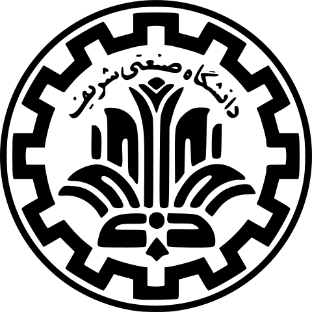
بسم الله الرحمن الرحیم



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:

400200919

بهمن 1400

**فهرست**

[**1. مقدمه** 3](#_Toc94900145)

[**1-1 دادگان و ابزارها و روش‌های مورد استفاده** 3](#_Toc94900146)

[**2-1 تنظیمات و موارد اولیه** 4](#_Toc94900147)

[**2. کنترل کیفیت داده‌ها** 5](#_Toc94900148)

[**1-2. نرمال‌سازی و نمودار جعبه‌ای** 6](#_Toc94900149)

[**3-2. بررسی همبستگی بین نمونه‌ها** 15](#_Toc94900150)

[**1-3-2. بررسی همبستگی بین تمامی نمونه‌ها** 15](#_Toc94900151)

[**2-3-2. بررسی همبستگی بین نمونه‌های سالم** 17](#_Toc94900152)

[**3. بررسی تمایز در بیان ژن‌ها** 18](#_Toc94900153)

[**4. آنالیز Gene ontology و pathway‌ها** 23](#_Toc94900154)

[**1-4. بررسی pathway‌ها** 23](#_Toc94900155)

[**1-1-4. بررسی pathway‌های مرتبط با دسته AML در مقابل Monocytes که افزایش بیان داشته‌اند** 24](#_Toc94900156)

[**2-1-4. بررسی pathway‌های مرتبط با دسته AML در مقابل Monocytes که کاهش بیان داشته‌اند** 26](#_Toc94900157)

[**1-4. بررسی Gene ontology** 27](#_Toc94900158)

[**2-1-4. بررسی Gene ontology مرتبط با دسته AML در مقابل Monocytes** 27](#_Toc94900159)

# **1. مقدمه**

سرطان خون انواع مختلفی دارد، یکی از انواع آن لوسمی حاد میلوئیدی[[1]](#footnote-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]](#footnote-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 معنی داری داشته‌اند می‌پردازیم.

## **1-1 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 گرفته است. 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 نگرفته است. به منظور بررسی ژن‌های افزایش و کاهش یافته از دیتابیس‌های Enrichr استفاده شده‌است.

## **2-1 تنظیمات و موارد اولیه**

به منظور تحليل داده‌ها ابتدا مي بايست، داده‌ها را دانلود و در دايركتوري مناسب قرار داده و نمونه‌هاي نرمال و تست را مشخص نمود. براي اين منظور از طريق قطعه كد زير، محل فعلي دايركتوري را به عنوان مرجع مشخص مي كنيم و سپس براساس آن آدرس دهي‌هاي بعدي را انجام مي دهيم:

curD <- dirname(rstudioapi::getActiveDocumentContext()$path)

setwd(sub(paste0("/",sub("(.+)/","",curD)),"",curD))

سپس مي توان شماره سري داده‌ها و پلتفرم آن‌ها را به صورت مجزا تعيين نمود، البته الزامي به اين كار نيست و مي توان داخل كد دستوري قرار داد، پس از آن مي بايست داده‌ها را در دايركتوري مورد دلخواه خود دانلود نمود:

series <- "GSE48558"

platform <- "GPL6244"

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

براي اين منظور از تابعي در كتابخانه GEOquery به منظور دانلود استفاده مي شود.

پس از آن درصورتي كه داده چندين پلتفرم داشته باشد، صرفا پلتفرم خاص تعيين شده توسط سوال را مي خواهيم. براي اين منظور مي بايست يك فيلتر روي داده صورت پذيرد:

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

gset <- gset[[idx]]

البته درصورتي كه از داده length گرفته شود و صرفا يك پلتفرم موجود باشد، كافيست همان جايگزين شود و ديگر به عنوان ليستي از پلتفرم‌ها نباشد. در نهايت مي بايست گروه‌هاي نرمال و بيمار و غيره را مشخص كنيم و گروه‌هاي به جز بيمار و نرمال را از داده‌ها حذف كنيم:

gsms <- paste0("1111111111111XXXXXXXXXXXXXXXXXXXXXXXXXXX0XXX0XXXXX",

"XXXXXXXXXXXXXXXXXX0X0XXX0X0000X0XX00XX00X0X0X0X0X0",

"XXX0XXX0XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",

"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 طريق قطعه كد بالا مي توان داده‌ها را دسته بندي كرده و يك ستون جديد به داده‌ها افزود و گروه بندي مورد نظر را در آن گنجاند. (البته گروه‌هاي غيراز سالم و بيمار، از داده‌ها حذف مي شوند.)

# **2. 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.

## **1-2. 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[[3]](#footnote-3) 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.

داده‌های موجود در این پروژه در صورت رسم نمودار جعبه‌ای نرمال هستند. (شکل 1) 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 1 نیازی به نرمال‌سازی نیست، درصورت اجرای نرمال‌سازی بالا و رسم دوباره نمودار جعبه ای، نمودار حاصل، شکل 2 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 جعبه‌اي و ايجاد قابليت تفكيك بين نمونه‌هاي سالم و بيمار، با استفاده از قطعه كد زير مي توان اين تفاوت را ايجاد نمود:

dev.new(width=3+ncol(gset)/6, height=5)

ord <- order(gs)

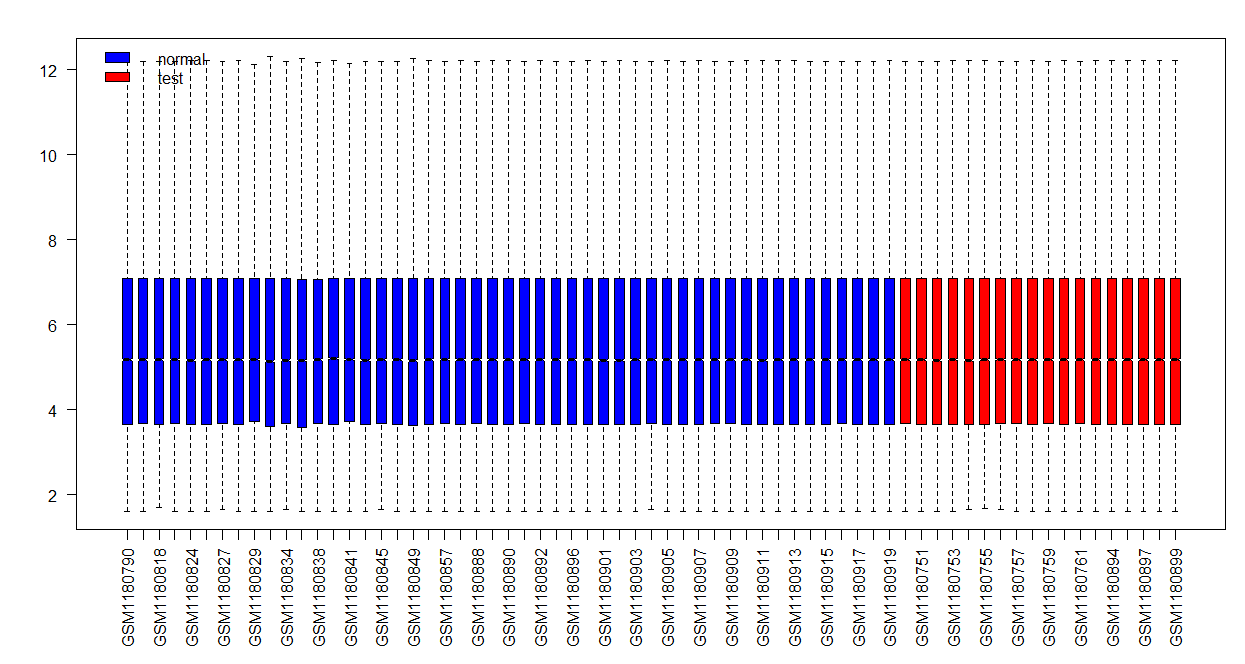
palette(c("blue", "red"))

par(mar=c(7,4,2,1))

