مقدمه ای بر بیوانفورماتیک

دانشگاه صنعتی شریف

پاییز 1401

اساتید: دکتر کوهی – دکتر شریفی

اعضای گروه:

ارسلان مسعودی‌فرد-99105718

امیرحسین محمودی-98108779

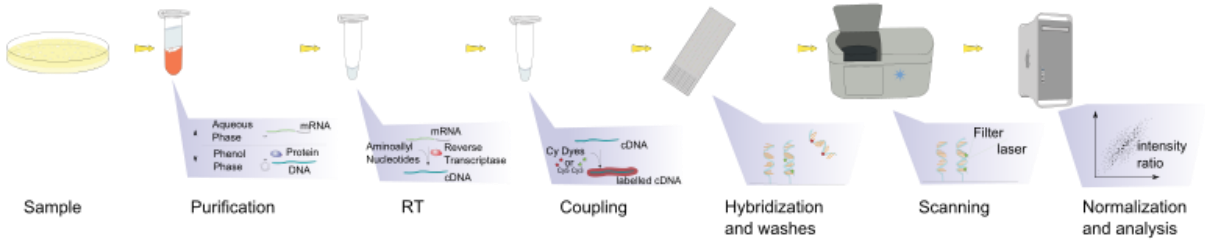
آرش ملکپور-98108821

بخش اول

ریزآرایه یک ابزار آزمایشگاهی است که برای تشخیص حالات هزاران ژن به صورت همزمان به کار می‌رود.

از آنها در تفسیر داده‌های تولید شده روی آزمایش روی DNA، RNA و پروتئین‌ها استفاده می‌شود.

تحلیل داده‌ی ریزآرایه آخرین مرحله در خواندن و پردازش داده‌های بدست آمده از چیپ ریزآرایه است، مانند شکل زیر یک نمونه تحت انواع پروسه‌هایی شامل پاکسازی(purification)، hybridization و اسکن توسط میکروچیپ قرار می‌گیرد که حاصل، داده‌های پرشماری است که تحلیل آن باید به کمک کامپیوتر صورت گیرد.



بسته به اینکه مراحل زیر چگونه و به چه ترتیبی انجام شوند خروجی متفاوتی خواهیم داشت و شرکت‌های مختلف در کنار محصول ریزآرایه خود نرم‌افزارهای تحلیلی نیز ارائه می‌دهند که تکنیک های زیر در آنها به کار می‌رود.

1.Aggregation and normalization

2.Identification of significant differential expression

3.Clustering

4.Pattern recognition

دیگر تکنیک آماری در تحلیل ریزآرایه‌ها SAM(significance analysis of microarrays) است که برای تشخیص تغییرات در حالات ژن به کار می‌رود به طوری که حالت هزاران ژن را در یک hybridization میتوان اندازه گرفت. پروتکل پایه انجام این روش به شرح زیر است.

1- انجام آزمایش‌های مربوط به ریزآرایه
2- دادن ورودی حالات به Microsoft excel
3- اجرای SAM به عنوان افزونه‌ی excel
4- تنظیم پارامتر دلتا برای بدست آوردن مشخصه‌ی# ژن ها در کنار FDR مورد قبول وارزیابی اندازه‌ی نمونه با محاصبه میانگین فاصله در حالات.
5- لیست کردن ژن‌های بیان شده.

بعد از اجرای SAM فرمت خروجی ها به حالات زیر خواهد بود.

- **Quantitative** — مقدار حقیقی
- **One class** — بررسی اینکه میانگین با صفر چقدر فاصله دارد
- **Two class** — دو دسته اندازه گیری
    - **Unpaired** — واحدهای اندازه‌گیری دو گروه متفاوتند
    - **Paired** — واحدها یکسانند
- **Multiclass** — حالت تعمیم یافته زوج نشده
- **Survival** — داده‌ی زمان تا یک رخداد یک اتفاق
- **Time course** — هر کدام از واحدهای آزمایشی در بیش از یک زمان اندازه‌گیری شده
- **Pattern discovery** — بدون پارامتر خروجی مشخص

بخش دوم

برای دریافت فایل های داده شده از وبسایت NCBI کتابخانه GEOquery و پکیج BiocManager را نصب می‌کنیم.

