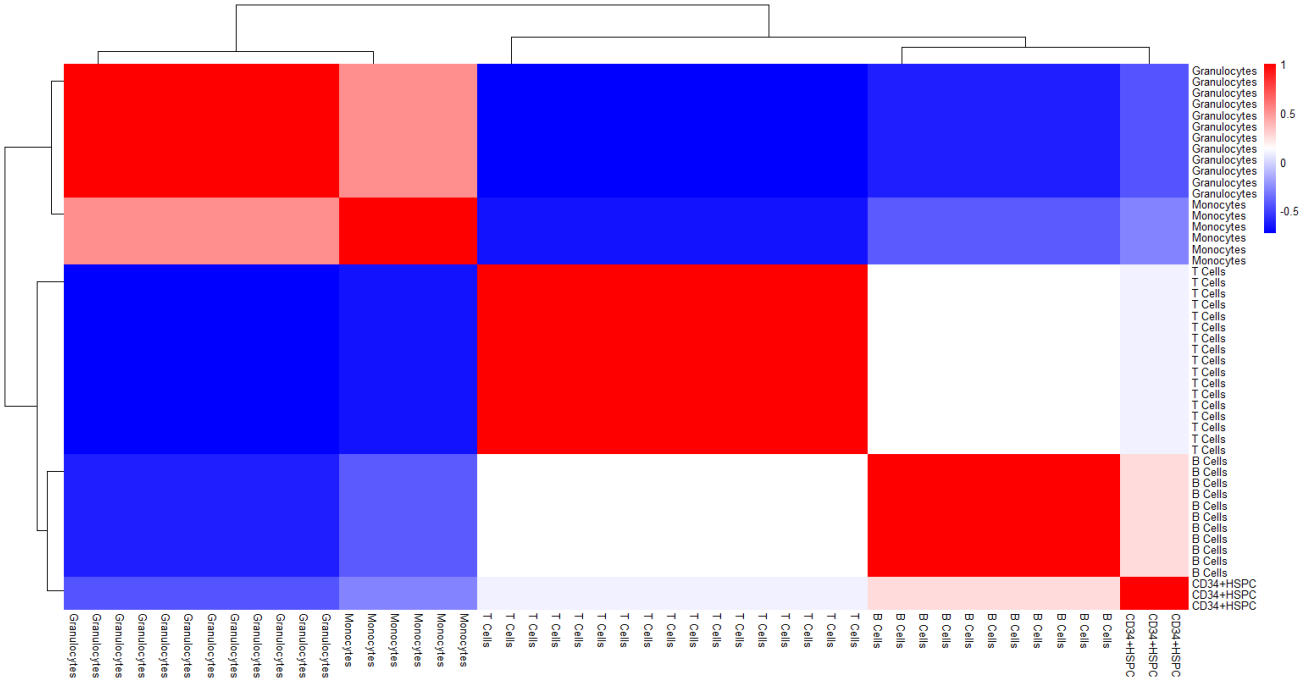
به منظور بررسی همبستگی صرفا بین نمونه‌های سالم می‌بایست، از ماتریس بیان، تمامی نمونه‌های AML Patient را حذف نمود. برای این منظور از طریق قطعه کد زیر اقدام می‌شود:

```
df <- data.frame(ex.scale.cor)
a <- t(df)
colnames(a) <- gset$source_name_ch1
a <- t(a)
colnames(a) <- gset$source_name_ch1
cols <- gset$source_name_ch1
cols <- cols[cols != "AML Patient"]
b <- subset(a, select=cols)
b <- t(b)
cols <- gset$source_name_ch1
cols <- cols[cols != "AML Patient"]
b <- subset(b, select=cols)
b <- t(b)
```

از طریق قطعه کد بالا با دوبارترانهاده کردن ماتریس بیان، تمامی نمونه‌های AML Patient حذف شده و صرفا نمونه‌های نرمال در ماتریس b ذخیره می‌شوند. سپس کافیست همانگونه که در حالت کلی نمودار heatmap رسم شد، در این جا نیز برای نمونه‌های موجود در ماتریس b، همبستگی محاسبه شده و سپس نمودار رسم شود. نمودار خروجی در شکل۱۱ قابل مشاهده است. قطعه کد اجرای محاسبه همبستگی و رسم نمودار در ادامه آمده است.

```
b.cor <- cor(b)
pheatmap(b.cor, color = bluered(255), border_color = NA)
```

به منظور بررسی بهتر شکل۱۱ نمودار به صورت دقیق‌تری در فایل heatmap-normal.pdf موجود می‌باشد. همانگونه که در شکل۱۱ مشخص است، میزان همبستگی بین نمونه‌های Granulocytes و نمونه‌های Monocytes بسیار زیاد است و همچنین میزان همبستگی نمونه‌های B Cells و CD34+HSPC نیز بسیار زیاد است. همچنین بین گونه‌های T Cells و B Cells با گونه‌های Granulocytes و Monocytes همبستگی منفی جدی‌ای وجود دارد که در نمودار هم کاملا مشخص است، البته این هبستگی منفی به میزان همبستگی مثبت گونه‌ها با یکدیگر و با خودشان نیست.



شکل ۱۱ – بررسی میزان همبستگی نمونه‌های سالم با یکدیگر از طریق Heatmap.

نکته قابل توجه در شکل ۱۱ که به کنترل کیفیت داده‌ها مربوط می‌شود، این است که میزان ارتباط بین نمونه‌های مختلف هر سلول با یکدیگر بسیار بالاست و این نشان دهنده کیفیت بسیار خوب داده‌هاست که گونه‌ها از هم مجزا نیستند.

۳. بررسی تمایز در بیان ژن‌ها

به منظور یافتن سلول‌های سالمی که همبستگی بیشتری با سلول‌های بیمار دارند، می‌بایست ماتریس نام سطرها و ستون‌های ماتریس بیان را به نام‌های کلی سلول‌ها (Source Name آن‌ها) تغییر داد و سپس از ستون‌ها تمامی AML Patientها را حذف نمود و همچنین از سطرها به جز AML Patient تمامی دیگر سلول‌ها را حذف نمود. در این حالت در سطرها فقط AML Patient باقی می‌ماند و در ستون‌ها دیگر سلول‌ها و از این طریق با ماکسیمم گیری سطرها می‌توان بیشترین همبستگی‌ها را معین نمود و مرتب کرد. با توجه به توضیحات بالا، کد مد نظر برای اجرای این عملیات به شرح زیر است:

```
df <- data.frame(ex.scale.cor)
a <- t(df)
colnames(a) <- gset$source_name_ch1
a <- t(a)
colnames(a) <- gset$source_name_ch1
```

```
cols <- gset$source_name_ch1
cols <- cols[cols != "AML Patient"]
b <- subset(a, select=cols)
b <- t(b)
cols <- gset$source_name_ch1
cols <- cols[cols == "AML Patient"]
b <- subset(b, select=cols)
b <- t(b)
bB <- b
maxCor <- data.frame(Genes =
unique(colnames(bB)[max.col(bB,ties.method="first")]),
                     CorWithAML = unique(rowMax(bB)))
for (x in 1:4) {
  cols <- colnames(bB)
  cols <- cols[cols != maxCor[[1]][x]]
  bB <- subset(bB, select=cols)
    maxCor <- rbind(maxCor, c(gene =
unique(colnames(bB)[max.col(bB,ties.method="first")]),
                              cor = unique(rowMax(bB))))
}
```

پس از اجرای کد بالا، خروجی یک جدول خواهد بود که بر اساس نام سلول‌ها و میزان همبستگی آن‌ها با AML Patient مرتب‌سازی شده‌است. خروجی جدول۱ قابل مشاهده است.

	نوع سلول	میزان همبستگی با AML Patient (عدد بین -۱ تا ۱)
۱	Monocytes	۰.۳۱۹۷٤
۲	CD34+HSPC	۰.۱۹۷۱۷
۳	B Cells	۰.۰۲۷٤٤
٤	Granulocytes	- ۰.۰۲٤۳۷
٥	T Cells	- ۰.۰٦٤۷٤

جدول ۱- مرتب‌سازی میزان همبستگی هرکدام از سلول‌ها با AML Patient. میزان همبستگی بین -۱ تا ۱ می‌باشد و هرچه به ۱ نزدیک‌تر باشد نشان دهنده همبستگی بیشتر است.

به منظور بررسی میزان تمایز بیان ژن‌ها ابتدا می‌بایست داده‌ها را مشخص کنیم که از کدام گروه هستند و سپس یک مدل خطی به آن‌ها فیت کنیم. این مدل خطی با استفاده از پکیج limma می‌باشد و بسیاری از تفاوت‌ها بین نمونه‌ها را مشخص می‌کند. پس از آن باید مشخص شود که تصمیم بر این است که تفاوت میان کدام گروه‌ها به دست آید و در نهایت یک مدل بیز با prior برابر با ۰.۰۱ به آن نسبت می‌دهد.

براساس خروجی‌های این مدل، جدول میزان تمایز بیان ژن‌ها براساس آماره B که مبتنی بر مدل‌های فیت شده از طریق پکیج limma به دست می‌آید مرتب‌سازی می‌کند. همچنین برای بررسی عدم اتفاق افتادن خطاهای نوع اول و دوم، از روش بنجامینی‌هاچبرگ[1] استفاده می‌شود. جدول خروجی با توجه به داده‌های اولیه‌ای که وجود دارد شامل ستون‌های متعددی است که انواع کدهای ژن و توضیحات آن را شامل می‌شود. به منظور ساده سازی جدول، صرفا بعضی از این پارامترها نگهداری می‌شود تا جدول ساده شود و سپس آن‌ها در یک فایل ذخیره سازی می‌شوند.

پس از این می‌بایست ژن‌هایی که در نمونه اولیه نسبت به نمونه دوم بیان بالایی داشته و همچنین به صورت عکس در نمونه دوم نسبت به نمونه اول بیان بالایی داشته‌است را به دست آورد. برای این منظور با توجه به حد تفاوت معنی‌دار 0.05 که برای adj.P.Val در نظر گرفته می‌شود، علاوه بر این می‌بایست محدودیت دیگری روی logFC گذاشته شود تا این تفاوت را پوشش دهد یعنی بالاتر بودن بیان ویا پایین‌تر بودن بیان در نقطه مقابل، که برای این حد نیز عدد ۱ و در نقطه مقابل عدد −۱ در نظر گرفته شده‌است. مقدار logFC نسبت لگاریتمی سرطانی به سالم می‌باشد که در صورتی که بیشتر از ۱ باشد به این معنی است که بیان آن ژن در نمونه سرطانی بیشتر بوده است و در نقطه مقابل در صورتی که کم‌تر از −۱ باشد به این معنی است که در نمونه سرطانی بیان کم‌تری داشته است و به‌ترتیب هر کدام از این موارد در دسته افزایش و کاهش بیان قرار می‌گیرند. به صورت کلی برای ژن‌هایی که در نمونه AML Patient بیان بیشتری دارند اصطلاح up را به کار می‌بریم و برای ژن‌هایی که کاهش بیان دارند down را به کار می‌بریم. این دو گونه ژن، عامل‌های اصلی ایجاد بیماری خواهند بود که در قسمت‌های بعد مورد بررسی قرار خواهند گرفت.