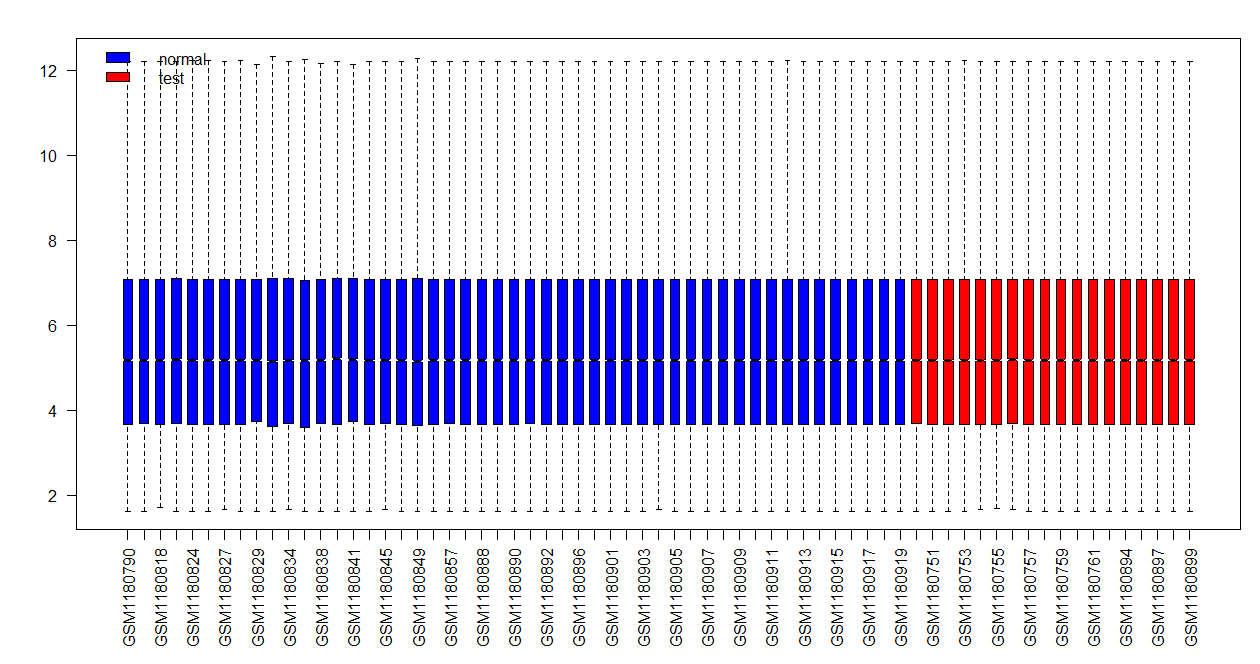
boxplot(ex[,ord], boxwex=0.6, notch=T, outline=FALSE, las=2, col=gs[ord])

legend("topleft", groups, fill=palette(), bty="n")

در كد بالا، ابتدا ابعاد نمودار خروجي مبتني بر نمونه‌ها مشخص مي‌شود، سپس‌ترتيب نمايش نمونه‌ها بر اساس نرمال به بيمار مرتب مي‌شود و پس از آن رنگ‌بندي نمودار براي جداسازي نرمال از بيار و در نهايت رسم نمودار و تعيين راهنما براي آن (مشخص كردن اين كه هر رنگ مرتبط به كدام نمونه است.)



**شکل 1**- 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.



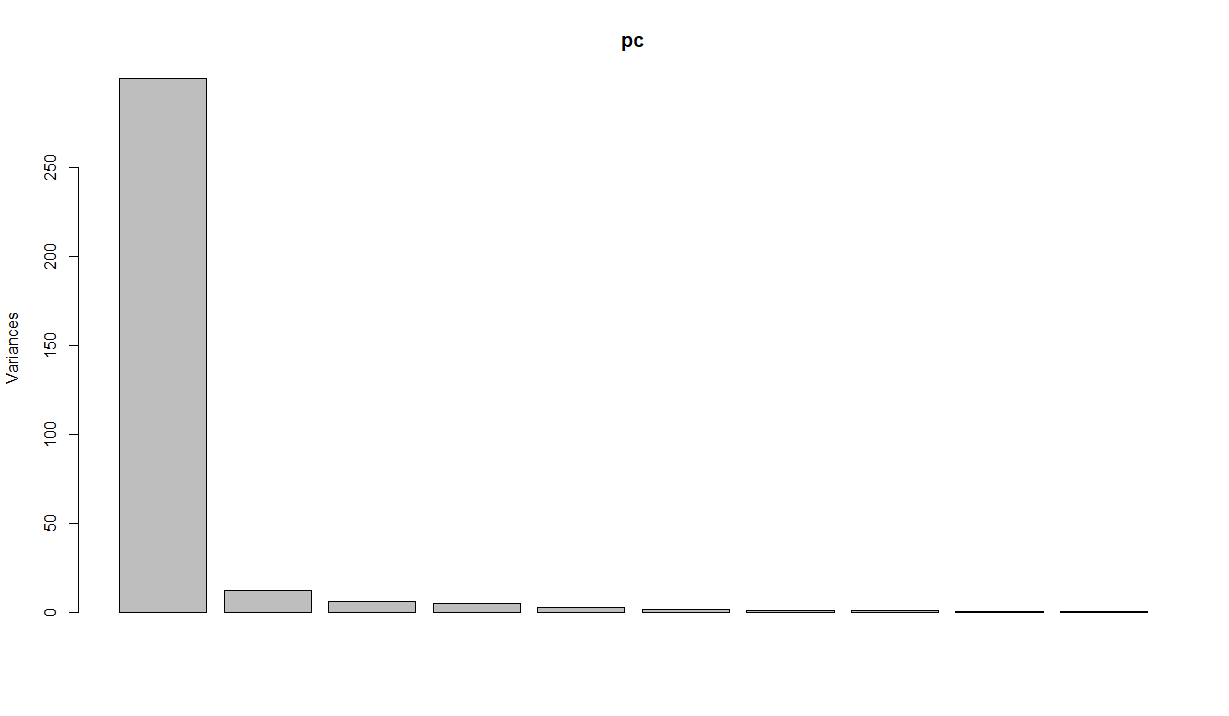
**شکل 2**- نمودار جعبه‌ای نمونه‌های سالم و بیمار. نمونه‌های بیمار با رنگ قرمز و نمونه‌های سالم با رنگ آبی مشخص شده‌اند. نمودار جعبه‌ای پس از اعمال نرمال‌سازی گسسته نشان دهنده این است که داده‌ها قبل از آن نیز نرمال بودنه‌اند چرا که در مقایسه با شکل1 تغییر محسوسی مشاهده نمی‌شود. از طرف دیگر شکل1 خود کاملا مشخص است که نرمال می‌باشد.