در دیتاست داده شده یک پلتفرم GPL6244 قرار دارد که نوع ریزآرایه را مشخص می‌کند، همچنین 170 تا GSM یا نمونه وجود دارد، پس فایل GSE که لیست تمامی آنها هست را دانلود می‌کنیم.

```
7  dataset <- getGEO( "GSE48558", GSEMatrix = TRUE,AnnotGPL = TRUE , destdir = ".")
8  dataset <- dataset[[1]]
9  print(dataset)
```

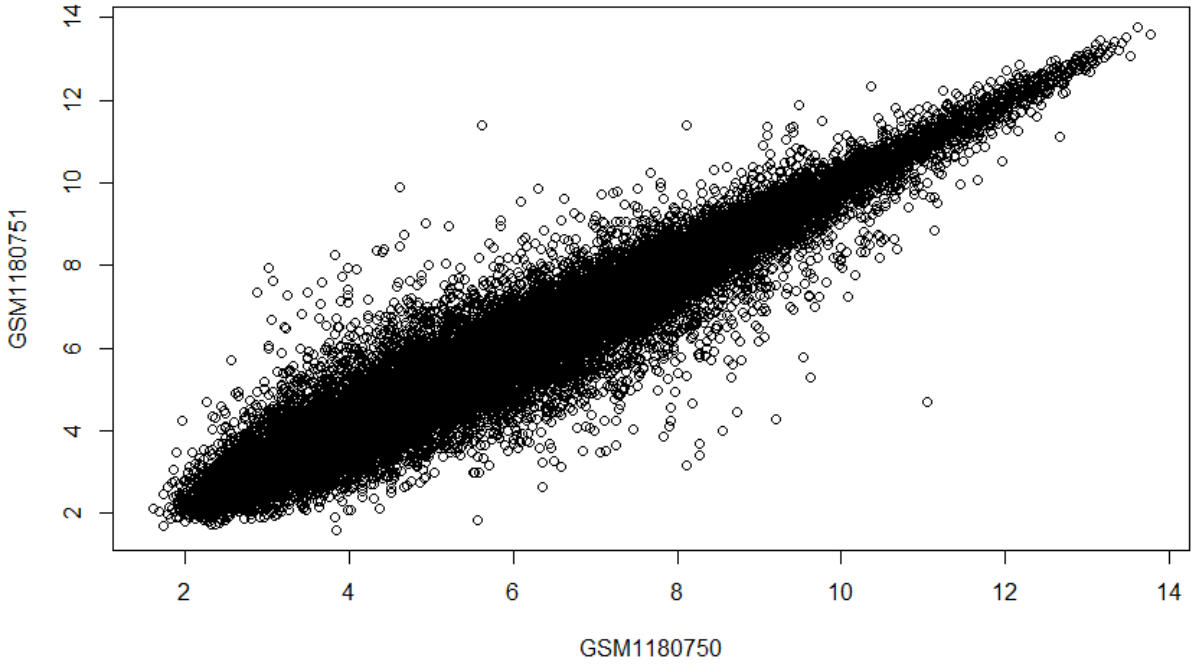
در ادامه داده هایی که source_name آنها AML Patient و یا phenotype نرمال دارند را نگه داشـــته و برای شـــهود بیشتر به بررسی تعداد آنها می‌پردازیم که 73% داده ها نرمالند.