با توجه به جدول۱، با سه مدل مختلف عملیات بالا انجام شده‌است تا خروجی‌های مختلف را مقایسه کنیم. ابتدا بین کلیه مدل‌های test یعنی AML Patient و تمامی مدل‌های سالم بررسی صورت گرفته‌است و ژن‌هایی که بیان بالا یا پایینی داشته‌اند، به دست آمده است. سپس به‌ترتیب بین AML Patient و Monocytes و سپس بین AML Patient و CD34+HSPC مقایسه صورت گرفته‌است و در هر دسته ژن‌های دارای افزایش بیان و کاهش بیان مشخص شده‌اند.

[1] Benjamini-Hochberg

برای مقایسه کلی بین سلول‌های بیمار و سالم:

```
##### based on all test-normal #####
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)
## fitting linear model to data
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(test-normal, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
## adjust by: false-discovery-rate or Benjamini-Hochberg
## sort by adj.P.Val
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
tT <- subset(tT,
select=c("Gene.symbol","Gene.ID","adj.P.Val","logFC", "B"))
write.table(tT, "result/dea/dea_test-normal_B.txt", row.names=F,
sep="\t", quote=F)
### Top Gene Expression mining
aml.up <- subset(tT, logFC > 1 & adj.P.Val < 0.05)
aml.up.genes <-
unique(as.character(strsplit2(unique(aml.up$Gene.symbol), "///")))
write.table(aml.up.genes, "result/dea/dea_test-normal_Up.txt",
            quote=F, row.names=F, col.names=F)
aml.down <- subset(tT, logFC < -1 & adj.P.Val < 0.05)
aml.down.genes <-
unique(as.character(strsplit2(unique(aml.down$Gene.symbol),
"///")))
write.table(aml.down.genes, "result/dea/dea_test-normal_Down.txt",
            quote=F, row.names=F, col.names=F)
```

برای مقایسه بین نمونه‌های بیمار و به صورت مجزا با هر نوع سلول، که برای دو نوع Monocytes و CD34+HSPC که میزان همبستگی بیشتری با نمونه‌های بیمار داشتند، صورت گرفته‌است:

```
##### based on top correlated cells #####
design <- model.matrix(~source_name_ch1 + 0, gset)
sfl <- factor(sname.gs)
colnames(design) <- levels(sfl)
## fitting linear model to data
fit <- lmFit(gset, design)
```

```
###### AMLPatient-Monocytes ######
cont.matrix <- makeContrasts(AMLPatient-Monocytes, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
## adjust by: false-discovery-rate or Benjamini-Hochberg
## sort by adj.P.Val
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
tT <- subset(tT, select=c("Gene.symbol",
"Gene.ID","adj.P.Val","logFC", "B"))
write.table(tT, "result/dea/dea_AMLPatient-Monocytes_B.txt",
            row.names=F, sep="\t", quote=F)
### Top Gene Expression mining
aml.up <- subset(tT, logFC > 1 & adj.P.Val < 0.05)
aml.up.genes <-
unique(as.character(strsplit2(unique(aml.up$Gene.symbol), "///")))
write.table(aml.up.genes, "result/dea/dea_AMLPatient-
Monocytes_Up.txt",
            quote=F, row.names=F, col.names=F)
aml.down <- subset(tT, logFC < -1 & adj.P.Val < 0.05)
aml.down.genes <-
unique(as.character(strsplit2(unique(aml.down$Gene.symbol),
"///")))
write.table(aml.down.genes, "result/dea/dea_AMLPatient-
Monocytes_Down.txt",
            quote=F, row.names=F, col.names=F)
###### AMLPatient-CD34+HSPC ######
cont.matrix <- makeContrasts(AMLPatient-CD34pHSPC, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
## adjust by: false-discovery-rate or Benjamini-hochberg
## sort by adj.P.Val
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
tT <- subset(tT, select=c("Gene.symbol",
"Gene.ID","adj.P.Val","logFC", "B"))
write.table(tT, "result/dea/dea_AMLPatient-CD34pHSPC_B.txt",
            row.names=F, sep="\t", quote=F)
### Top Gene Expression mining
aml.up <- subset(tT, logFC > 1 & adj.P.Val < 0.05)
aml.up.genes <-
unique(as.character(strsplit2(unique(aml.up$Gene.symbol), "///")))
```

```
write.table(aml.up.genes, "result/dea/dea_AMLPatient-
CD34pHSPC_Up.txt",
            quote=F, row.names=F, col.names=F)
aml.down <- subset(tT, logFC < -1 & adj.P.Val < 0.05)
aml.down.genes <-
unique(as.character(strsplit2(unique(aml.down$Gene.symbol),
"///")))
write.table(aml.down.genes, "result/dea/dea_AMLPatient-
CD34pHSPC_Down.txt",
            quote=F, row.names=F, col.names=F)
```

هرکدام از ژن‌هایی که افزایش یا کاهش بیان داشته‌اند در فایل‌های مجزا در دایرکتوری result/dea/ ذخیره شده‌است؛ همچنین به صورت یکپارچه ژن‌های هر دسته در فایل DifferentialExpressionAnalysis.xlsx قابل مشاهده است.

٤. آنالیز Gene ontology و pathway ها

ژن‌های به دست آمده در قسمت قبل شامل سه دسته مقایسه زیر هستند:

- مقایسه AMLها با تمامی نمونه‌های سالم
- مقایسه AMLها با نمونه‌های Monocytes که بیشتری میزان همبستگی را با نمونه‌های AML دارند
- مقایسه AMLها با نمونه‌های CD34+HSPC که بعد از Monocytes بیشترین میزان همبستگی را با نمونه‌های AML دارند.

در ادامه به بررسی ژن‌هایی که افزایش یا کاهش بیان در دسته مقایسه AML و Monocytes و همچنین AML و CD34 داشته‌اند، پرداخته شده‌است و pathwayها و Gene ontology مربوط به هر دسته به کمک وبسایت Enrichr مورد بررسی قرار گرفته‌اند.

٤-١. بررسی pathwayها

منظور از pathwayها مجموعه‌ای از فرآیندهای سلول و مولکولی است که عملکردهای سلول را تشکیل می‌دهند. در این قسمت با بررسی ژن‌های به دست آمده در قسمت‌های قبل و مقایسه آن‌ها با فرآیندهای مختلف سلولی و مولکولی بررسی می‌شود که این ژن‌ها در چه مسیرها و فرآیندهایی موثراند.

٤-١-١. بررسی pathwayهای مرتبط با دسته AML در مقابل Monocytes که افزایش بیان داشته‌اند

با بررسی ژن‌هایی که افزایش بیان در نمونه‌های بیمار داشته‌اند در مقایسه با Monocytesها(سالم)، ۱٤۸۹ ژن به دست می‌آید. در مقایسه با pathwayهای کتاب‌خانه Kegg[1] بیشترین مواردی که به عنوان ارتباط یافته شده‌است که با مقدار adj.P.Val بسیار پایینی هستند، موارد مرتبط با فرآیند تقسیم سلول است که شامل خود فرآیند تقسیم سلول، فرآیند رونویسی DNA و همچنین فرآیند غلط گیری از رونویسی DNA می‌باشد؛ این موارد شامل Cell Cycle ، DNA Replication، Fanconi anemia pathway ، Mismatch repair را می‌توان نام برد که همه در ارتباط با شروع و ادامه فرآیند تکثیر سلول هستند. (جدول۲ شامل ۱۳ مورد از کمترین adj.P.Valهای مرتبط با ژن‌های یافته شده در قسمت قبل می‌باشد.) ایجاد جهش‌هایی در ژن‌هایی که این فرآیندها را کنترل و مدیریت می‌کنند باعث بروز خطا در تکثیر سلول می‌شود. همچنین از جمله مواردی که در pathwayهای به دست آمده وجود دارد p53 است که یکی از ژن‌های مهم کنترل کننده و جلوگیری کننده از سرطان می‌باشد، با توجه به adj.P.Valهای به دست آمده نشان می‌دهد ژن‌های به دست آمده بر روی عملکرد این ژن موثر بوده و باعث بروز عملکرد نادرست این ژن شده و در نتیجه آن یکی از مکانیزم‌های مهم دفاعی سلول در مقابل سرطانی شدن خود دچار مشکل جدی شده‌است. از دیگر pathwayهای موثر یافته شده فرآیند اشتباه رونویسی در سرطان[2] می‌باشد که باعث بروز خطا در عملکرد فاکتورهای رونویسی و کنترل کننده‌های آن می‌شود، در این مورد در AML جهش‌های انجام شده در برخی ژن‌ها از جمله IL3 ، MPO باعث خطا در عملکردترجمه می‌شود.(شکل۱۲-قسمت سمت چپ بالا) (تصاویر برخی از pathwayهای مهم نام برده شده در دایرکتوری result/pathways/amu ضمیمه شده‌است.)