**2-2. کاهش ابعاد داده**

به منظور کاهش ابعاد روش‌های مختلفی وجود دارد، یکی از این روش‌ها PCA[[4]](#footnote-4) 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 (شکل3) می‌توان تفاوت‌های میزان تفکیک در هر کدام از این مولفه‌ها را مشاهده کرد. همانگونه که در شکل3 مشخص است میزان تمایز در مولفه اول از همه بیشتر و به‌ترتیب کاهش می‌یابد.



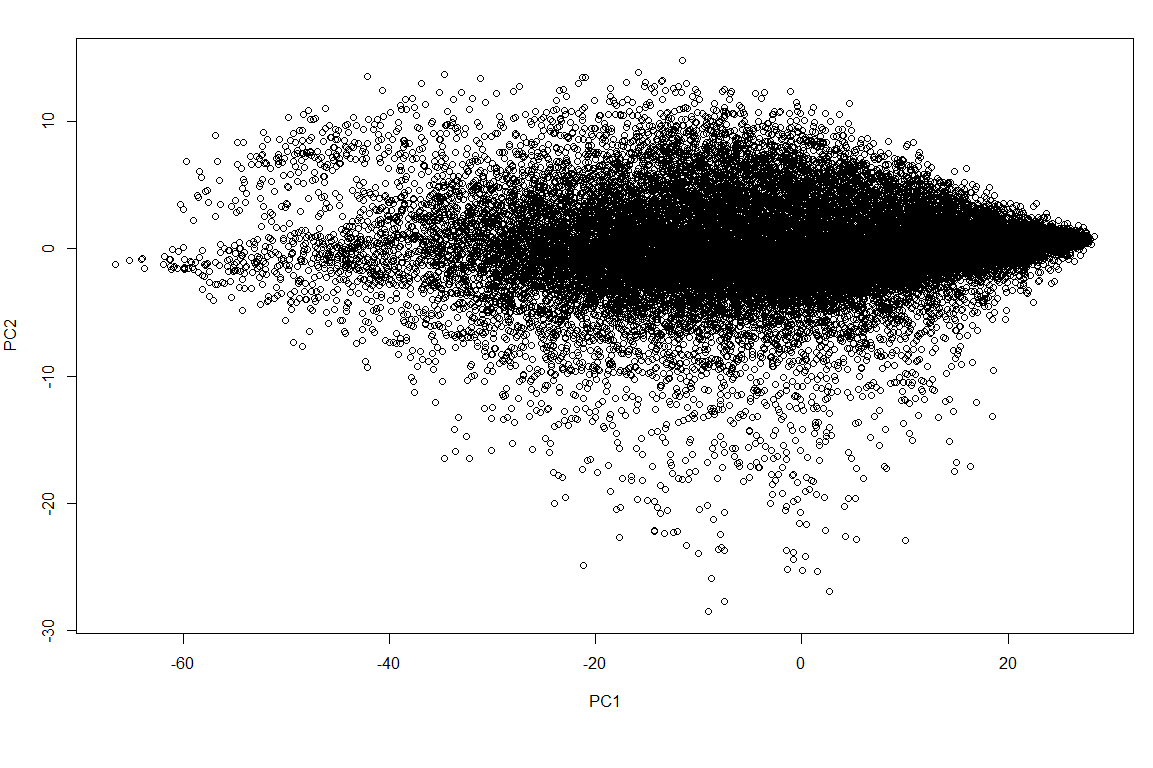
**شکل 3**- بررسی هرکدام از مولفه‌های خروجی تابع prcomp نشان دهنده میزان واریانس (تمایز) ایجاد شده میان ژن‌ها در هرکدام از این مولفه‌هاست. به‌ترتیب میزان تمایز از اولین به آخرین مولفه مرتب شده‌است. بنابراین بهترین گزینه برای رسم نمودار کاهش ابعاد مولفه اول و دوم می‌باشد. البته درصورت نیاز به رسم نمودار 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.(شکل4) مولفه اول در شکل4 نشان دهده میزان بیان ژن‌های مختلف در نمونه می‌باشد. همانگونه که درصورت رسم PCA برای نمونه‌ها مطابق شکل5، 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:

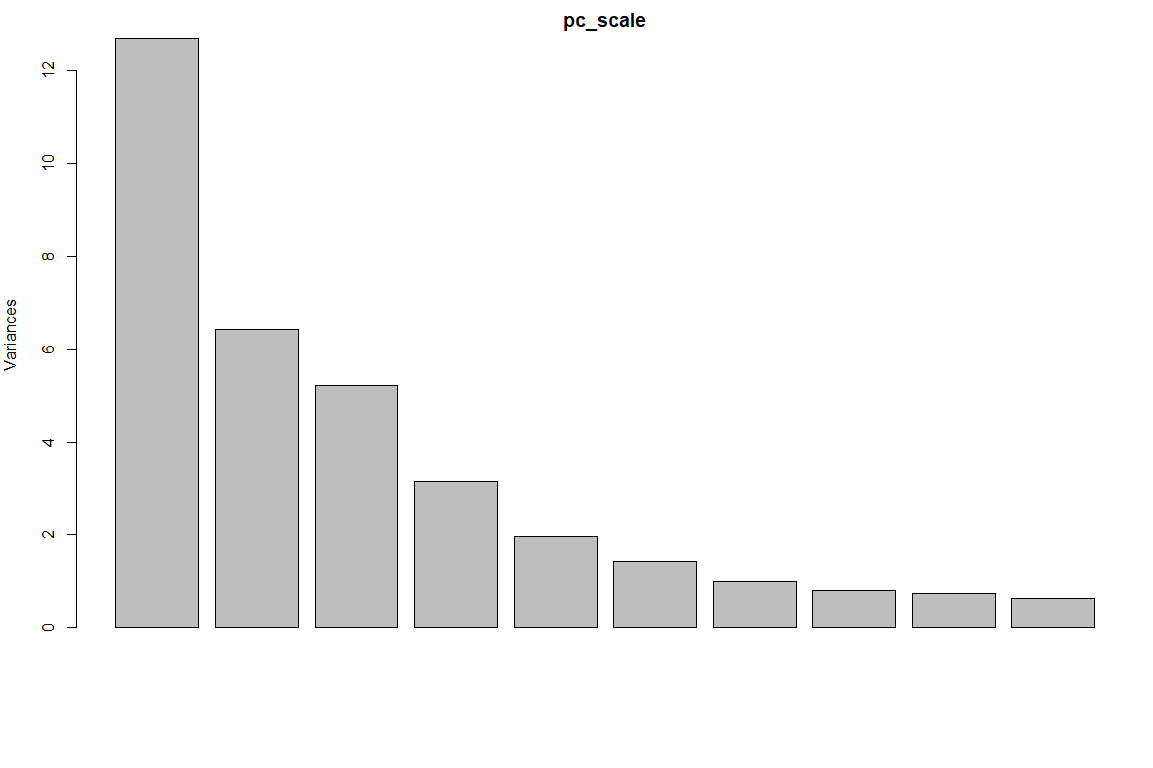
ex.scale <- t(scale(t(ex), scale = F))

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 متغیر scale در تابع، لغو نمود.