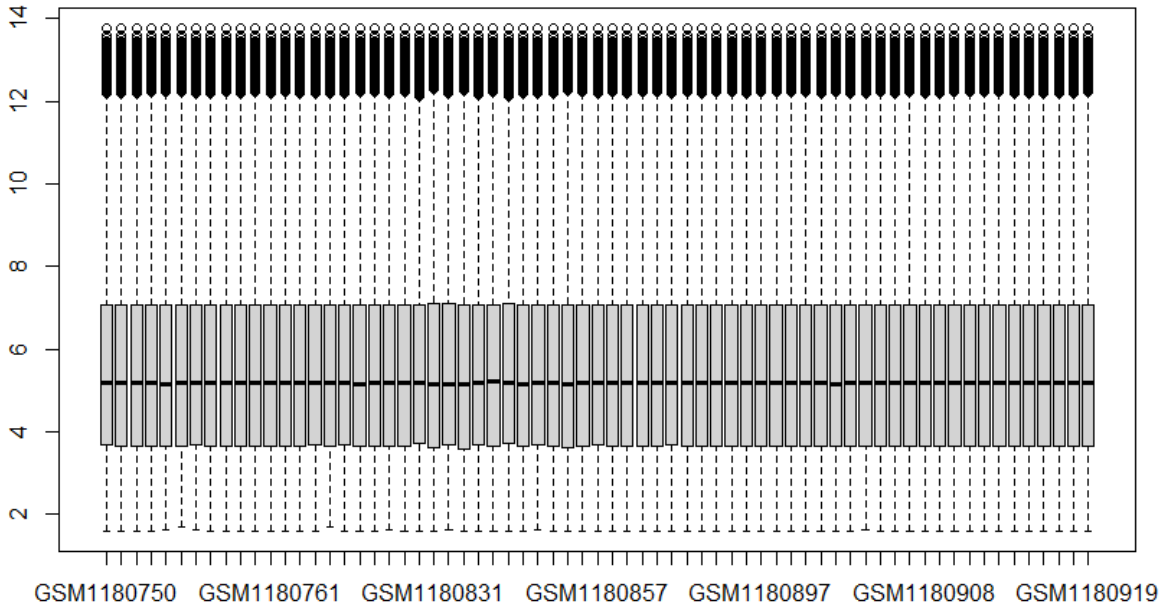
```
dataset<- dataset[,which(dataset$source_name_ch1 == "AML Patient"
                         | dataset$`phenotype:ch1` == "Normal")]
grouped = list()
for(i in 1:length(dataset$`phenotype:ch1`) ) {
  if (dataset$source_name_ch1[i] != "AML Patient") {
    grouped[[length(grouped) + 1]] <- "Normal"
  } else {
    grouped[[length(grouped) + 1]] <- "Test"
  }
}

a <- 0
for(i in 1:length(dataset$`phenotype:ch1`)){
  if (grouped[[i]] == "Normal"){
    a <- a + 1
  }
}
print(a / length(dataset$`phenotype:ch1`))
```

برای بدست آوردن بازه بیان ژن‌ها از قطعه کد زیر استفاده کرده و از آنجایی که کمتر از 14 هستند پس مقیاس لگاریتمیست.

```
print(min(exprs(dataset)))
print(max(exprs(dataset)))
```

نمودار Scatter آن را نیز رسم کرده تا دید بهتری نسبت به احتیاج آن به نرمالایز کردن پیدا کنیم. همانطور که مشخص است اکثر داده ها در حل یک خط قرار دارند و نمودار boxplot هم به همین موضوع اشاره می‌کند. چرا که چارک‌ها در موازات هم قرار گرفته‌اند.