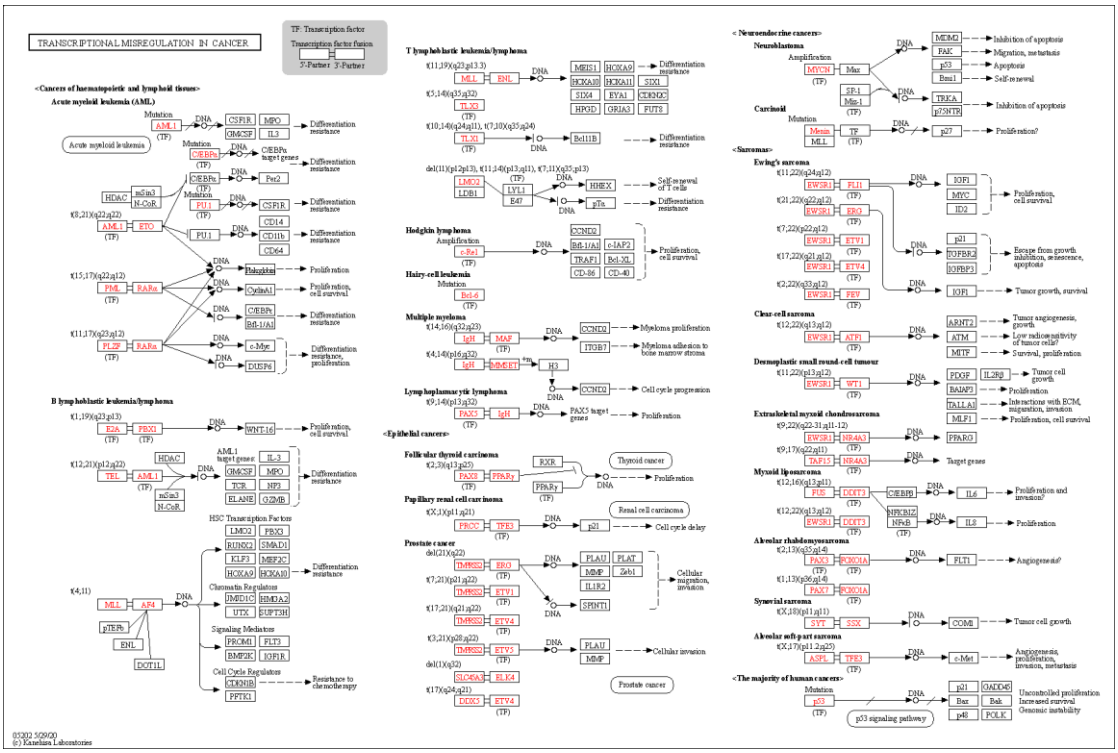
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cell cycle	1.96E-18	5.77E-16	6.77	275.87
2	DNA replication	1.11E-11	1.63E-09	12.57	317.1
3	Fanconi anemia pathway	0.000001149	0.0001126	5.28	72.22
4	p53 signaling pathway	0.000004729	0.0003475	4.11	50.35
5	Pyrimidine metabolism	0.000009816	0.0005772	4.58	52.86
6	Transcriptional misregulation in cancer	0.00001381	0.0006767	2.52	28.19

[1] KEGG: www.genome.jp/kegg/
[2] Transcriptional misregulation in cancer

7	Mismatch repair	0.00002148	0.0009023	8.03	86.36
8	Progesterone-mediated oocyte maturation	0.00004064	0.001494	3.14	31.71
9	Oocyte meiosis	0.00007473	0.002441	2.72	25.88
10	Base excision repair	0.00009597	0.002821	5.43	50.28
11	Hematopoietic cell lineage	0.0001123	0.003002	2.98	27.08
12	Human T-cell leukemia virus 1 infection	0.0008744	0.02142	1.99	14.04
13	Cellular senescence	0.001252	0.02831	2.17	14.49

جدول ۲- مجموعه‌ای از مرتبط‌ترین pathwayها با ژن‌های افزایش بیان یافته در سلول‌های بیمار در کتاب‌خانه Kegg.

مشابه مقایسه بالا در کتاب‌خانه WikiPathway[1] نیز انجام شده‌است و فایل جدول آن در دایرکتوری result/pathways/amu ضمیمه شده‌است.

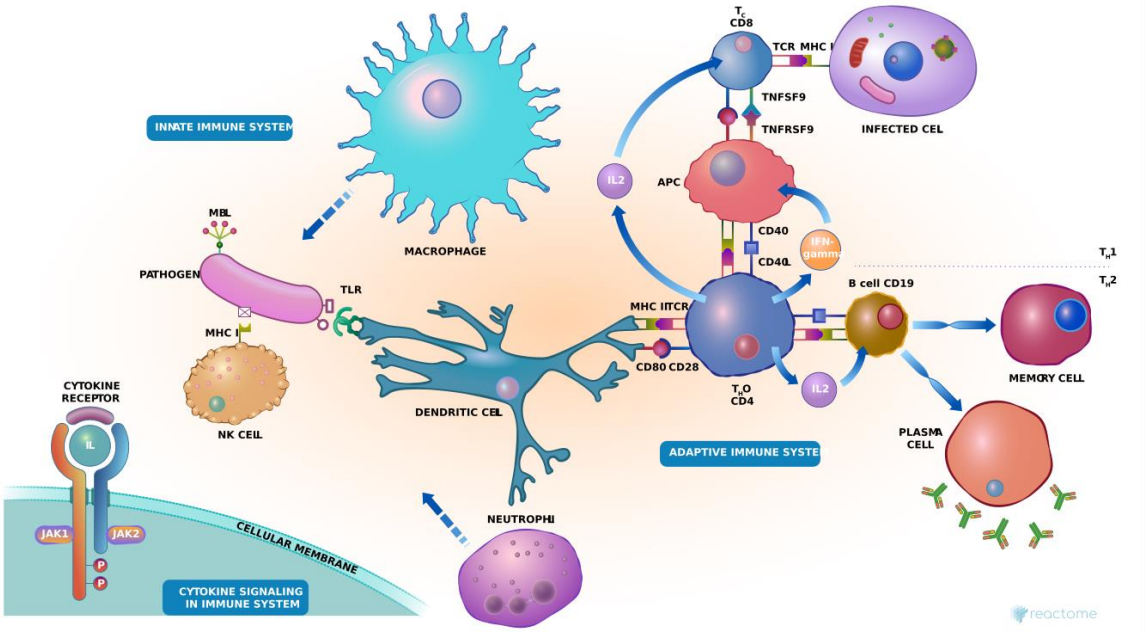


شکل ۱۲- برخی از جهش‌هایی که در اثر سرطان باعث ایجاد خطا در فرآیند رونویسی و کنترل آن می‌شود. سمت چپ بالا مربوط به AML می‌باشد.

[1] WikiPathway: www.wikipathways.org

٤-١-٢. بررسی pathwayهای مرتبط با دسته AML در مقابل Monocytes که کاهش بیان داشته‌اند

با بررسی ژن‌هایی که افزایش بیان در نمونه‌های بیمار داشته‌اند در مقایسه با Monocytesها(سالم)، ١٤١٥ ژن به دست می‌آید. در مقایسه با pathwayهای کتاب‌خانه Reactome[1] بیشترین مواردی که به عنوان ارتباط یافته شده‌است که با مقدار adj.P.Val بسیار پایینی هستند، همانگونه که در جدول٣ مشخص است هر ٨ مورد اول آن و بسیاری از موارد دیگر آن مرتبط با عملکرد سیستم ایمنی می‌باشد. از عملکرد ذاتی سیستم ایمنی تا آنزیم‌هایی مثل کیناز و ... که عملکرد ماکروفاژها و سلول‌های مختلف را تحت تاثیر قرار می‌دهد تحت تاثیر ژن‌های یافته شده هستند و از مقادیر adj.P.Val می‌توان فهمید که ارتباط بسیار زیادی با یکدیگر دارند. این ارتباط باعث می‌شود در نتیجه کاهش بیان این ژن‌ها، مشکلات جدی عملکردی در سیستم ایمنی ایجاد شده و در نتیجه نتواند فعالیت‌های غیرطبیعی سلول‌ها را کنترل و مدیریت نماید. نمونه‌های مختلفی از pathwayهای سیستم ایمنی همانند شکل١٣ در دایرکتوری result/pathways/amd ضمیمه شده‌است.



شکل ١٣- تصویری از pathway مرتبط با سیستم ایمنی (Immune System Homo sapiens R-HSA-168256) از کتاب‌خانه Reactome.

[1] Reactome: www.reactome.org

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Immune System Homo sapiens R-HSA-168256	1.30E-43	1.42E-40	3.08	303.76
2	Interferon gamma signaling Homo sapiens R-HSA-877300	4.20E-24	2.29E-21	11.12	598.37
3	Innate Immune System Homo sapiens R-HSA-168249	6.66E-22	2.41E-19	2.84	138.42
4	Cytokine Signaling in Immune system Homo sapiens R-HSA-1280215	1.53E-21	4.16E-19	3.13	150.05
5	Interferon Signaling Homo sapiens R-HSA-913531	7.92E-21	1.72E-18	5.57	257.81
6	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell Homo sapiens R-HSA-198933	9.93E-16	1.80E-13	5.25	181.24
7	Interferon alpha/beta signaling Homo sapiens R-HSA-909733	3.00E-15	4.66E-13	9.36	312.97
8	Adaptive Immune System Homo sapiens R-HSA-1280218	1.27E-14	1.72E-12	2.43	77.62
9	Toll-Like Receptors Cascades Homo sapiens R-HSA-168898	5.24E-14	6.33E-12	5.19	158.62
10	Toll Like Receptor 4 (TLR4) Cascade Homo sapiens R-HSA-166016	2.59E-10	2.82E-08	4.55	100.48

جدول ۳- عملکردهای مرتبط با ژن‌های کاهش بیان یافته در AML نسبت به Monocytes. جدول تکمیلی موارد فوق در فایل reactome_AML_Monocytes_Down.xlsx در دایرکتوری result/pathways/amd ضمیمه شده‌است.