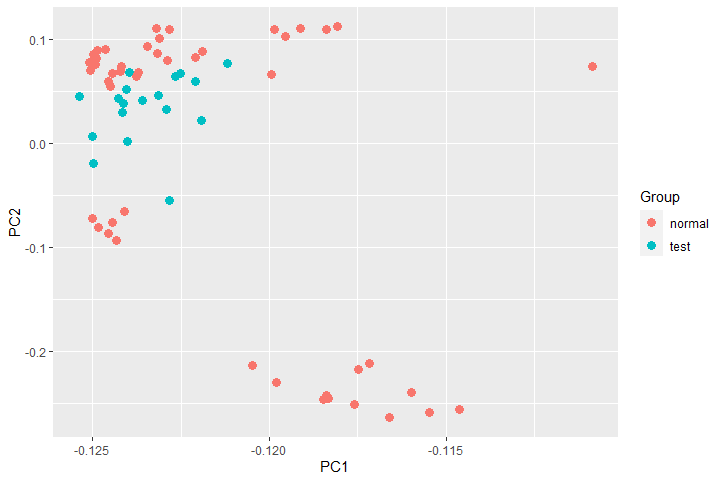
درصورتی که براساس مقادیر جدید، مولفه‌های مختلف روش PCA مورد بررسی قرار داده شود، شکل6 حاصل می‌شود که تمایز‌ها در مولفه‌های مختلف معقول‌تر شده‌اند و از طرفی در صورت رسم نمونه‌ها مطابق شکل7 ، می‌توان نتایج را مشاهده نمود که دیگر مولفه اول به سمت بیان کامل یا عدم بیان کامل گرایش ندارد و داده‌ها به صورت متفاوتی قرار گرفته‌اند. به منظور بررسی نمونه‌ها نیز می‌توان نمودار را مطابق روش قبلی برای نمونه‌ها رسم کرد (شکل8) 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.



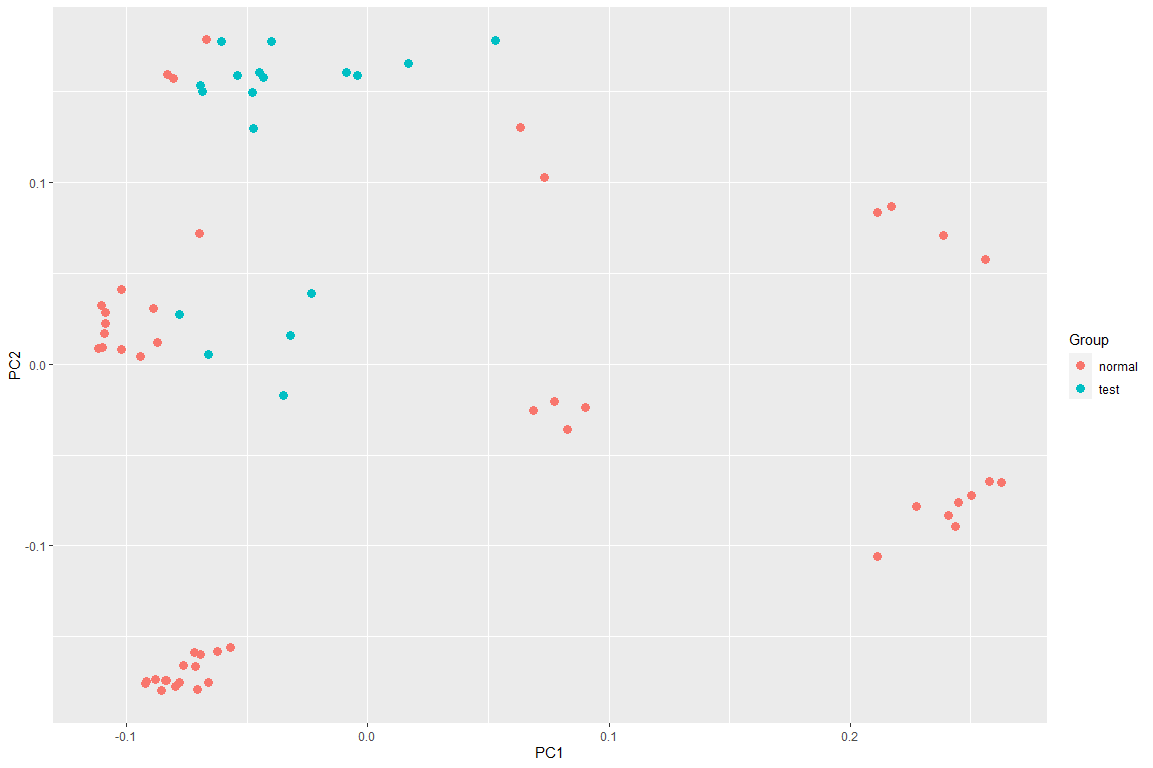
**شکل 4**- رسم داده‌ها مبتنی بر مولفه اول و دوم روش PCA. همانطور که در نمودار مشخص است تفاوت‌های موجود در مولفه اول به نوعی صرفا نشان می‌دهد که بعضی از ژن‌ها خیلی کم بیان شده‌اند یا اصلا بیان نشده‌اند و در مقابل بعضی دیگر بسیار زیاد بیان شده‌اند. این به نوعی داده‌ای ناکارآمد محسوب می‌شود.