بخش سوم

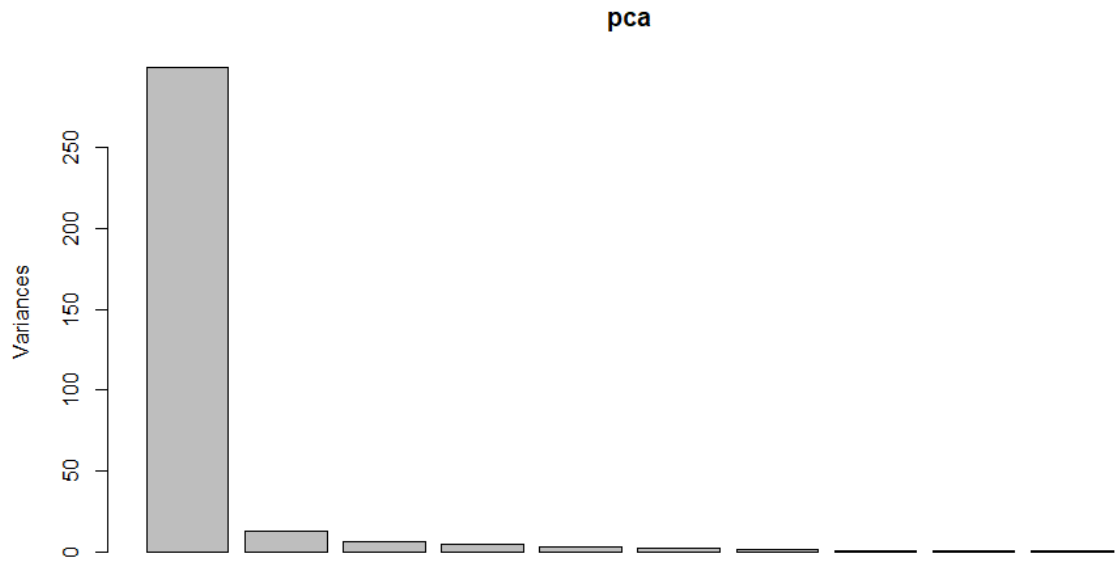
با کاهش ابعاد داده‌ها به دنبال متغیر هایی با کمترین اهمیت هستیم تا پیچیدگی داده‌ها را کاهش دهیم، همچنین ممکن است مقدار noise نیز کاهش یابد و در نهایت از overfit شدن داده ها جلوگیری می‌کند و تحلیل ساده‌تر آن‌ها را توسط نمودار scatter میسر می‌سازد.

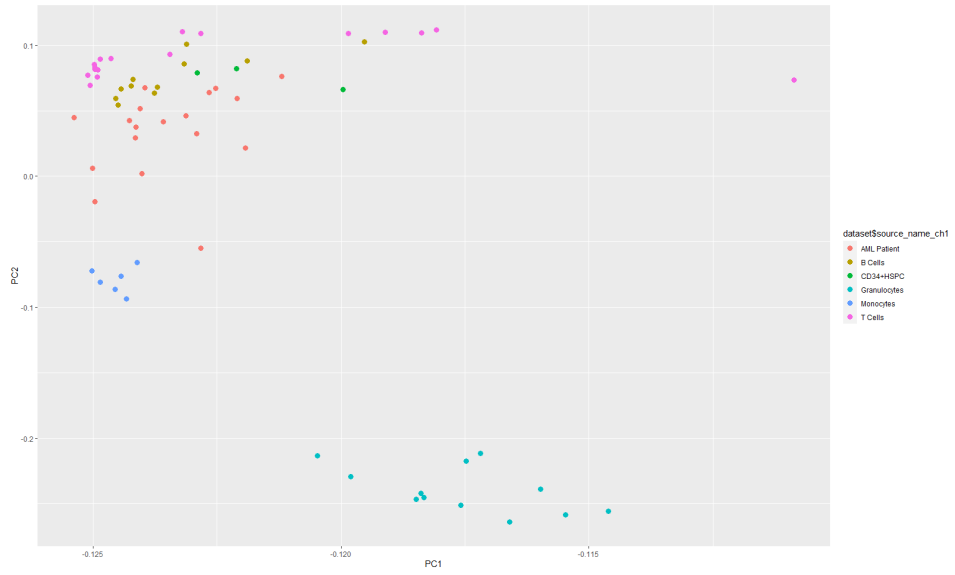
بدین منظور از سه روش PCA ، T-SNE و LLE استفاده می‌کنیم.

