به نام خدا

فاز اول پروژه مقدمه بر بایوانفورماتیک

اساتید درس

سمیه کوهی علی شریفی زارچی

اعضای تیم

رامین روشن 401206571

طاها محمدزاده 401202918

عرفان اسماعیلی 401212295

1 در مورد microarray ، روش کار آن و فرمت داده های خروجی آن به طور مختصر توضیح دهید.

دستگاه است که از یک چیپ تشکیل شده است که بر روی چیپ تعداد زیاد پیکسل قرار دارد و بر روی هر پیکسل توالی از DNA تک رشته ای که به سطح سیلیکونی چسبیده است . در این روش ابتدا ما یک سری نمونه از جامعه برای مثال افراد سالم یا غیر سالم بدست می اوریم سپس RNA نمونه ها را به cDNA تبدیل می کنیم و سپس به روش شیکینگ یا سانیکیشن آن را تکه تکه می کنیم و دقت شود در این روش هر دفعه مکان های متفاوت cDNA تکه می شود و سپس با فلورسانس تکه ها را رنگ می کنیم و بعد از آن با گرم کردن ، رشته از هم باز می شوند و به یک تک رشته تبدیل می شوند . در آخر تک رشته ای حاصل را بر روی ماکرو اررای می ریزیم و هر تکه به مکمل خود می چسبد و اگر نتوانست بچسبد یا مکمل نداشت آن را دور می ریزیم ، وسپس چیپ را به یک اسکنر لیزی می دهیم که نشان دهد هر پیکسل چقدر رنگ فلورسانت دارد و میزان رنگ هر پیکسل نشان گر میزان بیان پیکسل است.

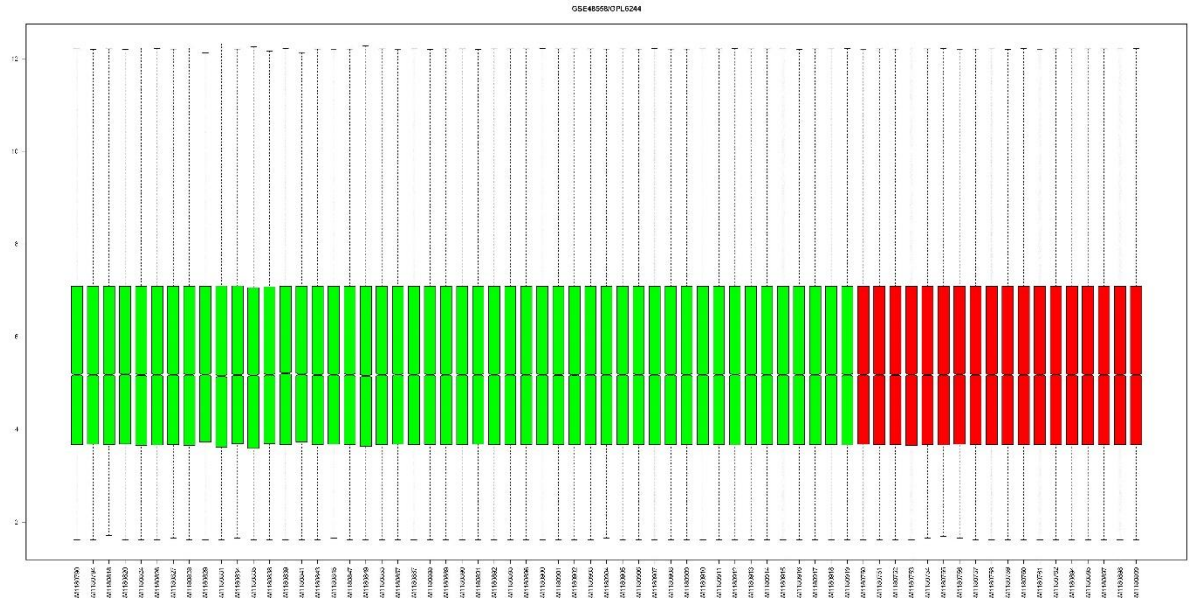
تصاویر ایجاد شده تبدیل به یک ماتریس از اعداد می شوند که نمونه ها برابر ستون ها و probe ها به عنوان سطر در نظر می گیریم و سپس بر روی این جدول تحلیل آزمایش انجام می دهیم.

دقت شود microarray برای تشخص جهش در در تکه خاص از DNA مورد استفاده نیز قرار می گیرد اما به جای RNA از آن تکه مورد نظر از DNA استفاده می شود.