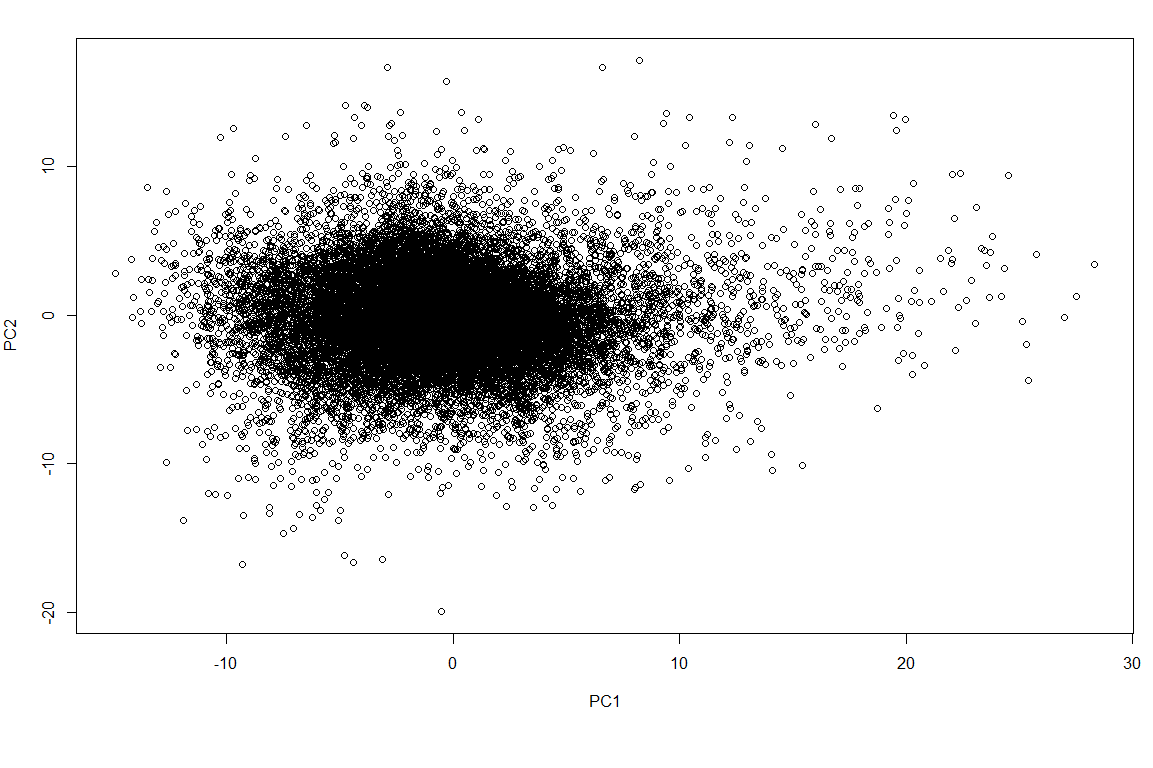
**شکل 6**- مولفه‌های مختلف به دست آمده مبتنی بر روش PCA پس از اعمال تغییر ابعاد. همانگونه که مشخص است دیگر تفاوت فاحش میان مولفه اول و دیگر مولفه‌ها نیست و با اعمال تغییر ابعاد، امکان این فراهم شد تا تاثیر بسیار بزرگی که ژن‌هایی که در تمامی نمونه‌ها بیان شده بودند و همچنین ژن‌هایی که اصلا بیان نشده بودند از بین برود و این روش بتواند تمایز بهتری متناسب با تفاوت بیان ژن‌ها ارائه بکند.



**شکل 5** - تفاوت نمونه‌ها مبتنی بر مولفه‌های اول و دوم به دست آمده از روش PCA، پیش از اعمال تغییر ابعاد.

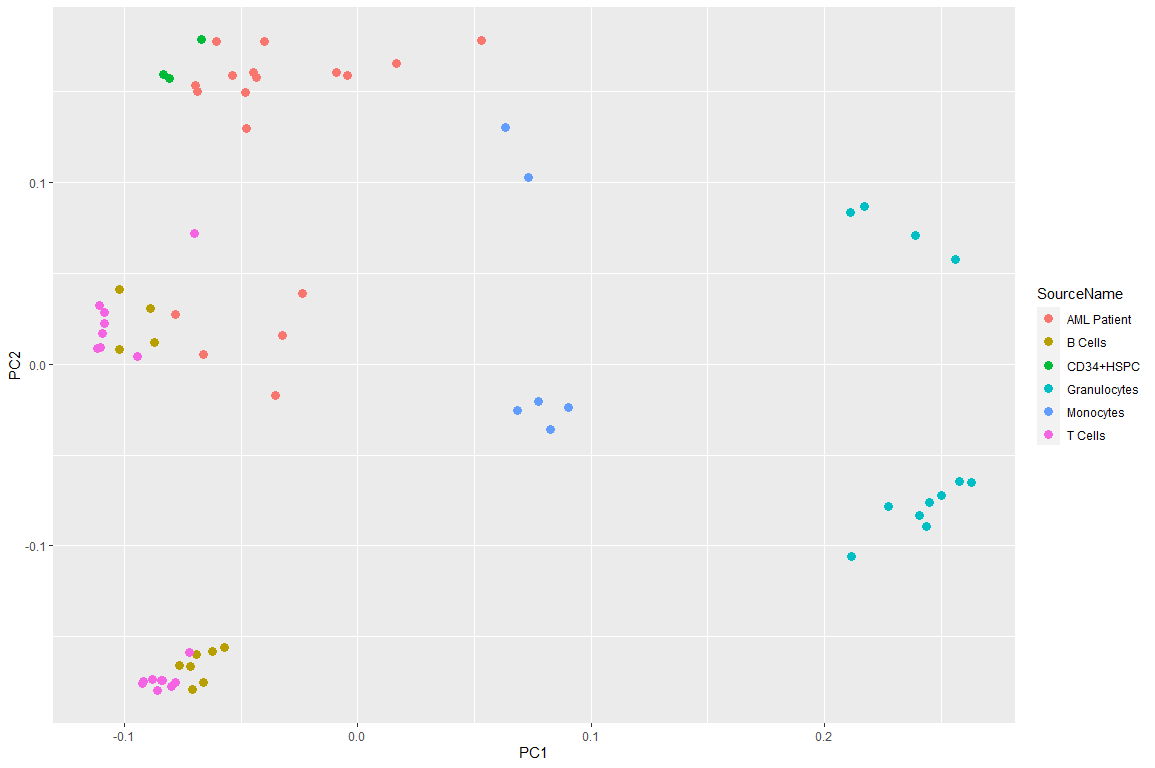


**شکل 8**- تفاوت نمونه‌ها مبتنی بر مولفه‌های اول و دوم به دست آمده از روش PCA، پس از اعمال تغییر ابعاد نمایش داده شده بر اساس دسته‌های سالم(normal) و بیمار(test).



**شکل 7**- 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.

همانگونه که در شکل9 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، نمی توانست به شکل مناسبی در ادامه به منظور بررسی تمایز ژن‌های بیان شده در نمونه‌های سالم و سرطانی مورد استفاده قرار گیرد.



**شکل 9**- 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.

به منظور رسم نمودار‌های 3 و 6 قطعه کد زیر مورد استفاده قرار می‌گیرد:

plot(pc)

به منظور رسم نمودار‌های 4 و 7 قطعه کد زیر مورد استفاده قرار می‌گیرد:

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

در قطعه کد بالا، x نشان دهنده ژن‌ها می‌باشد که بر اساس 2 مولفه اول برای رسم ارسال شده‌است.

به منظور رسم نمودار‌های 5 و 8 و 9 از کتابخانه ggplot2 استفاده شده‌است و قطعه کد زیر مورد استفاده قرار گرفته‌است (البته برای نمودار 9 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.