:PCA

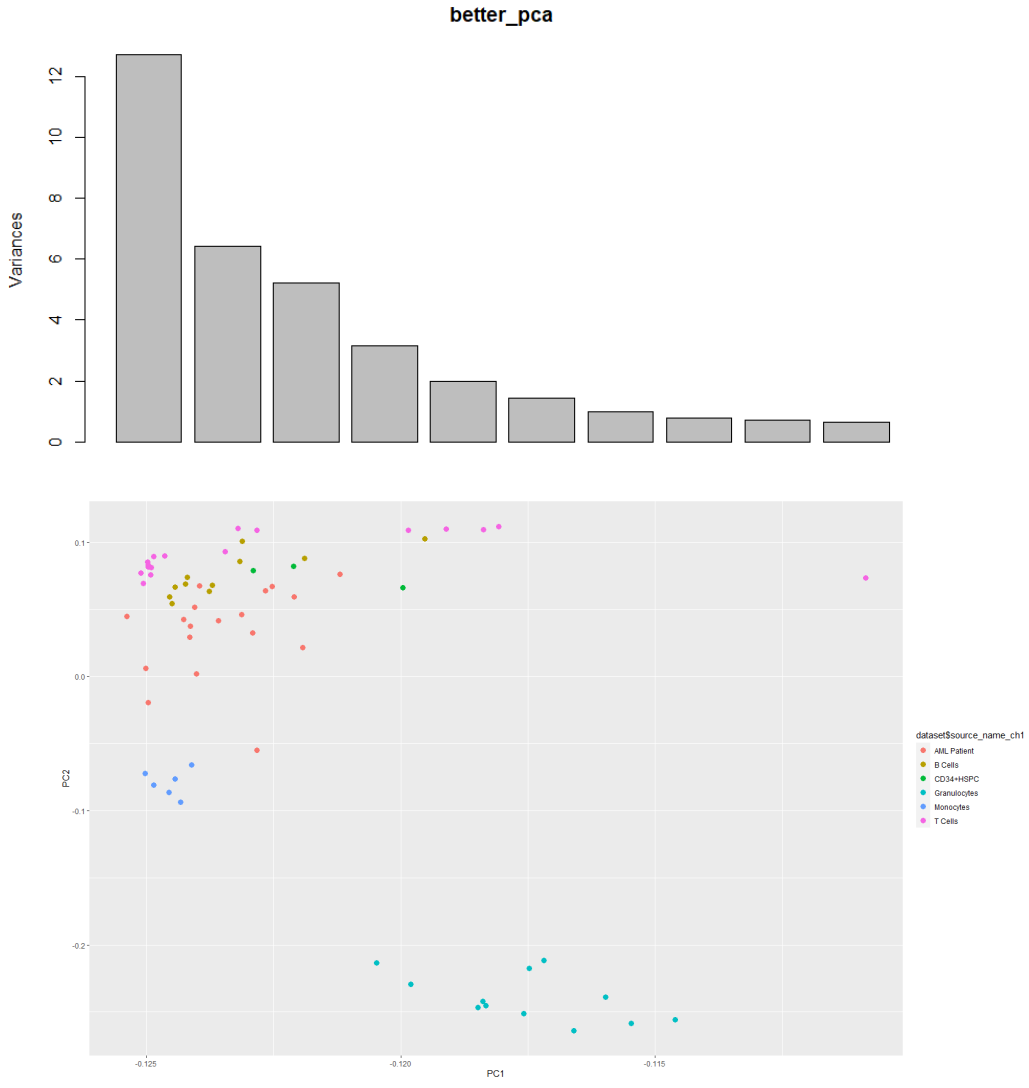
این روش جبرخطی از معمول‌ترین شیوه‌های کاهش ابعاد است. در بخش اول کد زیر از آنجایی که یکسری ژن‌ها بیان خیلی بالاتری نسبت به بقیه دارند نمودار scatter خوشه بندی مناسبی به ما نمی‌دهد و در قسمت دوم کد این مشکل را با scale کردن برطرف می‌کنیم.

```
69  pca <- prcomp(exprs(dataset))
70  plot(pca)
71  pcr <- data.frame(pca$r[,1:3] , Group = grouped)
72  ggplot(pcr , aes(PC1 , PC2 , size = 4, color=dataset$source_name_ch1)) + geom_point(size=3)
73
74
75
76
77  better_pca <- prcomp(t(scale(t(exprs(dataset)) , scale = FALSE)))
78  plot(better_pca)
79  better_pcr <- data.frame(pca$r[,1:3] , Group = grouped)
80  ggplot(better_pcr , aes(PC1 , PC2 , size = 4, color=dataset$source_name_ch1)) + geom_point(size=3)
```





