به نام خدا

درس مقدمه ای بر بیوانفورماتیک

اساتید: دکتر سمیه کوهی – دکتر علی شریفی زارچی

دانشجو: مهدی منوچهری

شماره دانشجویی:۴۰۰۲۱۱۵۹۲

مقدمه

Microarray یکی از روش های بیولوژی مولکولی است که برای ارزیابی سطح mRNA و بیان هزاران ژن به طور همزمان مورد استفاده قرار میگیرد. از طرفی سرطان بیماری ای است که در آن چندین ژن درگیر هستند.بنابراین استفاده از microarray برای اندازه گیری سطح بیان چندین ژن می‌تواند برای بررسی الگو بیان ژن در سرطان مفید باشد چرا که در سرطان الگو بیان ژن بسیاری از ژن ها تغییر می‌یابد.داده های خام اولیه از NCBI بدست آمده اند.به منطور ایجاد ست های همگون اطلاعاتی برخی اعمال مانند تقسیم بندی و نرمال کردن روی این داده ها انجام شد.به دنبال عمل دسته بندی اطلاعات نتایج ژن های مشترکی را نشان داد بیان همزمان داشتند در هر دو بافت. در نتایج دو لیست وجود دارد که ژن ها بر اساس بیانشان به دسته طبقه بندی می‌شوند. در یک لیست ژن هایی که بیش از حد نرمال بیان شده‌اند قرار دارند و در لیست دیگر ژن هایی که کمتر از حد معمول بیان شده‌اند.

پایپلاین اجرایی

دسته بندی نمونه ها

لوکمیا انواع مختلفی دارد. یکی از انواع این سرطان ، لوکمی حاد مغز استخوان AML است. با تحلیل دادگانی که در اختیار داریم ، ژن هایی را که در این نوع سرطان نقش موثرتری دارند را به دست آوریم . به این منظور در مجموعه دادگان ، داده هایی را که Phenotype آنها Normal است را گروه نرمال و نمونه هایی را که name Source آن ها patient AML است را گروه تست در نظر می‌گیریم

در این مرحله گروه تست شامل ۱۸ sample و گروه نرمال شامل ۴۹ sample شد. که مجموعا ۶۷ sample داریم.

کنترل کیفیت داده ها

تشخیص نرمالایز بودن ماترس بیان(ex) : با توجه به اینکه مقدار max value این ماتریس عدد ۱۳,۷۶۱۵۴ می‌باشد نیازی به لگاریتمی کردن مقادیر نمی‌باشد.
در صورت نیاز به صورت مقابل عمل می‌کنیم:

```
ex <- log(ex+1)
exprs(gset) <- ex
```

رسم boxplot:
به صورت مقابل boxplot رسم می‌کنیم.

```
pdf("Results/boxplot.pdf",width = 64,)
boxplot(ex)
dev.off()
```

نتیجه کد در فایل boxplot نشان دهنده این است که نمونه ها نرمالایز شده می‌باشند .
در صورتی که نرمالایز نشده بودن می‌تونستیم به صورت زیر عمل کنیم:

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

pdf("Results/boxplotnormal.pdf",width = 64,)
boxplot(ex)
dev.off()
```

که نتیجه در فایل boxplotnormal قابل مشاهده است.

کاهش ابعاد داده

با استفاده از قطعه کد مقابل نمودار های فایل pc بدست آورده ایم.

```
pc <- prcomp(ex)

pdf("Results/pc5.pdf")

plot(pc)