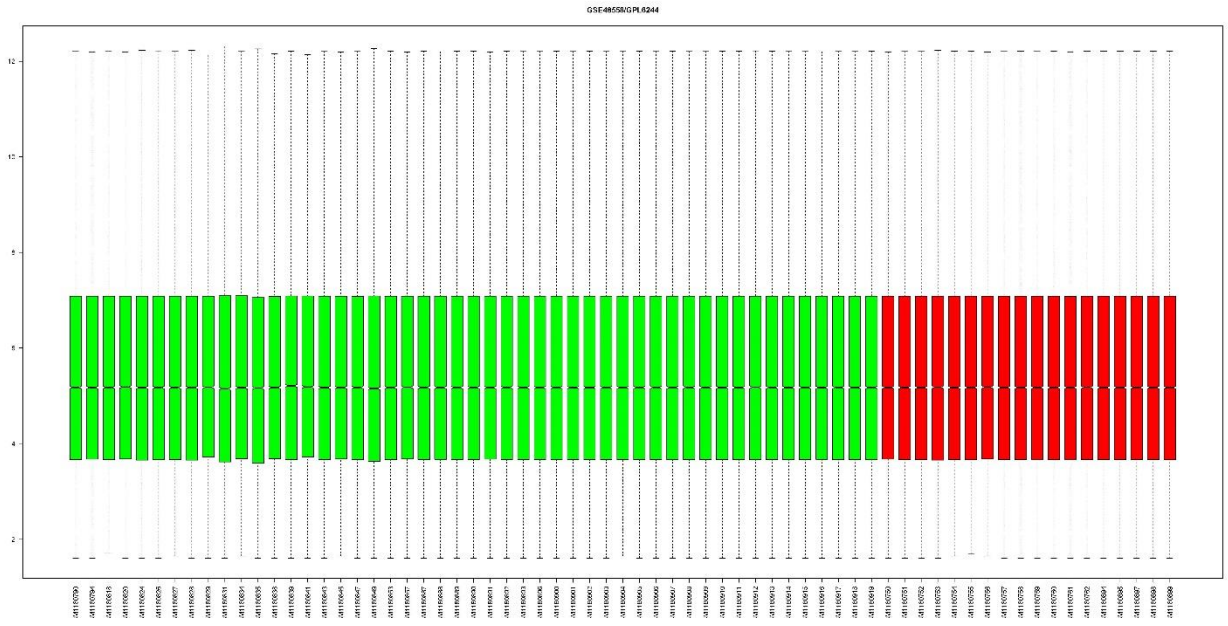
فرمت فایل ها

هر گونه فرمت که بتوان داده ها را به صورت spreadsheet یا matrix که به صورت tab-delimited بتوان ذخیره کرد برای مثال txt یا فایل پردازش native خود دستگاه برای مثال chp قابل قبول است اما فایل Excel قابل قبول نیست.

2. داده هایی که phenotype آنها normal است را به عنوان داده های گروه سالم و داده هایی که source name آن ها AML patient است را به عنوان گروه داده های بیمار در نظر بگیرید. داده های اولیه ممکن است برای تحلیل های بعدی آماده نباشند. کیفیت داده ها را از جنبه هایی که به نظرتان میرسد بررسی کنید و در صورت لزوم تغییرات لازم را روی آن ها اعمال کنید .(راهنمایی: برای مثال نرمال سازی داده ها.) برای هر ویژگی ای که کیفیتش را کنترل میکنید ذکر کنید که این کنترل چه لزومی دارد، مراحل بررسی و کنترل خود را گزارش کنید (برای مثال نمودارها و ...) و اگر لازم بود تغییری در داده ها ایجاد کنیم،تاثیر تغییرات را گزارش کنید.



Normalized data



Raw data

با بررسی داده ها می توان دریافت که میانه داده ها و چارک ها بسیار به هم نزدیک هستن و می توان با مشاهده از نمودار دریافت که داده ها همگی از یک توزیع پیروی می کنند . اگر ما به داده نرمالیز شده دقت کنیم تفاوت چندانی نسبت به داده خام ندارد.

به دلیل که این که مینیم و ماکسیمم داده ها تفاوت چندانی نداشتند پس نیازی به استفاده از Log2 transform نیست.

هدف از نرمالایز کردن داده ها حذف داده های بدون ساختار وتکراری است و درکل می توان گفت که قصد ما همسان سازی داده هاست در این سوال هم ما با نرمالایز کردن داده ها داده هارا در یک اسکیا یکسان قرار می دهیم

```
data.normalized = normalizeQuantiles(data)



pdf("result/boxplot_health.pdf",width = 32,height = 16)
boxplot(data, boxwex=0.7, notch=T, main=title, outline=FALSE, las=2,col=colors)
dev.off()

pdf("result/boxplot_normalized.pdf",width = 32,height = 16)
boxplot(data.normalized, boxwex=0.7, notch=T, main=title, outline=FALSE, las=2,col=colors)
dev.off()
```

3. لزوم کاهش ابعاد داده ها چیست؟ سه روش مختلف برای کاهش ابعاد را انتخاب کرده و نتایج حاصل از هر سه روش را گزارش کنید. سپس با مقایسه این نتایج، روشی که بهترین خروجی را نتیجه داده است، انتخاب کنید. دلایل انتخاب روش بهتر را ذکر کنید. راهنمایی: برای مثال می توانید کاهش ابعاد را با سه روش PCA، MDS و tSNE انجام دهید.