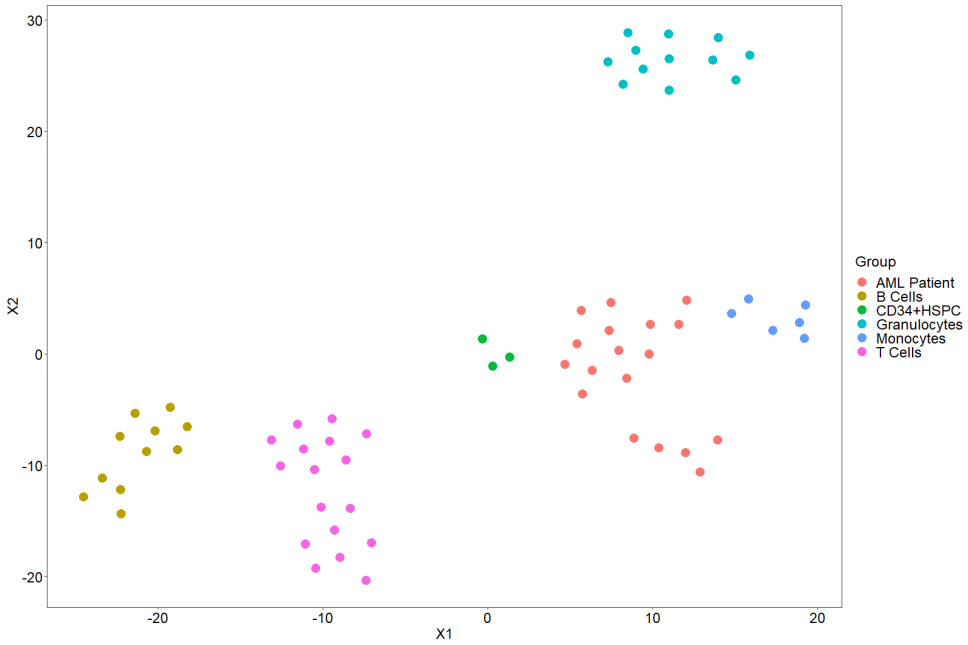
نمودار اصلاح شده با scale:



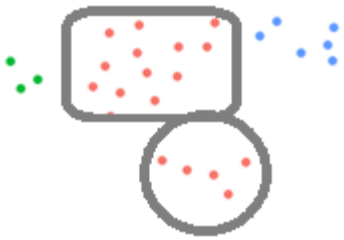
:T-SNE

این روش مدل توزیع احتمالی همسایه‌های هر نقطه را ارائه می‌دهد که توسط کتابخانه زیر آن را بر روی دیتاست خود پیاده می‌کنیم.

```
81  library(M3C)
82  tnse_model <- tsne(dataset,labels=as.factor(dataset$source_name_ch1))
83  plot(tnse_model)
```