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

dev.off()
```

هر نقطه در این نمودار بیانگر یگ ژن می‌باشد.و طوری کاهش بعد داریم(عکس گرفتن از فضا) که در راستای pc1 بیشترین variation را می‌بینیم.و بعد سپس عمود بر آن یعنی pc2 .

ژن‌هایی که اصلا بیان نشده اند (میانگین در همه جا صفر) ، در همه جا میانگین ثابت دارند این دو نوع دو سر طیف را در نمودار تشکیل می-دهند.

در نتیجه pc1 اطلاعات خوبی به ما نمی‌دهد.در صورتی که انتظار داریم مهمترین اطلاعات از داده به ما بدهد.

برای بهبود نمودار همه ژن ها را از میانگین بیان همان ژن کم می‌کنیم. در واقع میانگین بیان همه ژن ها را صفر می‌کنیم.

در این صورت pc1 فقط بر اساس تفاوت ها می‌باشد.

این تغییرات در ماتریس بیان اصلی انجام نمی‌دهیم.

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

با استفاده از t(ex) یک بار ex را Transpose می‌کنیم چون تابه scale فقط روی ستون ها کار می‌کند.(ژن ها را قصد داریم scale کنیم.)

تلبع scale برای هر ژن (ستون) میانگین صفر می‌کند.(همه مقادیر منها میانگین می‌کند.)

مجدد با استفاده از قطعه کد

```
ex.scale <- t(scale(t(ex),scale = F))
pc <- prcomp(ex.scale)
pdf("Results/pc_scaled.pdf")
plot(pc)
plot(pc$x[,1:2])
dev.off()
```

در فایل pc_scaled می‌بینیم که pc1 همه variation ها را در خود ندارد

همچنین توزیع ژن ها نیز منطقی تر شده است.

گام بعدی

در قدم بعدی باید نمودار pc ، sample ها را زسم کنیم

مراحل زیر را انجام می‌دهیم:

```
gr <- c(rep("AML Patient",13),"Granulocytes","Granulocytes","B Cells","T
Cells","Granulocytes","Granulocytes",
rep("Monocytes",2),"B Cells","T Cells",rep("T Cells",2),
rep("T Cells",2), "B Cells","T Cells","B Cells","T Cells",
CD34","CD34","CD34",rep("Granulocytes",7),rep("AML Patient",2)"
```

```
T Cells",rep("AML Patient",3),rep("B Cells",7),"T Cells",rep("Monocytes",4),"Granulocytes",rep("T ",
Cells",7))

pcr <- data.frame(pc$r[,1:3],Group=gr)
pdf("Results/pca_samples1.pdf")
ggplot(pcr,aes(PC1, PC2,color=Group)) + geom_point(size=3) + theme_bw()
dev.off()
```

خروجی در فایل pca_samples1

مشاهده می‌کنیم که pc1 به خوبی توانسته است Granulocytes و Monocytes را از سایر نمونه ها جدا کند.

همچنین مشاهده می‌کنیم که T Cells و B Cells ها شامل دو زیر گروه هستند که شباهت زیادی هم به یکدیگر دارند.

AML نیز به دو گروه تقسیم شده اند که یک گروه شباهت زیادی با CD34 شباهت زیادی دارند.

همچنین متوجه میشویم داده هایی که داریم داده های خوبی از لحاظ آماری است.

.

بررسی همبستگی بین نمونه ها

با استفاده از تابع heatmap همبستگی بین نمونه ها را بدست میاورم.

```
pdf("Results/CorHeatmap3.pdf",width = 20,height = 20)
pheatmap(cor(ex),labels_row = gr,labels_col = gr)
dev.off()
```

نتیجه در فایل CorHeatmap3 وجود دارد.

مشاهده میکنیم که Granulocytes ها به خودشون خیلی شبیه هستند و از سایرین تفاوت زیادی دارند در واقع گروه مناسبی برای مقایسه کردن نیستند.

بهترین مقایسه می تواند بین CD34 یا T Cells و AML Patient باشد.

همچنین شباهت دو به دو AML ها از شباهت دو به دو سایر گروه ها کمتر است دلیل میتوند بخاطر میزان تفاوت غده های سرطانی است که حتی در سلول های یک بدن هم بسیار زیاد است

بررسی تمایز در بیان ژن ها

با توجه به نمودار heatmap نمونه های aml و CD34 تشابخ بیشتری به همدیگر دارند بنابراین تمایز در بیان ژن ها را در این دو نمونه بررسی می‌کنیم.