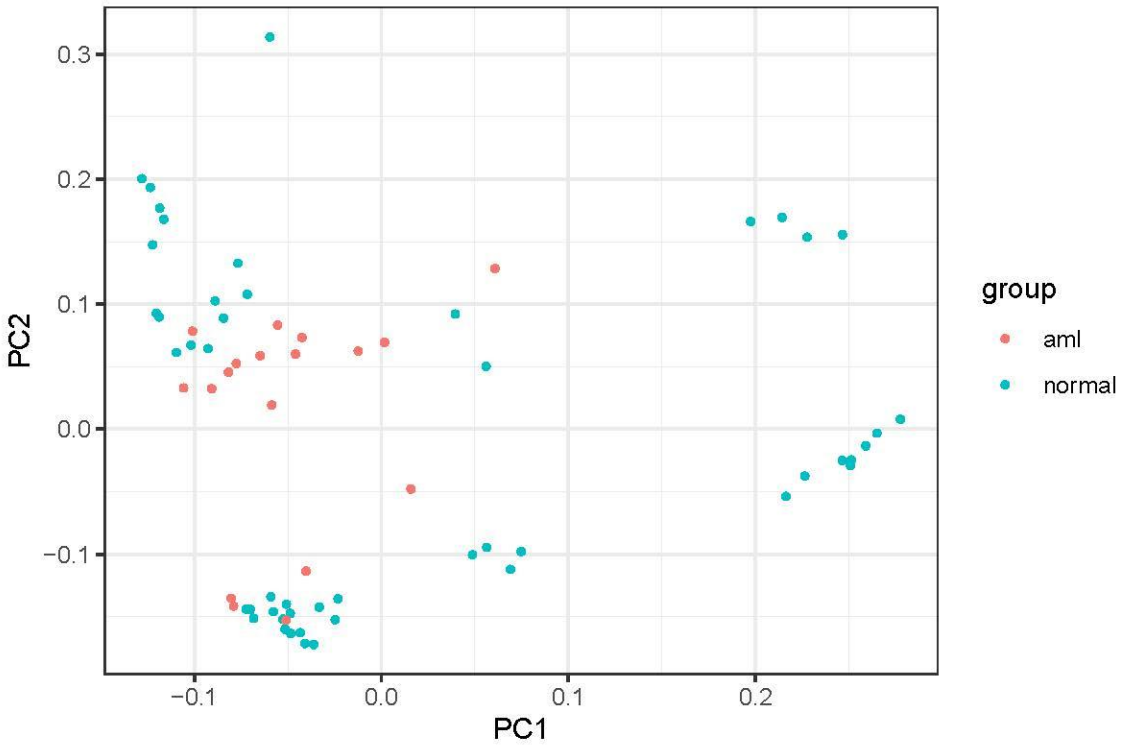
کاهش ابعاد باعث کاهش هزینه در محاسباتی و فضای ذخیره سازی می شود

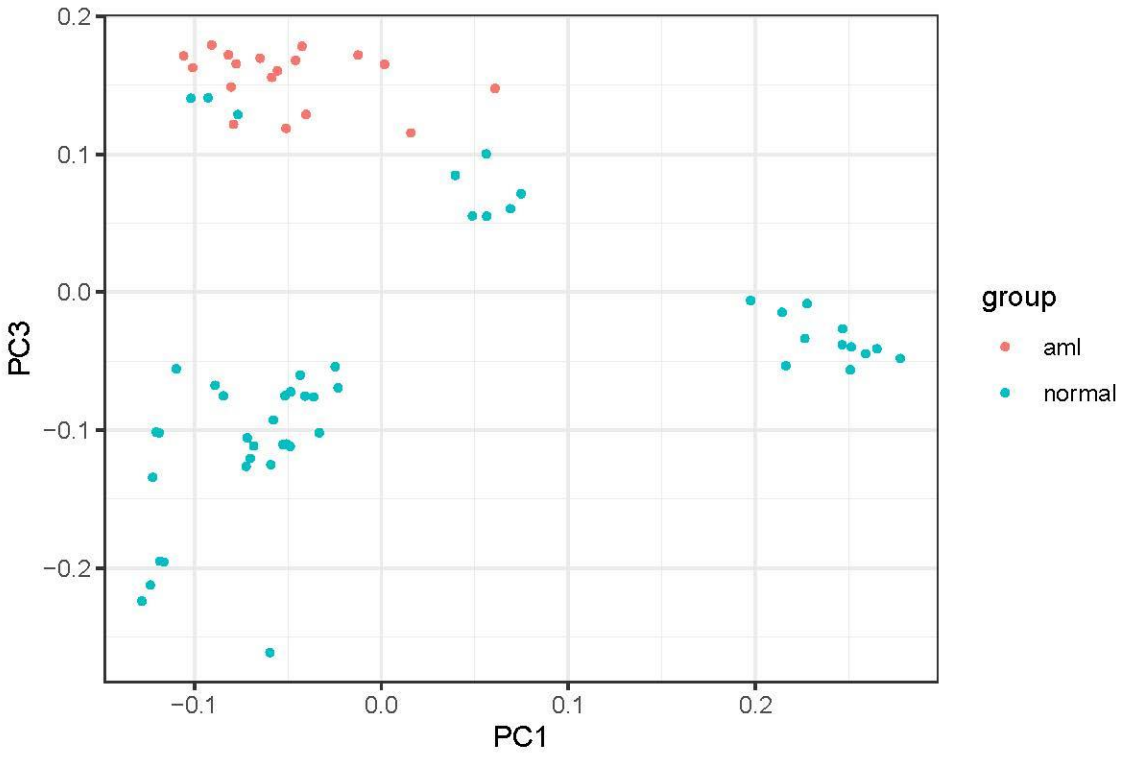
امکان رخداد over-fitting را کاهش می دهد و از Curse of dimensionality جلو گیری می کند