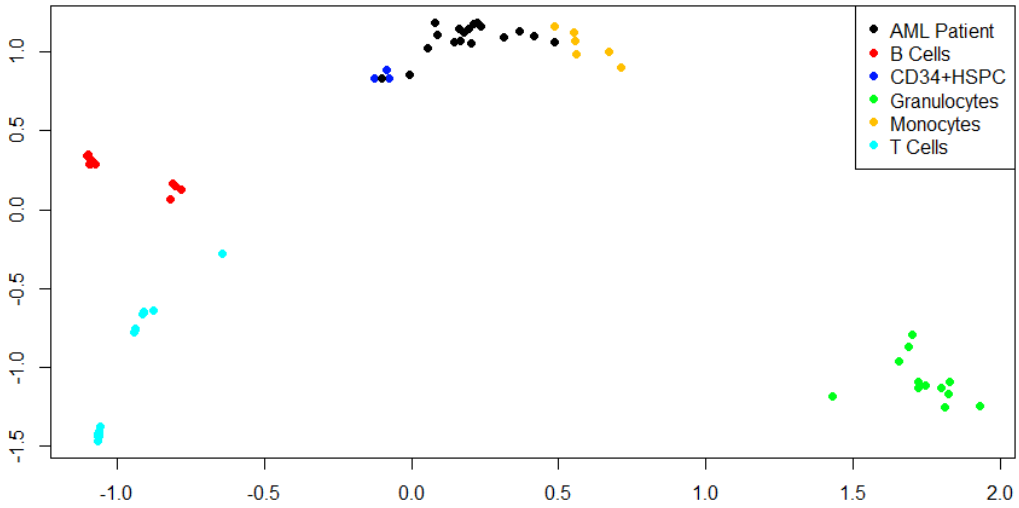
همانطور که از نمودار مشخص است خوشه(cluster) شدن داده‌ها به مراتب بهتر انجام شده، همچنین در خوشه‌ی تست یا بیمار هم دو دسته‌ی جدا از هم خیلی نزدیک‌تر نسبت به PCA قرار گرفته‌اند و فاصله آن‌ها با دو دسته CD34 و monocytes هم نشان‌گر همبستگی است که در بخش بعد به آن می‌پردازیم.



:LLE

این روش که مخفف Locally Linear Embedding است در واقع در کل غیرخطی بوده و با اینکه در نمونه زیر کلاسترینگ خیلی خوبی به ما ارائه می‌دهد اما باید توجه کرد که مقدار k یا همسایه‌های درنظر گرفته شده نسبتا زیاد است و درصورت داشتن تعداد داده بالاتر زمان اجرای خیلی زیادی خواهیم داشت.

```r
library(RDRToolbox)
lle_model = LLE(t(exprs(dataset)), dim=2, k=18)
plotDR(data=lle_model, labels=as.factor(dataset$source_name_ch1), axesLabels=c("", ""), legend=TRUE)
```