٤-١-٣. بررسی pathwayهای مرتبط با دسته AML در مقابل CD34+HSPC که افزایش بیان داشته‌اند

با بررسی ژن‌هایی که افزایش بیان در نمونه‌های بیمار داشته‌اند در مقایسه با CD34 ها(سالم)، ۸۵۹ ژن به دست می‌آید. در مقایسه با pathwayهای کتاب‌خانه Reactome بیشترین مواردی که به عنوان ارتباط، یافته شده‌است که با مقدار adj.P.Val بسیار پایینی هستند، موارد مرتبط با فرآیندهای سیستم ایمنی ذاتی، جلوگیری از خونریزی و انعقاد خون، فرآیندهای تطبیقی سیستم ایمنی و ساخت آنتی بادی‌های مناسب، همچنین فرآیندهای GTPase که تحت تاثیر این ژن‌ها می‌باشد و موارد متعدد دیگر که برخی از این موارد در جدول٤ لیست شده‌است. به عنوان مثال فرآیند انعقاد خون در شکل١٤، هموستاز یک پاسخ فیزیولوژیکی است که با توقف خونریزی از رگ آسیب دیده به اوج خود می‌رسد. در شرایط عادی، اندوتلیوم عروقی از گشاد شدن عروق پشتیبانی می‌کند، چسبندگی و فعال شدن پلاکت‌ها را مهار می‌کند و به تبع انعقاد را سرکوب می‌کند، شکاف فیبرین را افزایش می‌دهد و خاصیت ضد التهابی دارد. تحت ترومای حاد عروقی(پارگی رگ

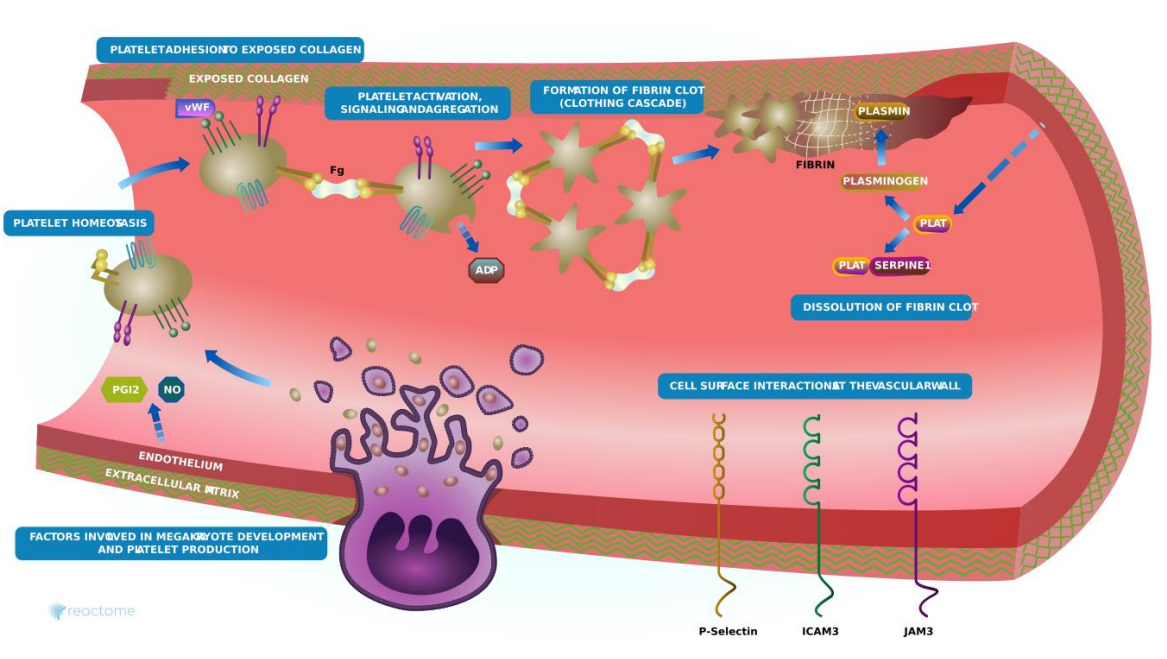
یا عواملی که باعث ایجاد ضعف دیواره رگ شود) ، مکانیسم‌های منقبض کننده عروق غالب می‌شوند و اندوتلیوم ماهیتی پیش‌انعقادی و پیش‌التهابی پیدا می‌کند. و با استفاده از پلاکت‌ها و مواد آزاد درون خون ساختارهای غیر حل شونده در خون تولید می‌کنند تا بتوانند دیواره رگ را ترمیم کنند[1]. این فرآیند به صورت کلی جزئی از فرآیند انعقاد خون می‌باشد. ژن‌هایی که افزایش بیان داشته اند به صورت جدی‌ای در این فرآیند موثراند.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Immune System Homo sapiens R-HSA-168256	2.94E-22	3.18E-19	2.71	134.41
2	Hemostasis Homo sapiens R-HSA-109582	1.48E-13	5.34E-11	3.14	92.63
3	Adaptive Immune System Homo sapiens R-HSA-1280218	1.13E-13	5.34E-11	2.78	82.86
4	Cell Cycle, Mitotic Homo sapiens R-HSA-69278	7.20E-13	1.95E-10	3.29	91.93
5	Cell Cycle Homo sapiens R-HSA-1640170	1.45E-12	3.13E-10	2.99	81.49
6	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell Homo sapiens R-HSA-198933	4.80E-12	8.64E-10	5.42	141.21
7	Innate Immune System Homo sapiens R-HSA-168249	9.87E-11	1.53E-08	2.44	56.32
8	Signaling by Rho GTPases Homo sapiens R-HSA-194315	7.95E-10	1.07E-07	3.15	65.91
9	G0 and Early G1 Homo sapiens R-HSA-1538133	3.68E-08	0.000004423	15.02	257.07
10	Cross-presentation of particulate exogenous antigens (phagosomes) Homo sapiens R-HSA-1236973	1.60E-07	0.00001577	67.31	1053.12

جدول ٤- بررسی pathwayهای مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های بیمار نسبت به CD34 مبتنی بر کتاب‌خانه Reactome. جدول تکمیلی در فایل reactome_aml_cd34_up.xlsx در دایرکتوری result/pathways/acu ضمیمه شده‌است.

فرآیندهای متعددی دیگری نیز تحت تاثیر این ژن‌ها هستند که در جدول لیست شده‌اند. از جمله آن‌ها GTPase است که در فرآیند تولید پروتئین‌ها و مواردی که نیاز به انرژی است مورد استفاده قرار می‌گیرد و همچنین فرآیندهای تکثیر سلولی که در قسمت‌های قبل نیز با ژن‌های به دست آمده ارتباط جدی‌ای داشته‌اند. (تصاویر برخی از این pathwayها در دایرکتوری result/pathways/acu ضمیمه شده‌است.)

[1] https://reactome.org/content/detail/R-HSA-109582



شکل ١٤- فرآیند مرتبط با انعقاد خون در کتاب‌خانه Reactome.

همچنین با بررسی کتاب‌خانه Kegg برخی از pathwayهای مهم نیز تحت تاثیر این ژن‌ها یافت می‌شود. از جمله Lysosome را می‌توان نام برد که با adj.P.Val برابر با 0.00002571 با این ژن‌ها در ارتباط است. فرآیند مربوط به Lysosome درون سلول بسیار حیاتی است که به نوعی وظیفه شناخت و دفع زباله‌های سلول را به عهده دارد. این زباله‌ها شامل پروتئین‌های اشتباه تا خورده یا اشتباه تولید شده می‌باشد که یکی از عوامل مهم در ایجاد و گسترش سرطان می‌باشد. جدول تکمیلی این کتاب‌خانه در فایل keg_aml_cd34_up.xlsx در دایرکتوری result/pathways/acu ضمیمه شده‌است.

٤-١-٤. بررسی pathwayهای مرتبط با دسته AML در مقابل CD34+HSPC که کاهش بیان داشته‌اند

با بررسی ژن‌هایی که افزایش بیان در نمونه‌های بیمار داشته‌اند در مقایسه با CD34 ها(سالم)، ٨٤٥ ژن به دست می‌آید. در مقایسه این ژن‌ها با کتاب‌خانه‌های pathway در Enrichr در کتاب‌خانه‌های kegg ، Reatome ، WikiPathway در adj.P.Val آن‌ها تفاوت معنی‌داری دیده نمی‌شود.

تنها کتاب‌خانه‌ای که تفاوت‌های معنی‌داری در آن مشاهده می‌شود، ARCHS4 Kinases Coexp می‌باشد؛ در بین خروجی‌های این کتاب‌خانه ۱۰ مورد اول آن تفاوت‌های معنی‌دار وجود دارد. (جدول۵) این موضوع نشان دهنده تاثیر ژن‌های کاهش بیان یافته در این موارد می‌باشد.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	PRKG2 human kinase ARCHS4 coexpression	1.63E-08	0.000007959	3.2	57.31
2	PTK2 human kinase ARCHS4 coexpression	0.000004099	0.0006654	2.68	33.29
3	TIE1 human kinase ARCHS4 coexpression	0.000004099	0.0006654	2.68	33.29
4	PAK5 human kinase ARCHS4 coexpression	0.00001117	0.00136	2.58	29.47
5	CASK human kinase ARCHS4 coexpression	0.00002934	0.002858	2.49	25.94
6	EPHB1 human kinase ARCHS4 coexpression	0.00007431	0.006031	2.39	22.7
7	MAPK10 human kinase ARCHS4 coexpression	0.0001812	0.008823	2.29	19.74
8	BMPR1A human kinase ARCHS4 coexpression	0.0001812	0.008823	2.29	19.74
9	EPHA7 human kinase ARCHS4 coexpression	0.0001812	0.008823	2.29	19.74
10	TEK human kinase ARCHS4 coexpression	0.0001812	0.008823	2.29	19.74

جدول ۵- بررسی ژن‌های کاهش بیان داشته در نمونه‌های بیمار نسبت به CD34 مبتنی بر کتاب‌خانه ARCHS4 Kinases Coexp.

به عنوان نمونه با بررسی PRKG2 در archs4 ، این ژن پروتئینی را کد می‌کند که متعلق به خانواده پروتئین‌های کیناز سرین است. پروتئین کدگذاری شده به چندین گیرنده تیروزین کیناز متصل می‌شود و از فعال شدن آن‌ها جلوگیری می‌کند. پروتئین متصل به غشاء، تنظیم کننده ترشح روده و رشد استخوان است[1].

٤-۲. بررسی Gene ontology

در این قسمت به بررسی اثر گذاری ژن‌های افزایش یا کاهش بیان داشته در بیماران نسبت به دسته‌های مختلف در پروسه‌های زیستی، عملکردهای مولکولی و کامپونت‌های سلولی پرداخته شده‌است. در ادامه در دو دسته مختلف مقایسه بین AML و Monocytes همچنین AML و CD34 صورت گرفته‌است که در هرکدام ژن‌های کاهش یا افزایش یافته به صورت مجزا بررسی شده‌اند.