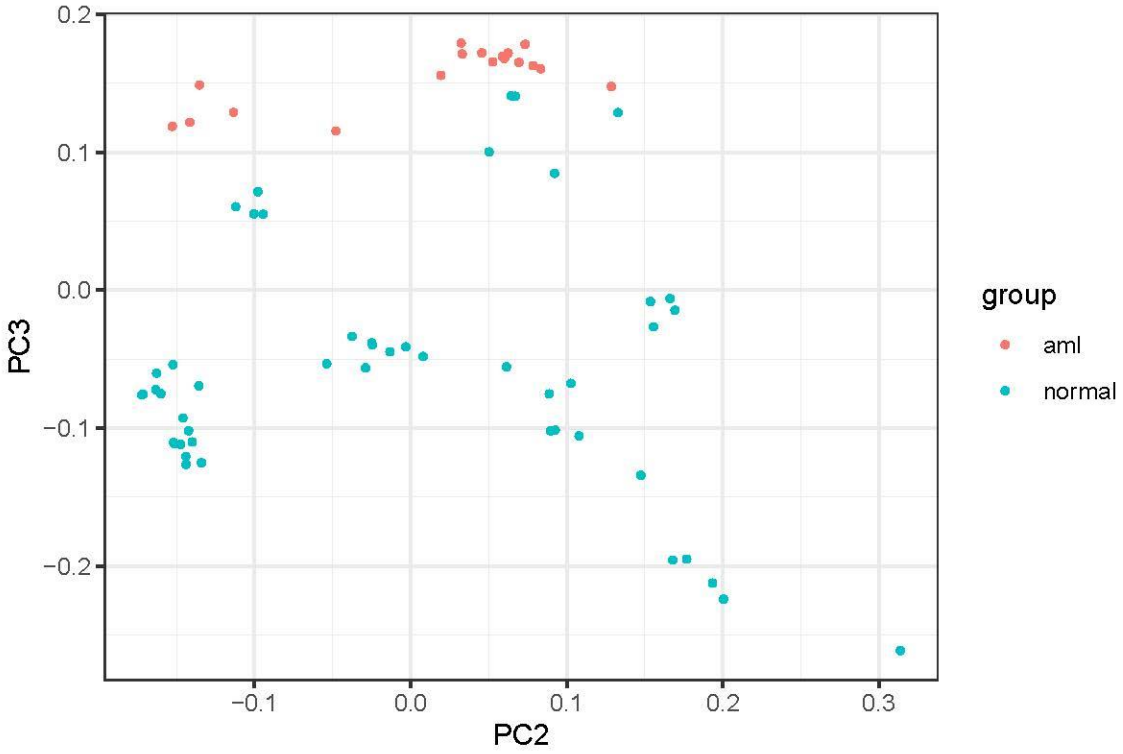
بعضی از ویژگی به وسیله بقیه ویژگی ها قابل محاسبه هستن(وابستگی) و با کاهش ابعاد ویژگی های اضافه حذف می شوند و علاوه بر آن ویژگی های بی تاثیر نیز حذف می شوند

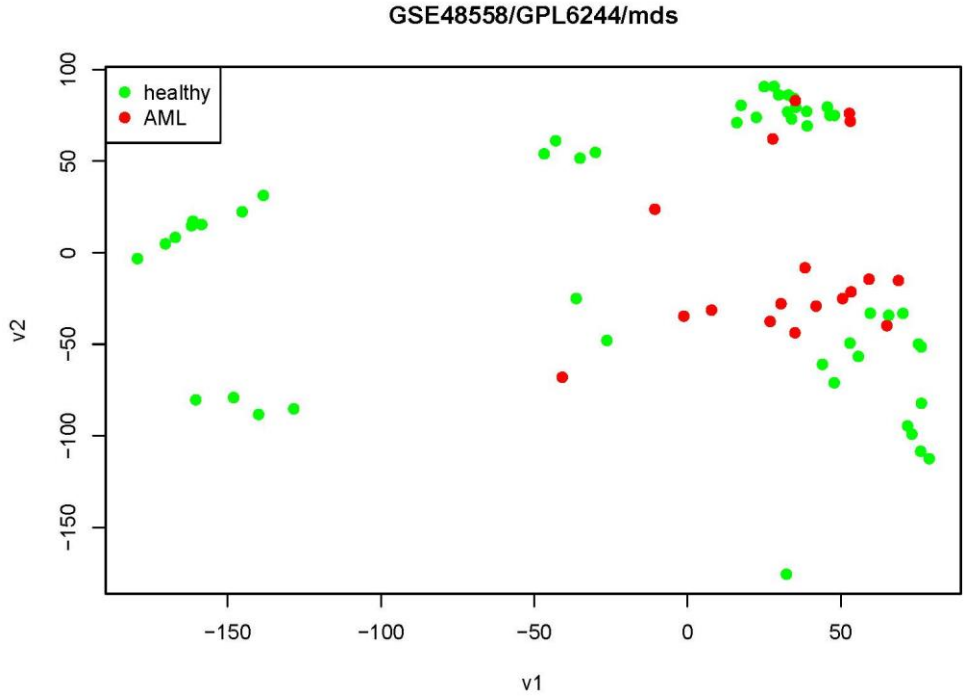
امکان رسم نمودار و اشکال هندسی در فضای 2 یا 3 بعدی را ایجاد می کند