## **3-2. 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.

### **1-3-2. 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 شکل9 قابل مشاهده است.

ex.scale.cor <- cor(ex.scale)

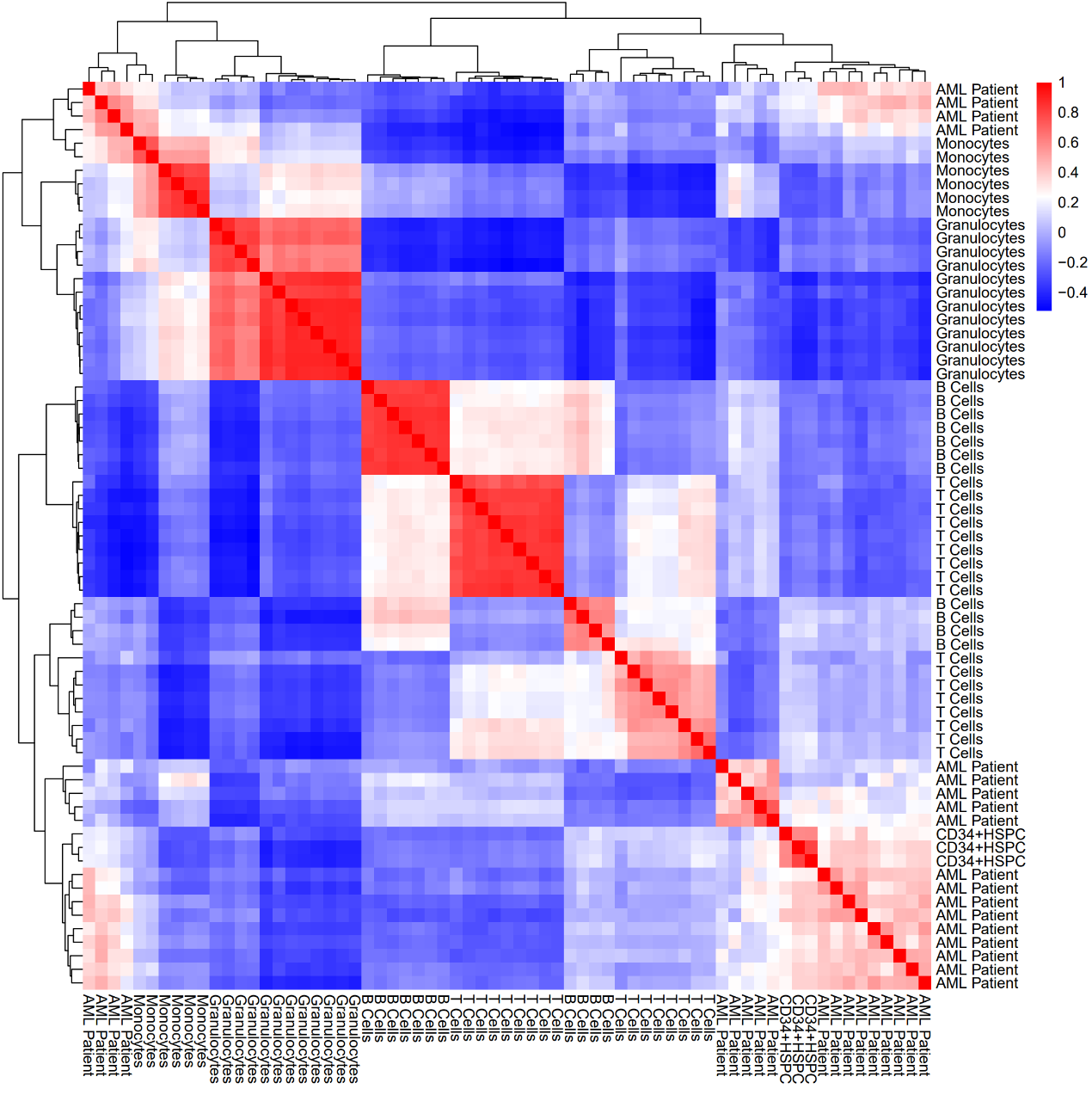
pheatmap(ex.scale.cor,

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 شکل9 مشخص است میزان همبستگی در نمونه‌های AML Patient و نمونه‌های سالم، به صورت کلی چهار نمونه AML Patient و 6 نمونه Monocytes و 12 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 10**- بررسی میزان همبستگی تمامی نمونه‌ها با یکدیگر از طریق Heatmap.

### **2-3-2. 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)

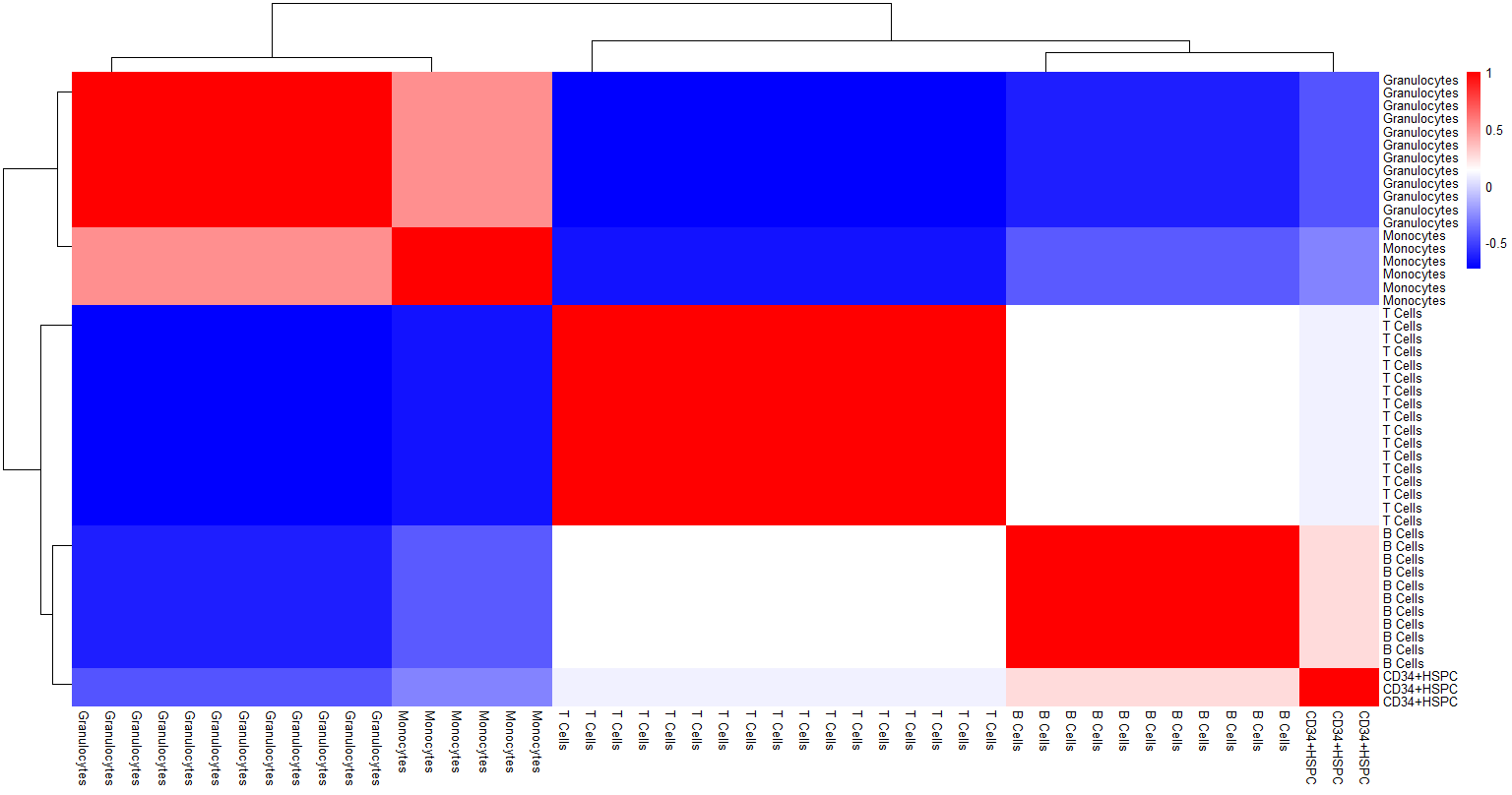
از طریق قطعه کد بالا با دوبار‌ترانهاده کردن ماتریس بیان، تمامی نمونه‌های AML Patient حذف شده و صرفا نمونه‌های نرمال در ماتریس b ذخیره می‌شوند. سپس کافیست همانگونه که در حالت کلی نمودار heatmap رسم شد، در این جا نیز برای نمونه‌های موجود در ماتریس b، همبستگی محاسبه شده و سپس نمودار رسم شود. نمودار خروجی در شکل11 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 شکل11 نمودار به صورت دقیق‌تری در فایل heatmap-normal.pdf موجود می‌باشد. همانگونه که در شکل11 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.

