notebook

December 9, 2022

Saman Dehestani	401208226
Sajede Fadaei	401208226
Aminreza Sefid	401208226

1 Loading libraries

[ ]:

```
library(GEOquery)
library(limma)
library(umap)
library(pheatmap)
library(gplots)
library(ggplot2)
library(reshape2)
library(plyr)
library(repr)
library(gridExtra)
library(ggpubr)
library(Rtsne)
library(MASS)
```

2 Loading the $GSE48558$ series

The `getGEO()` method from the `GEOquery` package is used in order to download the `GEO SOFT` format of the $GSE48558$ and then parse it to the `R` structure obtaining the `GSE Matrix` and the `GPL annotations`:

[17]:

```
gset <- getGEO("GSE48558", GSEMatrix =TRUE, getGPL=T, destdir='../Data/')
gset <- gset[[1]]
```

```
Found 1 file(s)

GSE48558_series_matrix.txt.gz

Using locally cached version: ../Data//GSE48558_series_matrix.txt.gz
Using locally cached version of GPL6244 found here:
../Data//GPL6244.soft.gz
```

[12]:
```
dim(gset)
```

Features 32321 **Samples** 170

3 Selecting proper samples

There are different sample types (mentioned as `sourc_ename`), including:

- AML cell line
- AML patient
- B ALL cell line
- B ALL patient
- B normal
- T ALL cell line
- T patient
- T normal
- granulocytes normal
- monocytes normal
- CD34+ normal

We will only select the `Leukemia` AML samples and those with the `normal` phenotype value. The selection and the grouping were done first in the `GEO2R` analysis tool provided by https://www.ncbi.nlm.nih.gov/. A string will be generated, which will be used to select and group the samples:

[18]:
```
gsms <- paste0("0000000000000XXXXXXXXXXXXXXXXXXXXXXXXXXX1XXX1XXXXX",
               "XXXXXXXXXXXXXXXXXX2X3XXX1X1442X3XX33XX33X2X3X2X3X5",
               "XXX5XXX5XXXXXXXXXXXXXXXXXXXXXXXXXXXXX1111111003000",
               "22222223444413333333")
sml <- strsplit(gsms, split="")[[1]]
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]
```

Samples are grouped according to the following rules:

- All the `AML` samples are grouped together.
- `Normal` samples are grouped based on their `source_name` (B, T, granulocytes, monocytes, and CD34+).

Each number in the provided string stands for a group, and the `X` character means the corresponding sample won't be selected.
Next we will create a `factor` from these class numbers and add the `group` column to the `gset`.

[20]:

```
gs <- factor(sml)
groups <- make.names(c("AML","Granulocytes","B Cells","T␣
  ↪Cells","Monocytes","CD34"))
levels(gs) <- groups
gset$group <- gs
```

4 Pre-Processing the data

write some bullshit about the essence of the preprocessing. a brief about what are we going to do for the preprocessing.

4.1 Range of the *expression* values

[22]:

```
ex <- exprs(gset)
print(min(ex))
print(max(ex))
```

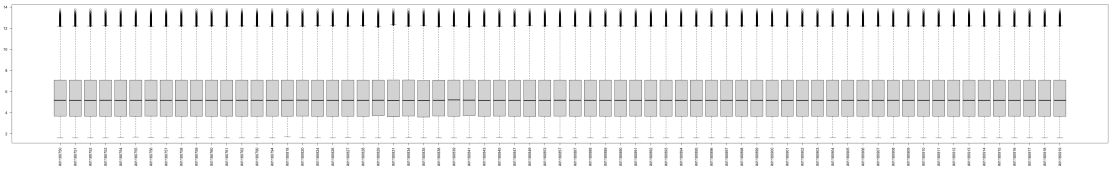
```
[1] 1.611473
[1] 13.76154
```

According to this, min value is 1.6 and max value is 13.76, so they are logarithmic already and data is normalized.

4.2 Plot the *expression* matrix

[6]:

```
options(repr.plot.width=50, repr.plot.height=8)
boxplot(ex, las = 2)
```



Question 4)