ما از 4 روش pca,mds,tsne,umap برای کاهش ابعاد استفاده کردیم که نتایج به شرح زیر است.

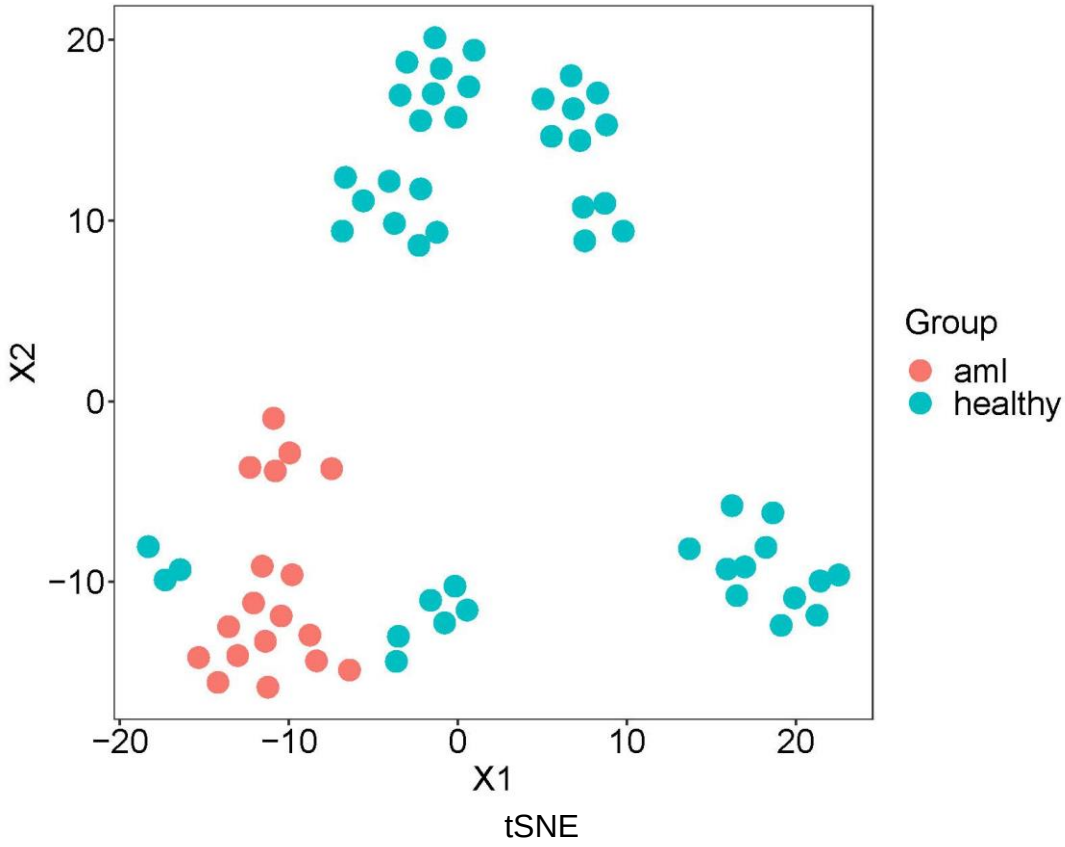

0.3
0.2
0.1
0.0
−0.1
PC2
−0.1
0.0
0.1
0.2
PC1
group
aml
normal

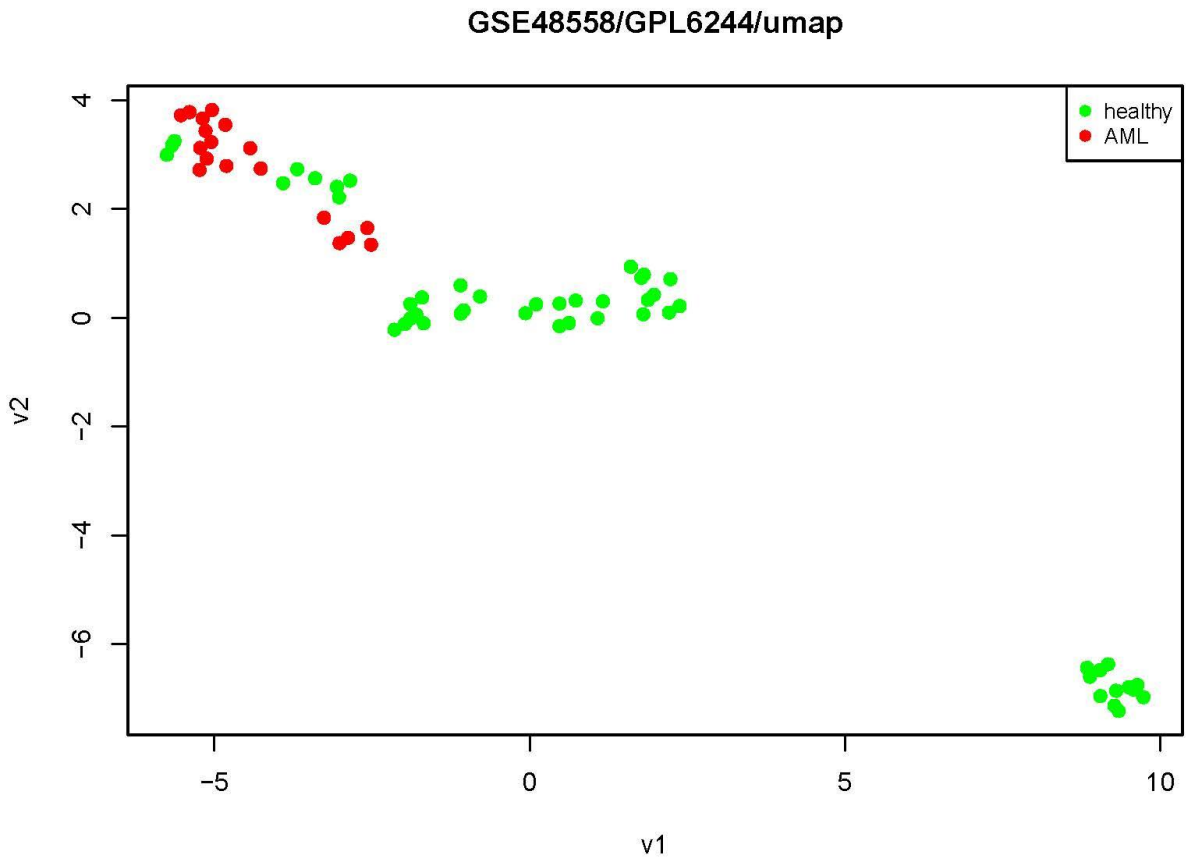

0.2
0.1
0.0
−0.1
−0.2
PC3
−0.1
0.0
0.1
0.2
PC1
group
aml
normal





mds_classical


20
10
0
−10
X2
−20
−10
0
10
20
X1
tSNE
Group
aml
healthy



Uniform Manifold Approximation and Projection(umap)

با توجه به نمودار های بالا در میابیم که tsne در تفکیک داده ها بهترین عملکرد را نسبت به بقیه روش ها از خود نشان داده است و بعد از آن PC2 و PC3 بهترین تفیک را داشتند، به وسیله این دو روش می توان به خوبی افراد سالم و بیمار را از هم جدا کرد.