نکته قابل توجه در شکل11 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.



**شکل 11** - بررسی میزان همبستگی نمونه‌های سالم با یکدیگر از طریق Heatmap.

# **3. 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 جدول1 قابل مشاهده است.

|  |  |  |
| --- | --- | --- |
|  | نوع سلول | میزان همبستگی با AML Patient (عدد بین -1 تا 1) |
| 1 | Monocytes | 0.31974 |
| 2 | CD34+HSPC | 0.19717 |
| 3 | B Cells | 0.02744 |
| 4 | Granulocytes | 0.02437- |
| 5 | T Cells | 0.06474- |

**جدول 1**- مرتب‌سازی میزان همبستگی هرکدام از سلول‌ها با AML Patient. میزان همبستگی بین -1 تا 1 می‌باشد و هرچه به 1 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 0.01 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[[5]](#footnote-5) 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 معنی دار 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 1 و در نقطه مقابل عدد -1 در نظر گرفته شده‌است. مقدار logFC نسبت لگاریتمی سرطانی به سالم می‌باشد که در صورتی که بیشتر از 1 باشد به این معنی است که بیان آن ژن در نمونه سرطانی بیشتر بوده است و در نقطه مقابل در صورتی که کم‌تر از -1 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 جدول1، 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.

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.

# **4. آنالیز Gene ontology و pathway‌ها**

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 داشته‌اند، پرداخته شده‌است و pathway‌ها و Gene ontology مربوط به هر دسته به کمک وبسایت Enrichr مورد بررسی قرار گرفته‌اند.

### **1-4. 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.

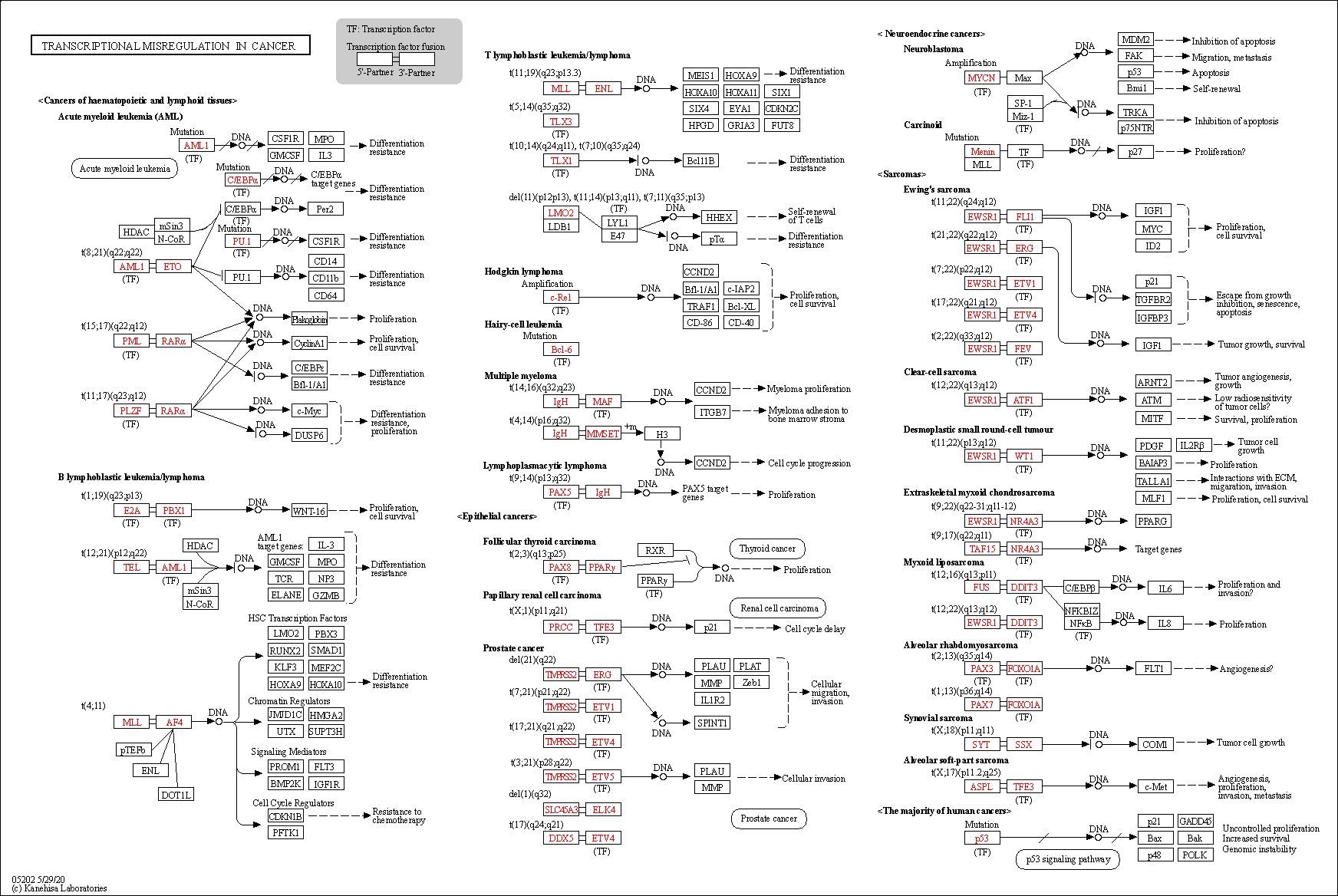
### **1-1-4. بررسی pathway‌های مرتبط با دسته AML در مقابل Monocytes که افزایش بیان داشته‌اند**

با بررسی ژن‌هایی که افزایش بیان در نمونه‌های بیمار داشته‌اند در مقایسه با Monocytes‌ها(سالم)، 1489 ژن به دست می‌آید. در مقایسه با pathway‌های کتاب‌خانه Kegg[[6]](#footnote-6) 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. (جدول2 شامل 13 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[[7]](#footnote-7) 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.(شکل12-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 |
| 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 2**- مجموعه‌ای از مرتبط‌ترین pathway‌ها با ژن‌های افزایش بیان یافته در سلول‌های بیمار در کتاب‌خانه Kegg.