[1] https://www.ncbi.nlm.nih.gov/gene/5593

٤-٢-١. بررسی Gene ontology مرتبط با دسته AML در مقابل Monocytes که افزایش بیان داشته اند

با درج ژن‌های افزایش بیان داشته در نمونه‌های سرطانی نسبت به Monocytes در وبسایت Enrichr و بررسی Ontologyهای ارائه شده؛ این ژن‌ها در فرآیندهای متعددی اثر بالایی دارند و همانگونه که انتظار می‌رود تعداد زیادی از موارد که دارای adj.P.Val بسیار کوچکی هستند، در ارتباط با فرآیند کپی سازی DNA و رفع اشکالات کپی سازی و همچنین به صورت کلی تقسیم سلولی می‌باشند. جدول٦ شامل پروسه‌های زیستی‌ای می‌باشد که بیشتری ارتباط را با ژن‌های یافته شده دارند.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	mitotic DNA replication (GO:1902969)	6.41E-10	1.32E-07	112.56	2382.65
2	double-strand break repair via break-induced replication (GO:0000727)	2.92E-10	7.20E-08	62.57	1373.73
3	DNA strand elongation involved in DNA replication (GO:0006271)	3.49E-13	1.85E-10	43.91	1259.62
4	mitotic DNA replication initiation (GO:1902975)	0.00001279	0.0008449	62.37	702.64
5	nuclear cell cycle DNA replication initiation (GO:1902315)	0.00001279	0.0008449	62.37	702.64
6	microtubule cytoskeleton organization involved in mitosis (GO:1902850)	1.09E-27	4.01E-24	9.38	582.11
7	DNA replication initiation (GO:0006270)	1.50E-14	1.11E-11	15.56	495.41
8	mitotic spindle elongation (GO:0000022)	0.000004142	0.0002978	37.44	464.07
9	mitotic spindle midzone assembly (GO:0051256)	0.000004142	0.0002978	37.44	464.07
10	pre-replicative complex assembly involved in nuclear cell cycle DNA replication (GO:0006267)	0.000004142	0.0002978	37.44	464.07

جدول ٦- پروسه‌های زیستی مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های سرطانی نسبت به Monocytes. جدول تکمیلی در فایل BiologicalProcess_AML_Monocytes_Up.xlsx در دایرکتوری result/ontology/amu ضمیمه شده‌است.

اکثر فرآیندهای بالا مرتبط با عملیات‌های کپی برداری از DNA می‌باشد که این نشان دهده ایجاد تغییرات در این فرآیندها در طی ابتلا به AML می‌باشد.

با بررسی عملکردهای مولکولی، می‌توان دریافت که ژن‌های افزایش بیان داشته با مولکول‌های مرتبط با کپی برداری DNA از جمله مولکول‌های متصل شونده به نقطه شروع کپی برداری، مولکول‌های دخیل در DNA Polymerase و همچنین اتصالات بین مولکولی در پروتئین‌ها و DNA و همچنین آبکافت

ATP مرتبط‌اند. به همین دلیل با افزایش بیان این ژن‌ها، این عملکردها دچار مشکلات متعددی می‌شوند. (جدول۷ مهم‌ترین عملکردهای مولکولی مرتبط با ژن‌های یافته شده را نشان می‌دهد.)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	DNA replication origin binding (GO:0003688)	6.53E-11	4.53E-08	19.51	457.6
2	single-stranded DNA helicase activity (GO:0017116)	2.52E-07	0.00008713	13.9	211.22
3	four-way junction DNA binding (GO:0000400)	9.63E-07	0.0002226	14.06	194.84
4	DNA polymerase binding (GO:0070182)	0.000001799	0.0002603	12.5	165.37
5	DNA secondary structure binding (GO:0000217)	0.000002254	0.0002603	7.15	93.02
6	single-stranded DNA binding (GO:0003697)	0.000001972	0.0002603	3.69	48.43
7	5'-flap endonuclease activity (GO:0017108)	0.000004142	0.0004101	37.44	464.07
8	flap endonuclease activity (GO:0048256)	0.00001164	0.001008	24.96	283.58
9	3'-5' DNA helicase activity (GO:0043138)	0.0000785	0.006044	9.71	91.78
10	RNA-DNA hybrid ribonuclease activity (GO:0004523)	0.0001439	0.009068	49.86	441.05

جدول ۷- عملکردهای مولکولی مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های بیمار نسبت به Monocytes. جدول تکمیلی در فایل MolecularFunction_AML_Monocytes_Up.xlsx در دایرکتوری result/ontology/amu ضمیمه شده‌است.

با بررسی کامپونت‌های سلول، می‌توان دریافت که این ژن‌ها در cmg که وظیفه باز کردن رشته‌های DNA را از یکدیگر دارند موثر است، در spindle که ساختمان سلول را حفظ کرده و ارگانیزم‌ها را مدیریت کرده و همچنین در تقسیم سلول وظیفه جداسازی اجزا را به عهده دارد، در Golgi که وظیفه مدیریت و بسته بندی خروجی‌ها و ورودی‌های سلول را به عهده دارد و همچنین در kinase که یکی از آنزیم‌های مهم در تسهیل فرآیندهای درون سلولی می‌باشد و موارد متعدد دیگر موثر است. این تاثیر به نحوی است که با افزایش بیان این ژن‌ها، بروز خطا در عملکرد این موارد را شاهد هستیم و باعث ایجاد مشکلات سرطانی در سلول می‌شود. مرتبط‌ترین کامپونت‌های سلولی به ژن‌های افزایش یافته در جدول۸ لیست شده‌اند.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	CMG complex (GO:0071162)	5.09E-12	1.60E-09	185110	4813717.69
2	spindle (GO:0005819)	6.27E-10	9.87E-08	3.44	72.95
3	nuclear chromosome (GO:0000228)	2.26E-08	0.000002345	4.82	84.95
4	intracellular non-membrane-bounded organelle (GO:0043232)	2.98E-08	0.000002345	1.74	30.08

5	Golgi cis cisterna (GO:0000137)	5.79E-08	0.000003647	11.56	192.65
6	cyclin-dependent protein kinase holoenzyme complex (GO:0000307)	7.39E-08	0.000003878	9.58	157.34
7	Golgi cisterna membrane (GO:0032580)	1.00E-07	0.00000451	10.73	172.99
8	serine/threonine protein kinase complex (GO:1902554)	0.000001341	0.00005282	6.78	91.74
9	mitotic spindle (GO:0072686)	0.000002155	0.00007543	3.31	43.19
10	Golgi cisterna (GO:0031985)	0.000004668	0.0001337	4.95	60.72

جدول ۸ – کامپوننت‌های مرتبط با ژن‌های افزایش بیان یافته در بیماران نسبت به Monocytes. لیست تکمیلی این موارد در فایل CellComponent_AML_Monocytes_Up.xlsx در دایرکتوری result/ontology/amu ضمیمه شده‌است.

٤-۲-۲. بررسی Gene ontology مرتبط با دسته AML در مقابل Monocytes که کاهش بیان داشته اند

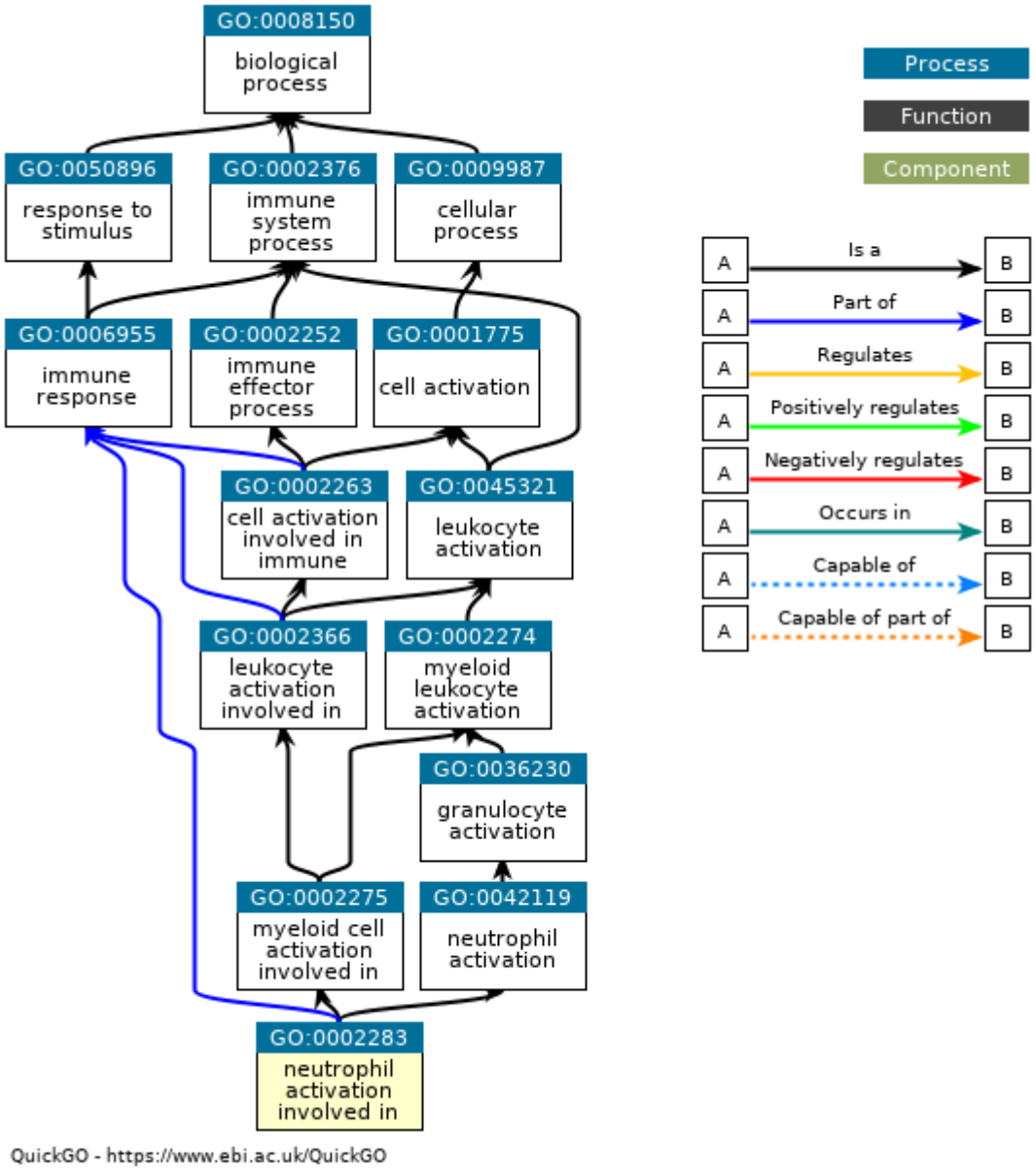
با درج ژن‌های کاهش بیان داشته در نمونه‌های سرطانی نسبت به Monocytes در وبسایت Enrichr و بررسی Ontologyهای ارائه شده؛ این ژن‌ها در فرآیندهای متعددی اثر بالایی دارند و همانگونه که انتظار می‌رود تعداد زیادی از موارد که دارای adj.P.Val بسیار کوچکی هستند بسیار زیاد است و در جدول۹ بعضی از این موارد آورده شده‌است. به منظور بررسی، مورد اول جدول۹، این مورد مربوط به تغییر در مورفولوژی و رفتار یک نوتروفیل ناشی از قرار گرفتن در معرض یک سیتوکین، کموکاین، لیگاند سلولی یا عامل محلول، که منجر به شروع یا تداوم یک پاسخ ایمنی می‌شود[1](شکل۱۵). دیگر موارد نیز به صورت مستقیم یا غیر مستقیم در پروسه‌های ایمنی بدن موثر هستند. (تصاویر برخی از پروسه‌های دیگر در دایرکتوری result/ontology/amd ضمیمه شده‌است.)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	neutrophil activation involved in immune response (GO:0002283)	2.05E-49	8.34E-46	5.81	650.83
2	neutrophil degranulation (GO:0043312)	4.00E-49	8.34E-46	5.81	647.55
3	neutrophil mediated immunity (GO:0002446)	2.66E-48	3.70E-45	5.69	623.53
4	cytokine-mediated signaling pathway (GO:0019221)	1.19E-30	1.24E-27	3.77	259.59
5	interferon-gamma-mediated signaling pathway (GO:0060333)	1.54E-27	1.28E-24	18.14	1119.67
6	cellular response to interferon-gamma (GO:0071346)	2.89E-25	2.00E-22	9.22	521.17
7	regulation of immune response (GO:0050776)	2.15E-21	1.28E-18	6.02	286.52