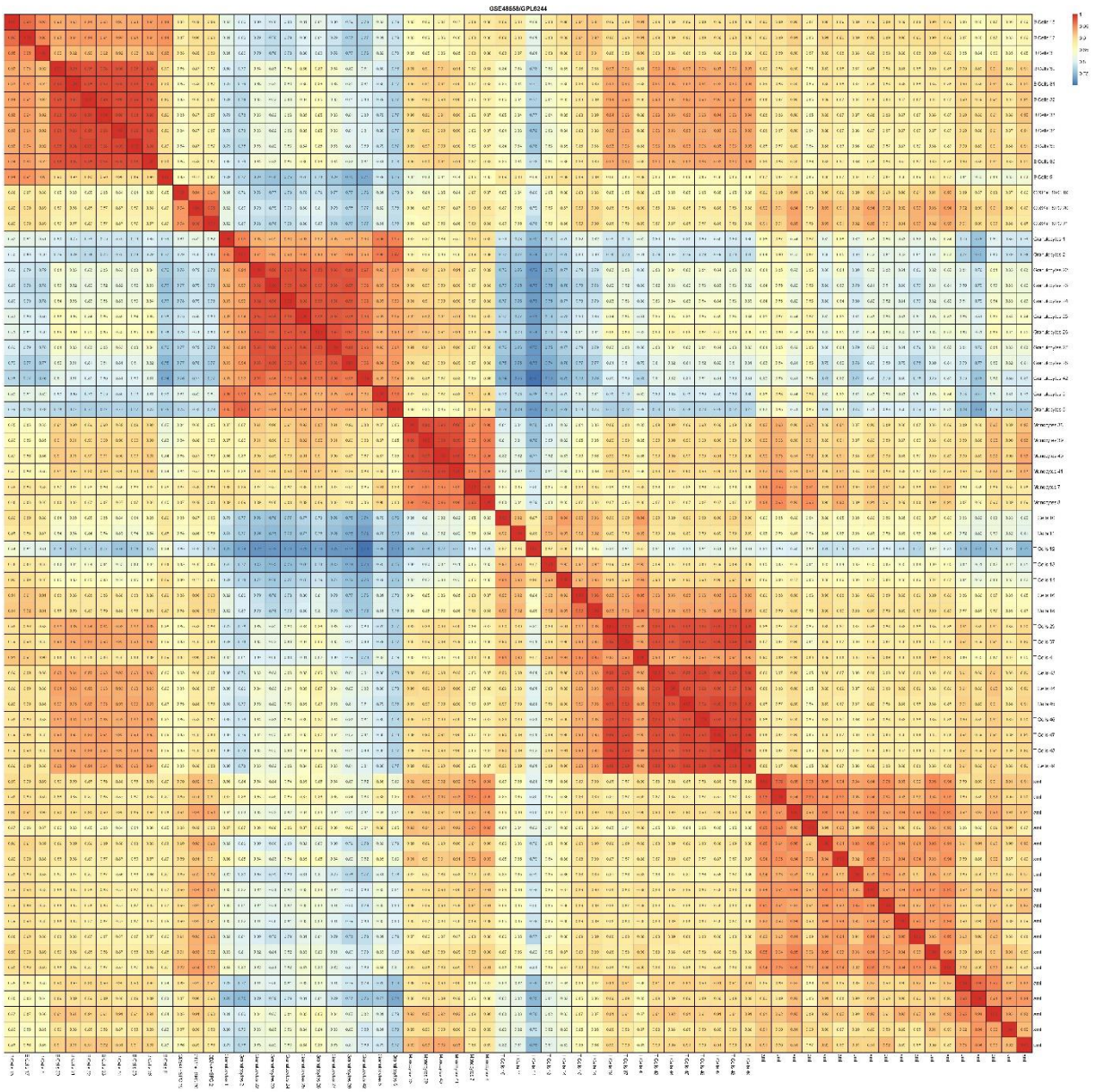
```
data.scaled = t(scale(t(data.normalized)))
```

```
data.umap = umap( data.scaled)
data.umap_col = data.umap[["data"]]
v1 = data.umap_col[,1]
v2 = data.umap_col[,2]

pdf("result/umap.pdf",width = 8,height = 6)
plot(v1,v2, main = paste(title,"/","umap",sep = ""), col = colors, pch = 19)
legend("topright", legend=c("healthy", "AML"),
       col=c("green", "red"), pch= 19, cex=0.8)
dev.off()

data.pca = prcomp(data.scaled,scale. = F)

pcr = data.frame(data.pca$rotation[,1:3], group = c(rep("normal",ncol(ph.normal)),rep("aml",ncol(source_name.aml))))

pdf("result/geom_point_pca12.pdf",width = 6,height = 4)
ggplot(pcr,aes(PC1,PC2,color = group)) + geom_point(size = 1) + theme_bw()
dev.off()

pdf("result/geom_point_pca13.pdf",width = 6,height = 4)
ggplot(pcr,aes(PC1,PC3,color = group)) + geom_point(size = 1) + theme_bw()
dev.off()

pdf("result/geom_point_pca23.pdf",width = 6,height = 4)
ggplot(pcr,aes(PC2,PC3,color = group)) + geom_point(size = 1) + theme_bw()
dev.off()

data.distance = dist(t(data.scaled))
data.mds_classical  = cmdscale(data.distance)
v1 = data.mds_classical[,1]
v2 = data.mds_classical[,2]

pdf("result/mds_classical.pdf",width = 8,height = 6)
plot(v1,v2, main = paste (title, "/", "mds", sep =""), col = colors, pch = 19)
legend("topleft", legend=c("healthy", "AML"),
       col=c("green", "red"),pch = 19 )
dev.off()


pdf("result/tSNE_Samples.pdf",width = 8,height = 6)
tsne(data,labels=group)
dev.off()
```