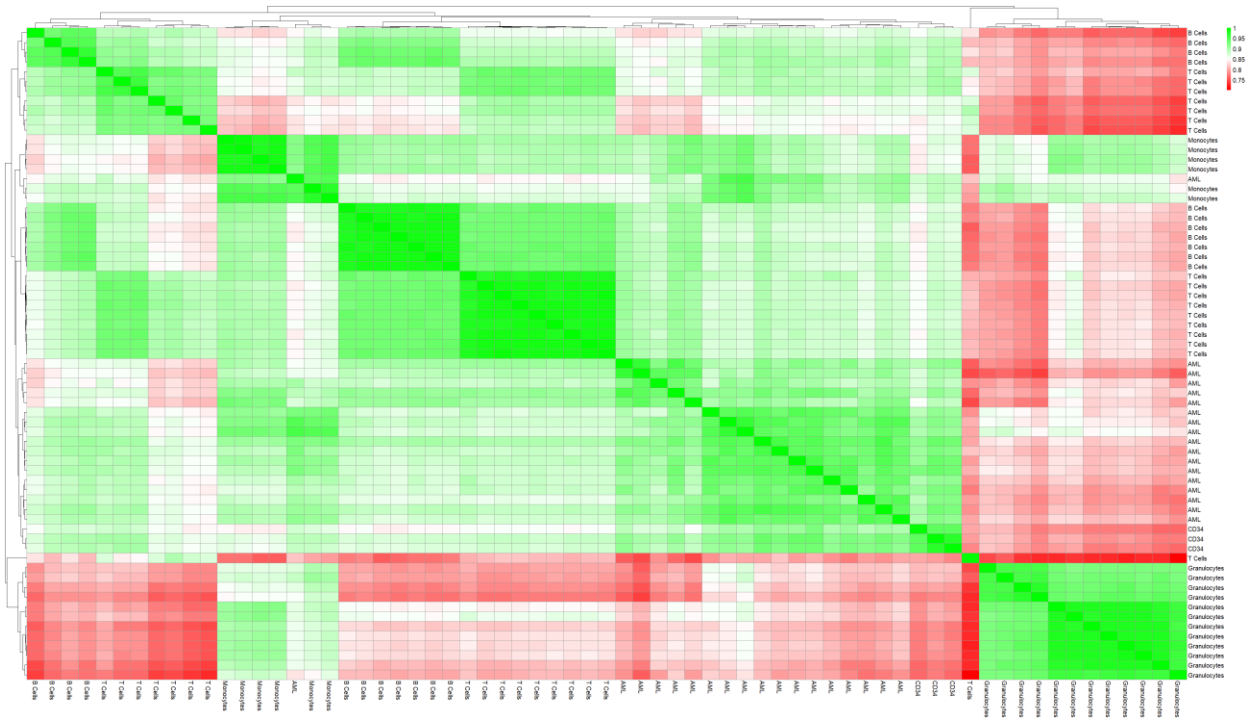


بخش چهارم

فیلد source_name در واقع بیانگر سلول نمونه است. به عنوان مثال T_cell نوعی سلول گلبول سفید، Granulocyte نوع دیگری از سلول های ایمنی (گلبول سفید) و عنوان AML هم که برای بیماران دارای سلول سرطانی است. پس عناوینی که AML ندارند مربوط به دسته نرمال و بقیه برای بیماران هستند و طبق کد زیر این دسته ها را از هم جدا کرده و نمودار heat map آن‌ها را رسم کرده تا مشخص شود کدام دسته از سلول ها همبستگی بیشتری نسبت به یکدیگر دارند.

```r
sourcenames = list()
for(i in 1:length(dataset$`phenotype:ch1`) ) {
  if (dataset$source_name_ch1[i] != "AML Patient") {
    sourcenames[[length(sourcenames) + 1]] <- strsplit2(dataset$source_name_ch1[i], "\\+")[1, 1]
  } else {
    sourcenames[[length(sourcenames) + 1]] <- "AML"
  }
}

library(pheatmap)

pheatmap(cor(exprs(dataset)), fontsize = 15
         ,color=colorRampPalette(c("red", "white", "green"))(50),
         labels_row =  sourcenames
         , labels_col =  sourcenames)
```

کد همبستگی



نمودار همبستگی

همانطور که از نمودار مشخص است دسته‌های سلولی یکسان همبستگی بیشتری دارند و گروه AML که دسته‌ی بیماران است به ترتیب با دو دسته‌ی CD34 و Monocytes بیشترین همبستگی را دارد، پی می‌توان این نتیجه را گرفت که پروتئین های CD34 هستند که برای بررسی AML حائز اهمیتند و می‌توان بقیه را نادیده گرفت. پس اگر به دنبال درمانی برای این بیماری هستیم بررسی‌های آینده باید به دسته‌ی CD34 معطوف شود.