[1] https://www.ebi.ac.uk/QuickGO/term/GO:0002283

8	positive regulation of cytokine production (GO:0001819)	4.10E-20	2.14E-17	4.02	179.28
9	regulation of interleukin-6 production (GO:0032675)	5.52E-18	2.56E-15	7.39	293.7
10	regulation of interleukin-8 production (GO:0032677)	1.48E-17	5.61E-15	9.22	357.35

جدول ۹- بررسی پروسه‌های زیستی مرتبط با ژن‌های کاهش بیان داشته در نمونه‌های بیمار نسبت به Monocytes. جدول تکمیلی در فایل BiologicalProcess_aml_monocytes_down.xlsx در دایرکتوری result/ontology/amd ضمیمه شده‌است.



شکل ۱۵- پروسه‌های زیستی مربوط به neutrophil activation involved in immune response مرتبط با ژن‌های کاهش بیان داشته در گونه‌های بیمار نسبت به Monocytes.

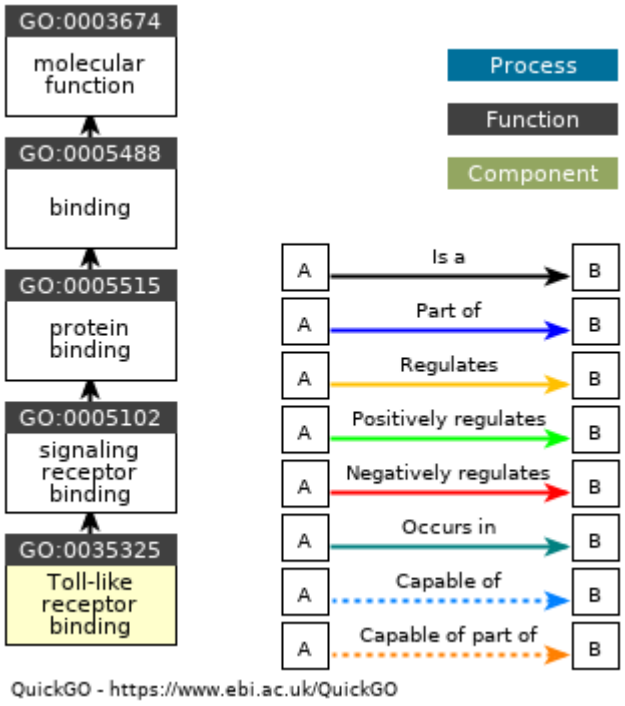
بررسی عملکردهای مولکولی مرتبط با ژن‌های کاهش بیان داشته در سایت Enrichr نتایج جدول۱۰ را ارائه می‌کند. مورد اول یعنی Toll-like receptor binding مرتبط با فعالیت سیستم ایمنی ذاتی بوده و در ارتباط با اتصال پروتئین‌های خاص به منظور شروع عملکرد سیستم ایمنی ذاتی می‌باشد[1](شکل۱۶). مورد دوم جدول فعالیت‌های بین سلولی و سیگنالینگ تغییرات در عملکرد سلول را به عهده دارد[2]. (تصاویر برخی دیگر از عملکردهای مولکولی در دایرکتوری result/ontology/amd ضمیمه شده‌است.)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Toll-like receptor binding (GO:0035325)	2.09E-09	0.000001545	59.48	1188.61
2	MHC class II receptor activity (GO:0032395)	2.44E-08	0.000007187	52.83	926.11
3	cytokine receptor activity (GO:0004896)	2.92E-08	0.000007187	4.71	81.67
4	complement receptor activity (GO:0004875)	8.38E-08	0.00001546	35.22	573.88
5	GTPase activator activity (GO:0005096)	1.88E-07	0.0000278	2.4	37.18
6	amyloid-beta binding (GO:0001540)	5.05E-07	0.00006206	4.43	64.18
7	GTPase regulator activity (GO:0030695)	0.000003003	0.0003166	2.52	32.03
8	icosanoid binding (GO:0050542)	0.00003277	0.003023	32.95	340.23
9	purine ribonucleoside triphosphate binding (GO:0035639)	0.00009239	0.007576	1.82	16.86
10	adenyl ribonucleotide binding (GO:0032559)	0.0002582	0.01906	1.94	16.07

جدول ۱۰- بررسی عملکردهای مولکولی مرتبط با ژن‌های کاهش بیان داشته در مقایسه با Monocytes. جدول تکمیلی در فایل MolecularFunction_aml_monocytes_down.xlsx در دایرکتوری result/ontology/amd ضمیمه شده‌است.

با بررسی کامپونت‌های سلول مرتبط با ژن‌های کاهش بیان داشته، جدول۱۱ حاصل شده‌است. بسیاری از موارد با سطوح غشایی اندامک ها، سلول، مواد ترشحی و ... در ارتباط هستند و بعضی دیگر با فرآیند از بین بردن مواد اضافی سلول از جمله Lysosomeها در ارتباط هستند.

[1] https://www.ebi.ac.uk/QuickGO/term/GO:0035325
[2] https://www.ebi.ac.uk/QuickGO/term/GO:0032395



شکل ۱۶- بررسی عملکرد مولکولی مربوط به Toll-like receptor binding تحت تاثیر کاهش بیان ژن‌های یافته شده.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	secretory granule membrane (GO:0030667)	9.26E-33	2.78E-30	6.22	458.82
2	tertiary granule (GO:0070820)	2.67E-24	4.01E-22	7.05	382.66
3	lysosome (GO:0005764)	6.42E-24	6.42E-22	3.72	198.78
4	ficolin-1-rich granule (GO:0101002)	1.55E-21	1.16E-19	5.94	284.71
5	cytoplasmic vesicle membrane (GO:0030659)	9.69E-21	5.82E-19	3.84	176.82
6	lysosomal membrane (GO:0005765)	1.04E-18	5.19E-17	3.88	160.64
7	ficolin-1-rich granule membrane (GO:0101003)	9.38E-17	4.02E-15	11.35	418.84
8	tertiary granule membrane (GO:0070821)	3.28E-16	1.23E-14	9.34	333.02
9	lytic vacuole membrane (GO:0098852)	6.09E-16	2.03E-14	3.93	137.74
10	endocytic vesicle (GO:0030139)	7.29E-16	2.19E-14	4.73	164.72

جدول ۱۱- بررسی کامپونت‌های سلول مرتبط با ژن‌های کاهش بیان داشته در بیماران نسبت به Monocytes. جدول تکمیلی در فایل CellComponent_aml_monocytes_down.xlsx در دایرکتوری result/ontology/amd ذخیره شده‌است.

٤-٢-٣. بررسی Gene ontology مرتبط با دسته AML در مقابل CD34+HSPC که افزایش بیان داشته اند

با بررسی ژن‌های افزایش بیان یافته در نمونه‌های AML نسبت به CD34، نشان می‌دهد یکی از موارد مرتبط همانند Monocytes، در مورد پروسه‌های زیستی مرتبط با ژن، neutrophil activation involved in immune response می‌باشد که در شکل۱۵ قابل مشاهده است. همچنین دیگر مواردی که در جدول۹ مورد بحث قرار گرفت اشتراک بسیار زیادی با جدول۱۲ که نتایج این افزایش بیان در مقایسه با CD34 می‌باشد، دارند.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	neutrophil activation involved in immune response (GO:0002283)	8.46E-33	3.08E-29	5.55	410.19
2	neutrophil degranulation (GO:0043312)	2.44E-32	4.43E-29	5.53	402.41
3	neutrophil mediated immunity (GO:0002446)	7.64E-32	9.25E-29	5.43	389.01
4	regulation of immune response (GO:0050776)	3.04E-11	2.21E-08	4.8	116.35
5	cytokine-mediated signaling pathway (GO:0019221)	2.96E-11	2.21E-08	2.74	66.34
6	cellular response to cytokine stimulus (GO:0071345)	1.22E-10	7.38E-08	2.93	66.95
7	positive regulation of cellular process (GO:0048522)	3.03E-08	0.00001574	2.37	41.07
8	mitotic sister chromatid segregation (GO:0000070)	6.02E-08	0.00002518	5.19	86.35
9	regulation of tumor necrosis factor production (GO:0032680)	6.93E-08	0.00002518	4.63	76.36
10	mitotic spindle organization (GO:0007052)	6.30E-08	0.00002518	4.11	68.11

جدول ۱۲- بررسی پروسه‌های زیستی مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های AML نسبت به نمونه‌های CD34+HSPC. جدول تکمیلی در فایل BiologicalProcess_aml_cd34_up.xlsx در دایرکتوری result/ontology/acu ضمیمه شده‌است.

با بررسی عملکردهای مولکولی مرتبط با این ژن ها، جدول۱۳ حاصل شده‌است. با بررسی خروجی‌های به دست آمده ارتباط جدی این ژن‌ها با عملکردهای سوخت و ساز در سلول را می‌توان مشاهده کرد از جمله این موارد تاثیر این ژن‌ها در عملکرد آنزیم کیناز می‌باشد که در تعداد زیادی از نتایج به صورت مستقیم یا غیرمستقیم تحت تاثیر این ژن هاست. علاوه بر کیناز موارد دیگری نیز وجود دارد که مرتبط با عملکردهای سایتواسکلتون‌های سلول می‌باشد.