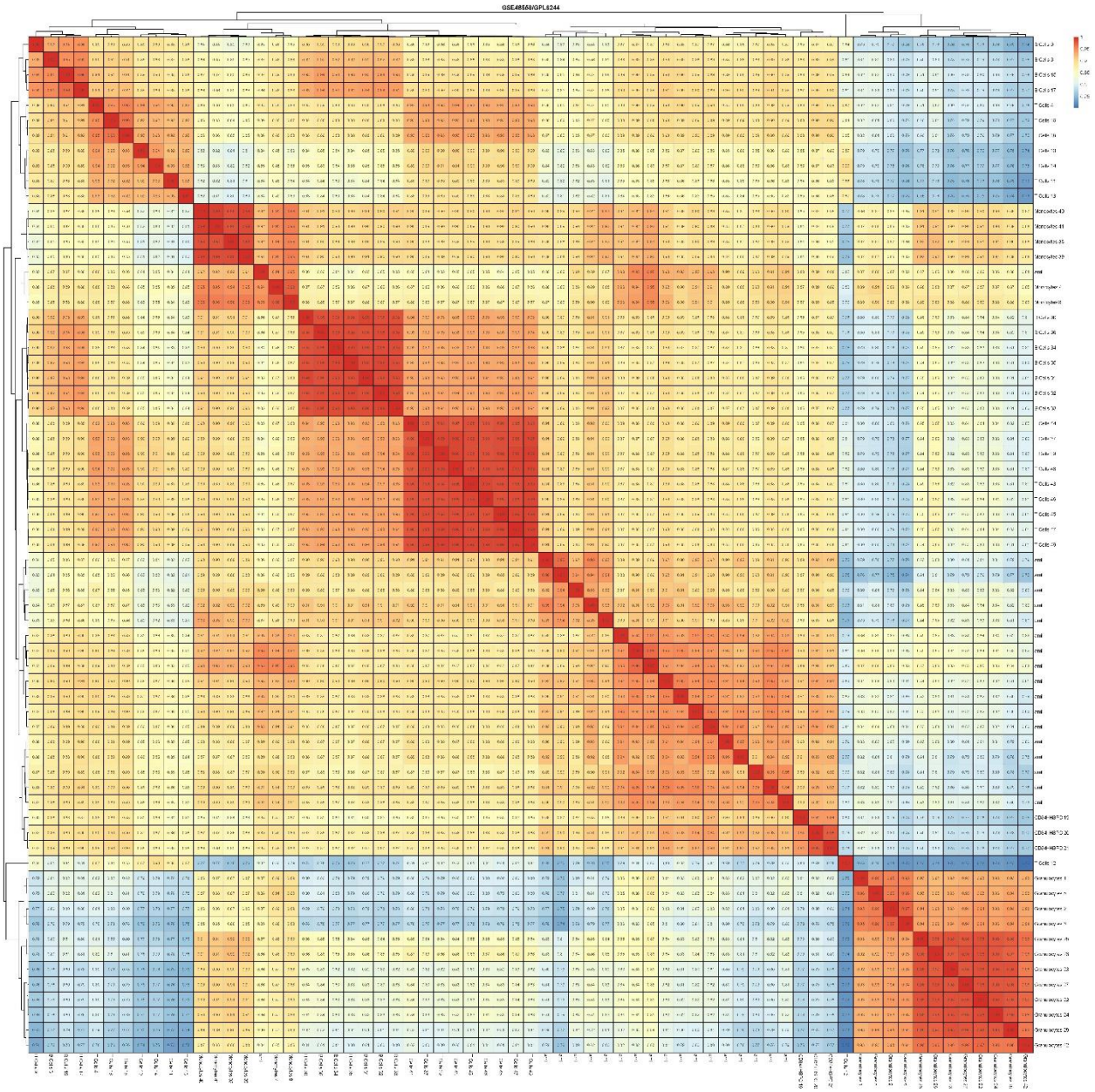
4. اگر دقت کنید source name نمونه های نرمال مختلف با یکدیگر تفاوت دارند. این فیلد بیانگر چیست؟

اگر داده ها را بر اساس source name گروه بندی کنیم (همه ی نمونه های بیمار در یک گروه هستند و هر نمونه ی سالم در گروه source name متناظر با خود است) همبستگی بین گروه ها با هم را بررسی کنید و به صورت یک نمودار نمایش دهید .گروهی از داده های سالم که بیشترین همبستگی با نمونه های گروه بیمار دارند را از روی نمودار مشخص کنید تا در مراحل بعد از این گروه برای تحلیل های بعدی استفاده شود. به نظر شما لزوم انجام این مرحله چیست؟

source name بیان گر نوع سلول ایمینی بدن است که مورد آزمایش واقع شده است می باشد که و این نام گذاری می تواند با توجه به فرد بیمار و سالم متفاوت باشد.



without clustering



With clustering

اگر به داده ها نگاه کنیم متوجه می شویم که CD34+HSPC بیش ترین همبستگی را با سلول ها فرد مبتلا به aml دارد.

```
newColName = function(names,matrix){
  paste(names,seq(1,ncol(matrix)),"")
}

healthy.new_name = exprs(ph.raw)
colnames(healthy.new_name) = newColName(ph.raw$source_name_ch1,healthy.new_name)
healthy.new_name = healthy.new_name[, sort(colnames(healthy.new_name))]


aml.new_name = exprs(source_name.raw)
colnames(aml.new_name) = rep("aml",ncol(aml.new_name))

data.cor = cbind(healthy.new_name,aml.new_name)
data.cor = cor(data.cor)


pdf("result/corheatmap_data.pdf",width = 32,height = 32)
pheatmap(data.cor,main=title, display_numbers = TRUE,border_color = "black")
dev.off()

pdf("result/corheatmap_data_no_cluster.pdf",width = 32,height = 32)
pheatmap(data.cor,main=title, display_numbers = TRUE,border_color = "black",cluster_rows = F,cluster_cols = F)
dev.off()
```