```
gr <- factor(gr)

gset$group <- gr

design <- model.matrix(~group + 0, gset)

colnames(design) <- levels(gr)

fit <- lmFit(gset, design)


cont.matrix <- makeContrasts(AML-CD34, levels=design)

fit2 <- contrasts.fit(fit, cont.matrix)

fit2 <- eBayes(fit2, 0.01)

tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)

tT <- subset(tT,
select=c("Gene.symbol","Gene.ID","adj.P.Val","P.Value","log
FC"))

write.table(tT,"Results/AML__CD34.txt", row.names=F,
sep="\t",quote=F)
```

در این کد از مدا بیزین استفاده کردیم و یک درصد از ژن هایی که به صورت خاص تماییز بیان ژنی داشتند را در فایل AML__CD34 بدست آورده ایم.

حالا می‌توانیم ژن هایی که به شکل معنی داری بیان کمتری یا بیشتری دارند را بدست بیاوریم.

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

aml.up.genes <- unique(aml.up$Gene.symbol)

write.table(aml.up.genes, file = "Results/AML_CD34_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(aml.down$Gene.symbol)
```

write.table(aml.down.genes, file = "Results/AML_CD34_down.txt",quote=F , row.names = F, col.names = F)

در قطعه کد اول p-value کمتر از 0.05 و LOGFC بیشتر از ۱ می‌باشد همچنین در قطعخ کد دوم p-value کمتر از 0.05 و LOGFC کمتر از ۱ می‌باشد.

نتیجه در فایل های AML_CD34_UP.txt و AML_CD34_down.txt وجود دارد.

آنالیز ontology gene و pathway ها

در این قسمت با استفاده از داده‌های بخش قبل درسایت Enrichr به آنالیز gene ontology و pathway می‌پردازیم.

ابتدا داده های مربوط به AML_UP_Gene را وارد می‌کنیم.

در دیتا بیس TRANSFAC and JASPAR PWMS می‌تونیم مشاهده کنیم که هر ژن کدوم ژن ها را UP_Down می‌کند.

نتیجه حاصل در فایل TRRUST_Transcription_Factors_2019_table گذاشته شده است.

برای مثال SPI1 human را در مقالات اخیر جستجو می‌کنیم و تاثیر آن بر aml را متوجه میشویم.

فاکتور رونویسی Spi1 یک تنظیم کننده کلیدی در بسیاری از مراحل خون سازی است و خود نوسازی سلول های بنیادی خونساز را محدود می کند. عدم تنظیم بیان یا فعالیت آن به سرطان خون کمک می کند، که در آن Spi1 می تواند یک انکوژن یا یک سرکوب کننده تومور باشد.

Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5664389/

در گام بعدی به pathway analyses می‌پردازیم.
دیتابیس Reactom به ما Pathway سیستم ایمنی را گزارش می‌کند که منطقی می‌باشد به دلیل اینکه aml مربوط به سیستم ایمنی بدن است.
خروجی مربوطه در فایل Reactome_2016_table گزارش شده است.

در قسمت ontology از پایگاه داده Jensen tissues مشاهده می‌کنیم که داده ها مربوط به blood می‌باشد که کاملا مورد انتظار است.
فایل مربوط به این گزارش: Jensen_TISSUES_table

همچنین مراحل فوق را برای داده های AML_UP_Gene می‌توانیم تکرار کنیم که اسامی فایل های مربوطه در قسمت زیر آمده است.

TRANSFAC_and_JASPAR_PWMs_table
WikiPathway_2021_Human_table
GO_Molecular_Function_2021_table

موارد دیگر

بهتر بود که نمونه های CD34 بیشتری برای تحلیل داشتیم تا به جواب های قابل اعتماد تری برسیم.
افزایش سلول های CD34 به نشانه بدخیم بودن aml و قابل بازگشت بودن آن است.
Aml بیماری است که سیستم ایمنی تحریک می‌کند در راستای یک التهاب.