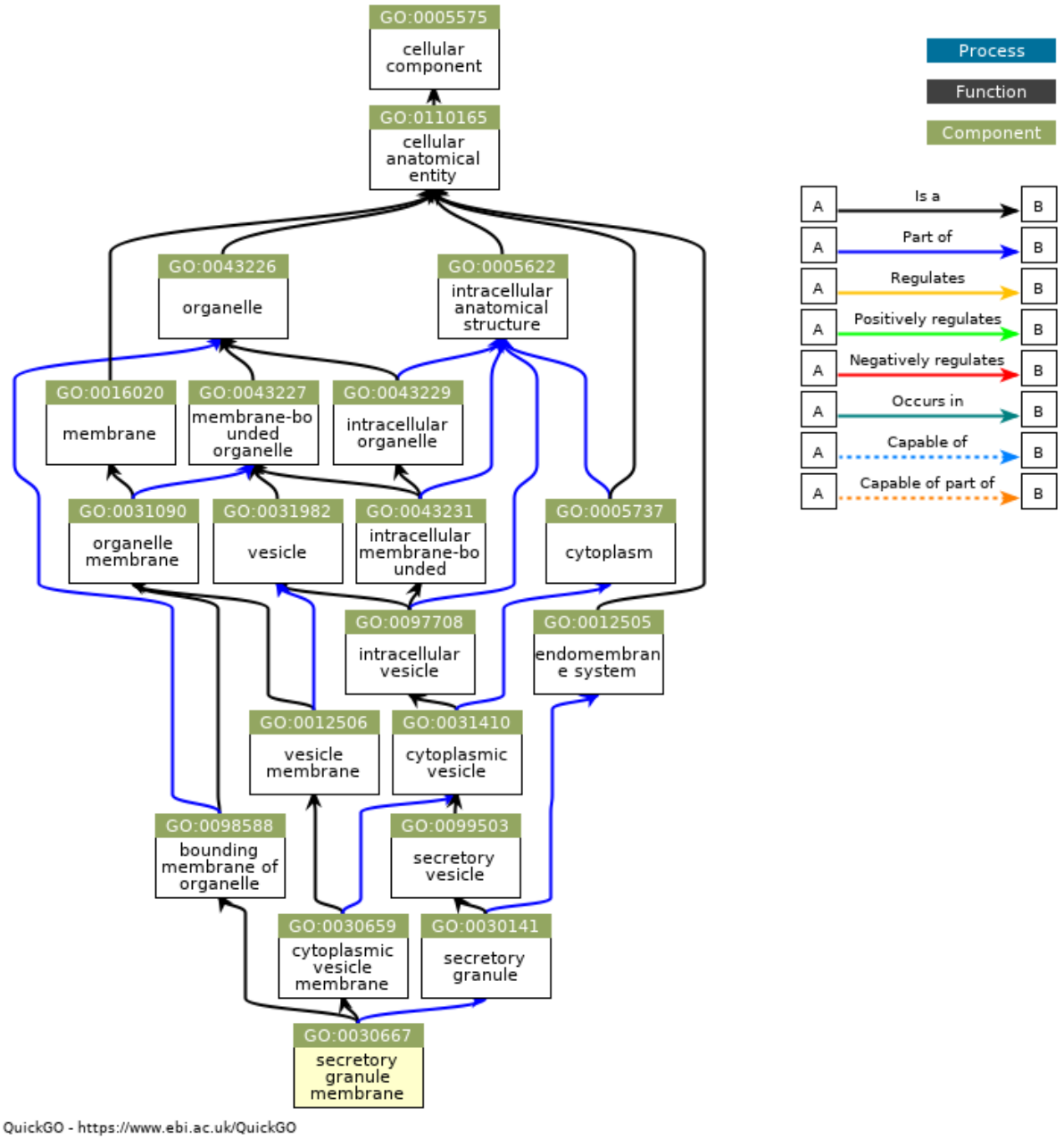
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	kinase binding (GO:0019900)	1.44E-08	0.000009299	2.68	48.45
2	cyclin-dependent protein serine/threonine kinase regulator activity (GO:0016538)	2.16E-07	0.00006961	8.46	129.84
3	protein kinase binding (GO:0019901)	6.15E-07	0.000132	2.36	33.69
4	superoxide-generating NAD(P)H oxidase activity (GO:0016175)	0.00005375	0.008654	18.67	183.57
5	microtubule motor activity (GO:0003777)	0.0001198	0.01543	4.89	44.15
6	amyloid-beta binding (GO:0001540)	0.0001493	0.01603	3.97	35.01
7	oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor (GO:0050664)	0.0003026	0.02436	11.2	90.76
8	GTPase activator activity (GO:0005096)	0.0003026	0.02436	2.14	17.37
9	cytokine receptor activity (GO:0004896)	0.0003721	0.02662	3.55	28.06
10	motor activity (GO:0003774)	0.0004821	0.03105	4.01	30.66

جدول ۱۳- بررسی عملکردهای مولکولی مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های AML نسبت به CD34+HSPC. جدول تکمیلی در فایل MolecularFunction_aml_cd34_up.xlsx در دایرکتوری result/ontology/acu ضمیمه شده‌است.

در بررسی کامپونت‌های سلولی تاثیر پذیر از ژن‌های به دست آمده، جدول۱٤ حاصل شده‌است. تعداد زیادی از خروجی‌ها مرتبط با گرانول (ذرات ترشحات داخل سلولی که در یک محفظه از فسفولیپیدها حمل می‌شوند.)های ترشحی و همچنین لیزوزوم‌ها هستند که می‌توان نتیجه گرفت ژن‌های موثر در این عملکرد، با ایجاد مشکل در بیانشان باعث بروز خطاهایی در این عملکردهای تخریبی و همچنین کنترلی برای ترشح و حفاظت از مواد داخل سلول می‌شوند. نمونه‌ای از این گرانول‌ها در شکل۱۷ قابل بررسی می‌باشد. بعضی دیگر از این موارد در دایرکتوری result/ontology/acu ضمیمه شده‌اند.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	secretory granule membrane (GO:0030667)	7.80E-15	2.07E-12	4.57	148.31
2	azurophil granule (GO:0042582)	1.60E-14	2.12E-12	6.23	197.86
3	specific granule (GO:0042581)	2.51E-13	2.22E-11	5.75	166.76
4	secretory granule lumen (GO:0034774)	3.68E-13	2.45E-11	3.95	113.22
5	specific granule membrane (GO:0035579)	3.56E-12	1.89E-10	7.72	203.43
6	azurophil granule lumen (GO:0035578)	2.15E-11	9.54E-10	7.37	181.08
7	tertiary granule (GO:0070820)	7.84E-11	2.98E-09	4.92	114.46
8	vacuolar lumen (GO:0005775)	2.48E-10	8.25E-09	4.82	106.51
9	ficolin-1-rich granule (GO:0101002)	1.32E-09	3.89E-08	4.28	87.52
10	cytoplasmic vesicle membrane (GO:0030659)	7.23E-09	1.92E-07	2.94	55.12

جدول ۱٤- بررسی کامپونت‌های سلولی تاثیر گرفته از ژن‌های افزایش بیان داشته در نمونه‌های بیمار نسبت به CD34+HSPC. جدول تکمیلی در فایل CellComponent_aml_cd34_up.xlsx در دایرکتوری result/ontology/acu ضمیمه شده‌است.



شکل ۱۷- بررسی secretory granule membrane که تحت تاثیر جدی ژن‌های افزایش بیان داشته در نمونه‌های بیمار AML نسبت به CD34+HSPC می‌باشد.

٤-٢-٤. بررسی Gene ontology مرتبط با دسته AML در مقابل CD34+HSPC که کاهش بیان داشته اند

با بررسی پروسه‌های زیستی، عملکردهای مولکولی و کامپونت‌های سلولی تفاوت‌های معنی‌داری مشاهده نمی‌شود؛ خروجی‌های هرکدام از جداول در دایرکتوری result/ontology/acd ضمیمه شده‌است. تنها موردی که با توجه به خروجی‌های قسمت‌های قبل می‌توان کمی به آن اعتماد کرد ارتباط این ژن‌ها با GTPase می‌باشد.

٥. بررسی مقالات مرتبط

با توجه به pathwayها و Gene ontology به دست آمده در قسمت قبل، در این بخش به بررسی موارد به دست آمده و مقایسه آن‌ها با مقالات پرداخته شده‌است.

یکی از موارد مهمی که در نتایج به دست آمده در قسمت‌های قبل به آن اشاره شد، فرآیند مرتبط به سیگنالینگ p53 می‌باشد. در مقاله [١] راجع به فرآیند مرتبط با سیگنالینگ p53 بحث شده‌است و نتیجه گیری نهایی این بوده است که به صورت مشهودی ژن‌های به دست آمده در مقاله که مربوط به مقایسه بیماران مبتلا به AML در مقایسه با نمونه‌های سالم می‌باشد، ارتباط جدی‌ای با فرآیند سیگنالینگ p53 دارند و نتیجه گیری انجام شده این است که «مسیر p53 در زیرگروه‌های مختلف AML غیرفعال می‌شود. تجزیه و تحلیل متمرکز ژن و پروتئین مسیر p53 در بیماران AML و APL نشان می‌دهد که غیرفعال‌سازی عملکردی پروتئین p53 را می‌توان به استیلاسیون مختل آن نسبت داد.» این اختلال گزارش شده در استیلاسیون، در نتایج قسمت‌های قبلی نیز وجود دارد که تاییدی بر نتایج به دست آمده در تحلیل داده‌ها می‌باشد. همچنین در مقاله [٢] و [٣] راجع به ارتباط‌های مستقیم و غیرمستقیم p53 با ایجاد و گسترش AML بحث شده‌است.

یکی از فاکتورهای رونویسی‌ای که مرتبط با داده‌های به دست آمده است، E2F4 می‌باشد که در مقاله [٤] توضیح می‌دهد که افزایش بیان ژن‌های مرتبط با این فاکتور رونویسی از روند تکثیر و گسترش AML جلوگیری می‌کند.

یکی از موارد مرتبط دیگر که در قسمت‌های قبل به دست آمده است GTPase می‌باشد که در مقاله [٥] GRAF که تنظیم کننده GTPase می‌باشد دچار عملکرد غیر طبیعی می‌شود و در نتایج تاکید می‌کند تاثیر جدی‌ای بر ایجاد و گسترش AML می‌باشد.