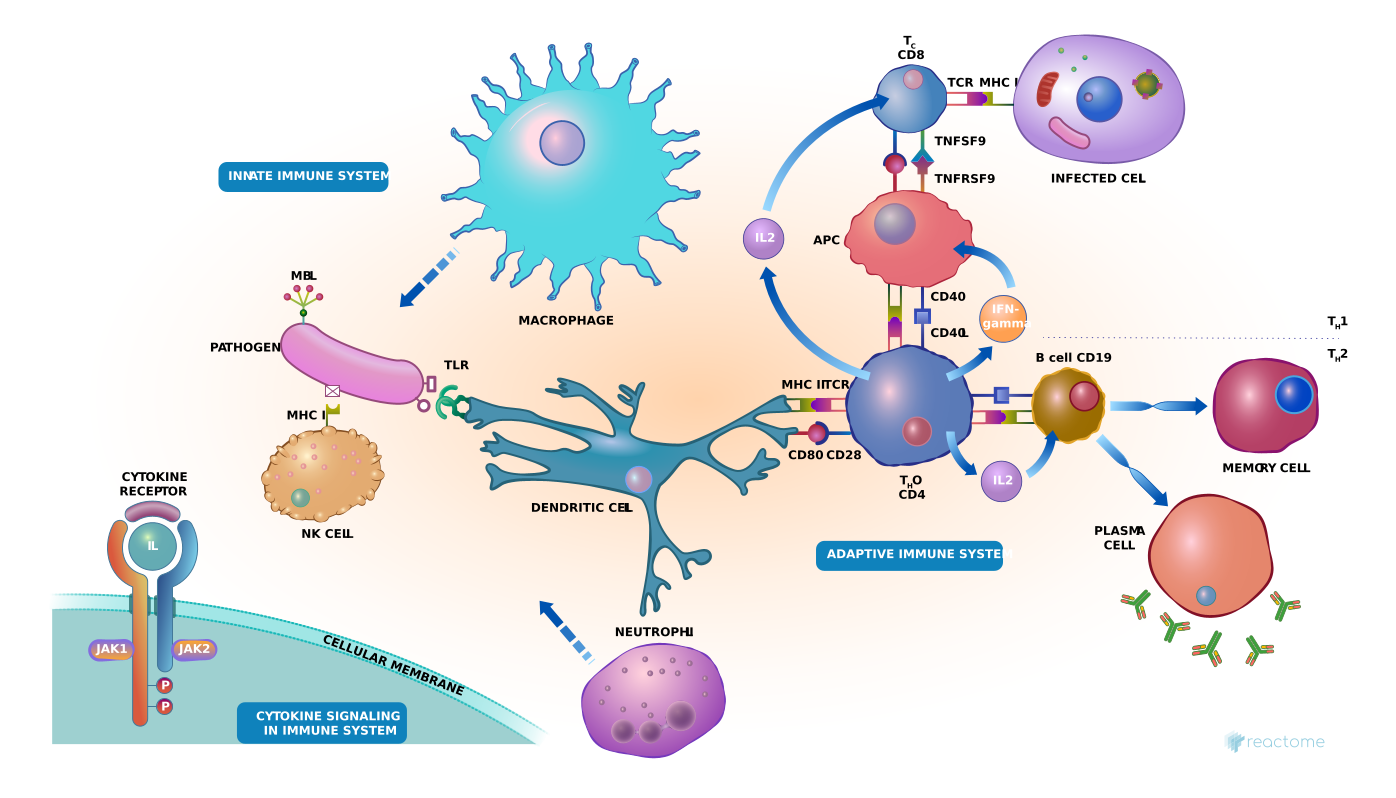
مشابه مقایسه بالا در کتاب‌خانه WikiPathway[[8]](#footnote-8) نیز انجام شده‌است و فایل جدول آن در دایرکتوری result/pathways/amu ضمیمه شده‌است.



**شکل 12**- برخی از جهش‌هایی که در اثر سرطان باعث ایجاد خطا در فرآیند رونویسی و کنترل آن می‌شود. سمت چپ بالا مربوط به AML می‌باشد.

### **2-1-4. 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)، 1415 ژن به دست می‌آید. در مقایسه با pathway‌های کتاب‌خانه Reactome[[9]](#footnote-9) بیشترین مواردی که به عنوان ارتباط یافته شده‌است که با مقدار adj.P.Val بسیار پایینی هستند، همانگونه که در جدول3 مشخص است هر 8 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 شکل13 در دایرکتوری result/pathways/amd ضمیمه شده‌است.



**شکل 13**- تصویری از pathway مرتبط با سیستم ایمنی (Immune System Homo sapiens R-HSA-168256) از کتاب خانه Reactome.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 3**- عملکرد‌های مرتبط با ژن‌های کاهش بیان یافته در AML نسبت به Monocytes. جدول کامل‌تر موارد فوق در فایل reactome\_AML\_Monocytes\_Down.xlsx در دایرکتوری result/pathways/amd ضمیمه شده‌است.

### **1-4. بررسی Gene ontology**

منظور از Gene ontology

در ادامه در سه زیربخش به دسته‌های مختلف مقایسه و بررسی pathway‌های به دست آمده در هرکدام پرداخته شده‌است.

### **2-1-4. بررسی Gene ontology مرتبط با دسته AML در مقابل Monocytes**

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. جدول4 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 4**- پروسه‌های زیستی مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های سرطانی نسبت به Monocytes. جدول کامل‌تر در فایل BiologicalProcess\_AML\_Monocytes\_Up.xlsx در دایرکتوری result/ontology 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. (جدول5 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 5**- عملکرد‌های مولکولی مرتبط با ژن‌های افزایش بیان داشته در نمونه‌های بیمار نسبت به Monocytes. جدول کامل‌تر در فایل MolecularFunction\_AML\_Monocytes\_Up.xlsx در دایرکتوری result/ontology ضمیمه شده‌است.

با بررسی کامپوننت‌های 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 کامپوننت‌های سلولی به ژن‌های افزایش یافته در جدول6 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 6** - کامپوننت‌های مرتبط با ژن‌های افزایش بیان یافته در بیماران نسبت به Monocytes. لیست کامل‌تر این موارد در فایل CellComponent\_AML\_Monocytes\_Up.xlsx در دایرکتوری result/ontology ضمیمه شده‌است.

1. AML: Acute myeloid leukemia [↑](#footnote-ref-1)
2. https://en.wikipedia.org/wiki/Acute\_myeloid\_leukemia [↑](#footnote-ref-2)
3. Q[uantile Normalization](https://www.biostars.org/p/296992/) [↑](#footnote-ref-3)
4. PCA: Principal Component Analysis [↑](#footnote-ref-4)
5. Benjamini-Hochberg [↑](#footnote-ref-5)
6. KEGG: www.genome.jp/kegg/ [↑](#footnote-ref-6)
7. Transcriptional misregulation in cancer [↑](#footnote-ref-7)
8. WikiPathway: www.wikipathways.org [↑](#footnote-ref-8)
9. Reactome: www.reactome.org [↑](#footnote-ref-9)