همچنین یکی از موارد ژن‌های مرتبط با فاکتور رونویسی EZH2 می‌باشد که در مقاله [٦] و [٧] در مورد تاثیر جدی این فاکتور رونویسی در کنترل AML بحث شده‌است.

یکی از موارد مهم امکان اتصال و گیرنده‌های کیناز هستند که در جهش‌های بیمارگونه AML این نقاط دچار مشکل می‌شوند. در مقاله [٨] در ارتباط با یک از این موارد بحث شده‌است.

لیزوزوم‌ها به عنوان یکی از اندامک‌های مهم درون سلول نقشی اساسی در حیات سلول ایفا می‌کنند. با توجه به نتایج قسمت‌های قبل یکی از موارد مرتبط به ژن‌های آن افزایش بیان داشتند، ژن‌های مرتبط با لیزوزوم‌ها بودند. در مقاله [٩] نیز لیزوزوم‌های بزرگتر درون سلول را به عنوان یکی از وجوه مهم تمایز میان AML و دیگر سلول‌ها ارائه کرده است.

یکی دیگر از مواردی که در قسمت‌های قبل ارائه شد ارتباط ژن‌ها با شبکه سیتوکین بود؛ در مقاله [١٠] نقش شبکه سیتوکین را در ایجاد و گسترش AML مورد بررسی قرار می‌دهد.

٦. بررسی برخی درمان‌ها و داروهای مرتبط

با توجه به pathwayها و Gene ontology به دست آمده در قسمت‌های قبل، درمان‌ها و داروهای مرتبط در این بخش مورد بررسی قرار گرفته‌است.

در قسمت قبل اشاره شد که یکی از موارد که در مقالات مورد بررسی قرار گرفته‌است گیرنده‌ها و اتصالات کیناز می‌باشد، در مقاله [٨] از AC220 به عنوان مهار کننده جهش‌های FLT3 در شرایط آزمایشگاهی استفاده شده و نتایج مطلوبی ارائه کرده است.

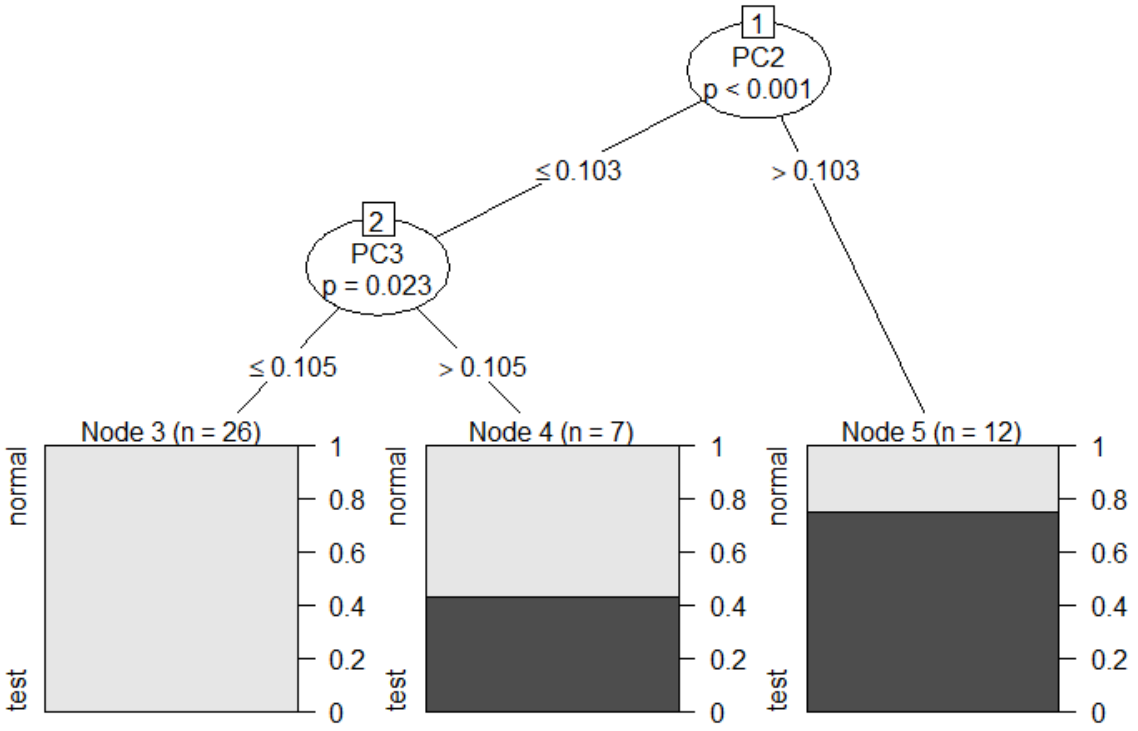
در مقاله [٩] با توجه به ویژگی خاص سلول‌های AML و ارتباط آن‌ها با مالاریا (در قسمت‌های قبل میان ژن‌های این بیماری و AML ارتباط جدی‌ای مشاهده شد.)، میان کتاب‌خانه‌های داروهای مرتبط با مالاریا بررسی انجام شده‌است و یکی از مهم‌ترین نتایج داروی ضد مالاریای مفلوکین می‌باشد که دیواره لیزوزوم را هدف قرار داده و آن را تضعیف و ازبین می‌برد. در نتیجه منجر به کاهش عمر سلول و از بین رفتن سلول خواهد شد. این راهکار در آزمایشگاه برای هدف قرار دادن سلول‌های بیمار AML نتیجه بخش بوده است.

مقاله [١٠] که در بخش قبل ارائه شد، راهکارهای حمله به این شبکه سیتوکین ای به منظور درمان بیماری AML را مورد بحث قرار می‌دهد.

یکی از جدی‌ترین موارد IL3 می‌باشد که تغییر در بیان آن یکی از اشتراکات مهم بین بیماران AML است، در مقاله [١١] از آنتی بادی‌ها برعلیه IL3 استفاده می‌کند تا بتواند سلول‌های بیمار را هدف قرار دهد.

۷. اجرای روش یادگیری ماشین برای داده ها

یکی از روش‌های مناسب، استفاده از درخت تصمیم می‌باشد. به منظور استفاده از درخت تصمیم بر روی داده ها، ابتدا داده‌ها تغییر ابعاد داده شدند و سپس بر اساس ۳ مولفه اول روش PCA درخت تصمیم رسم شده است. به منظور داده یادگیری ۷۵ درصد داده‌ها و به منظور تست ۲۵ درصد دیگر داده‌ها مورد استفاده قرار گرفته است. پس از بررسی خروجی میزان قابل اتکا بودن درخت ۰.۸۱ برآوورد شده است. (البته این نکته حائز اهمیت است که با توجه به کم بودن تعداد داده‌ها این خروجی قابل اعتماد به منظور استفاده در تصمیم‌گیری نیست.) خروجی درخت تصمیم در شکل۱۸ ضمیمه شده است. همچنین کدهای مورد استفاده برای این الگوریتم در فایل dtree.R در دایرکتوری src ذخیره شده است.



شکل ۱۸- درخت تصمیم خروجی.

یکی از روش‌های دیگر روش SVM است، بعد از اجرای روش SVM (کد در فایل svm.R در دایرکتوری src ضمیمه شده است. همچنین خروجی نمودارهای آن در دایرکتوری result/svm ضمیمه شده است.) خروجی‌های آن بسیار با هم متفاوت است و اصلا قابل اعتماد نیستند.

٨ مراجع

[١] J. Abramowitz, T. Neuman, R. Perlman, and D. Ben-Yehuda, "Gene and protein analysis reveals that p53 pathway is functionally inactivated in cytogenetically normal Acute Myeloid Leukemia and Acute Promyelocytic Leukemia," *BMC medical genomics,* vol. 10 ,no. 1, pp. 1-16, 2017.

[٢] K. Kojima *et al.*, "The dual PI3 kinase/mTOR inhibitor PI-103 prevents p53 induction by Mdm2 inhibition but enhances p53-mediated mitochondrial apoptosis in p53 wild-type AML," *Leukemia,* vol. 22, no. 9, pp. 1728-1736, 2008.

[٣] Y. Lyu *et al.*, "Dysfunction of the WT1-MEG3 signaling promotes AML leukemogenesis via p53-dependent and-independent pathways," *Leukemia,* vol. 31, no. 12, pp. 2543-2551, 2017.

[٤] Y. Feng, L. Li, Y. Du, X. Peng, 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,* vol. 24, no. 3, pp. 2157-2168, 2020.

[٥] S. Bojesen *et al.*, "Characterisation of the GRAF gene promoter and its methylation in patients with acute myeloid leukaemia and myelodysplastic syndrome," *British journal of cancer,* vol. 94, no. 2, pp. 323-332, 2006.

[٦] B. Salvatori *et al.*, "Critical role of c-Myc in acute myeloid leukemia involving direct regulation of miR-26a and histone methyltransferase EZH2," *Genes & cancer,* vol. 2, no. 5, pp. 585-592, 2011.

[٧] J. Wang *et al.*, "Analysis of TET2 and EZH2 gene functions in chromosome instability in acute myeloid leukemia," *Scientific reports,* vol. 10, no. 1, pp. 1-11.٢٠٢٠ ,

[٨] T. Grafone, M. Palmisano, C. Nicci, and S. Storti, "An overview on the role of FLT3-tyrosine kinase receptor in acute myeloid leukemia: biology and treatment," *Oncology reviews,* vol. 6, no. 1, 2012.

[٩] M. A. Sukhai *et al.*, "Lysosomal disruption preferentially targets acute myeloid leukemia cells and progenitors," *The Journal of clinical investigation,* vol. 123, no. 1, 2012.

[١٠] H. Reikvam, K. J. Hatfield, H. Fredly, I. Nepstad, K. A. Mosevoll, and Ø. Bruserud, "The angioregulatory cytokine network in human acute myeloid leukemia-from leukemogenesis via remission induction to stem cell transplantation," *European cytokine network,* vol. 23, no. 4, pp. 140-153, 2012.

[١١] D. Kirchhoff *et al.*, "IL3RA-Targeting Antibody–Drug Conjugate BAY-943 with a Kinesin Spindle Protein Inhibitor Payload Shows Efficacy in Preclinical Models of Hematologic Malignancies," *Cancers,* vol. 12, no. 11, p. 3464, 2020.