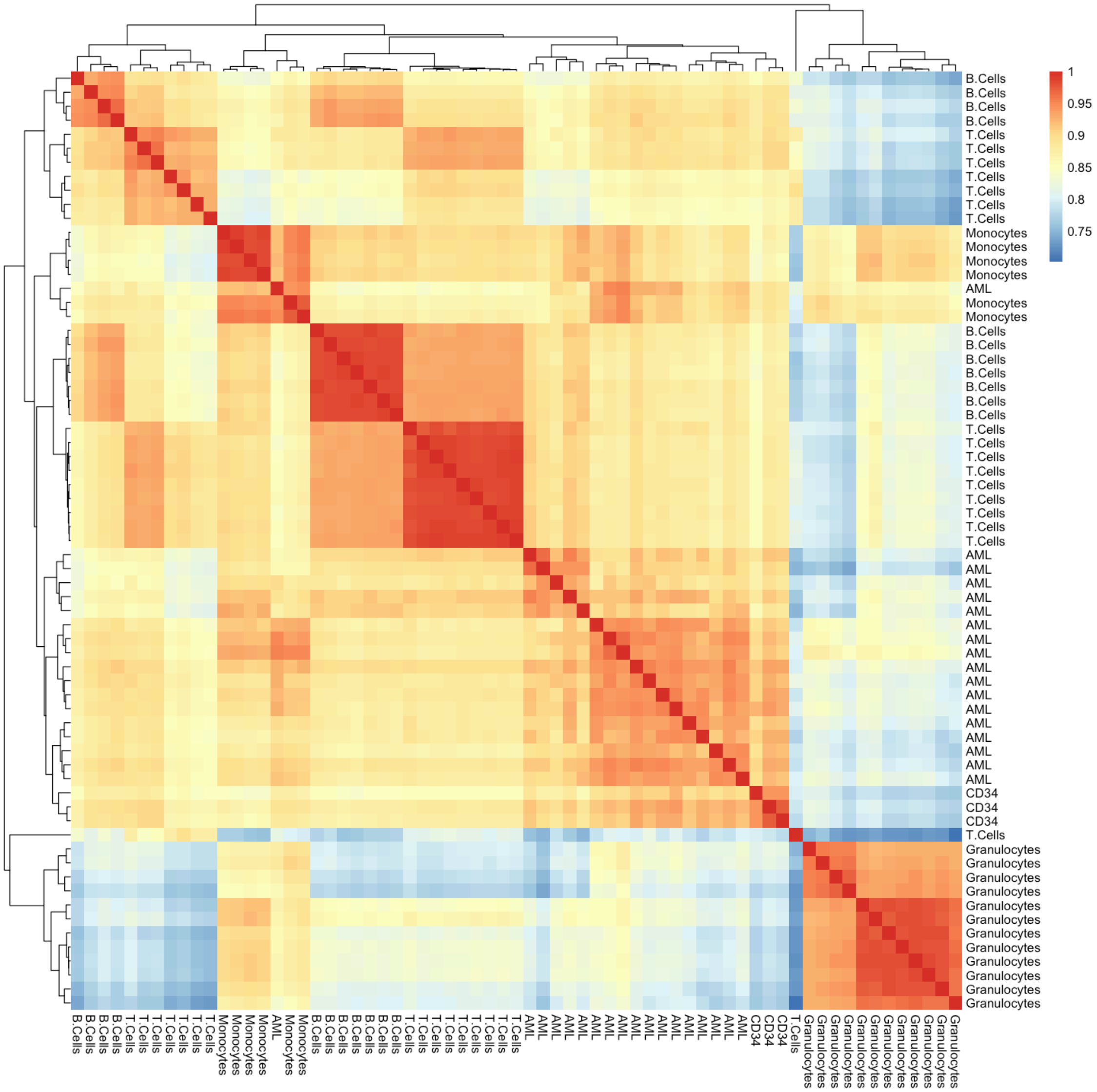
Heat map shows correlations between different samples, for example each sample has high corealation with itself that is determined with red color or granulocytes have low corelation between B-cells and T-celss that is determined with blue

According to the heatmap AML has high corolation with CD34,Monocytes

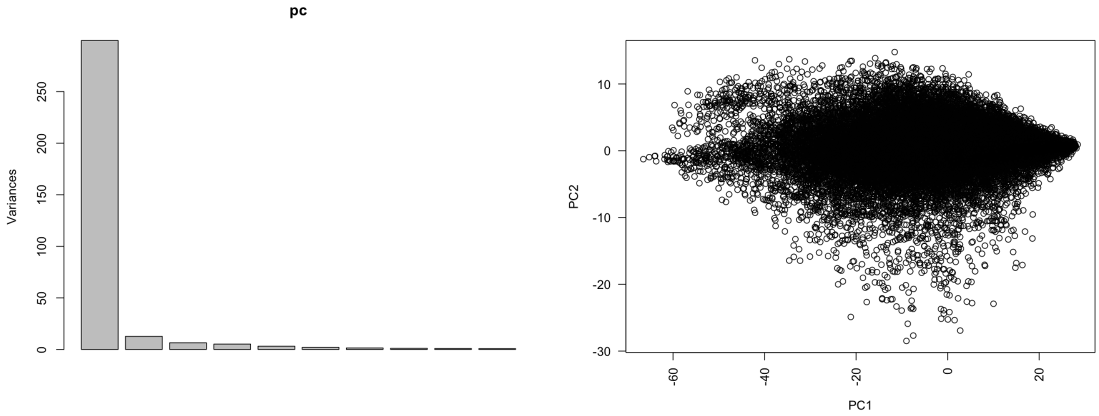
what is neccesecity?

[7]:

```
options(repr.plot.width=12, repr.plot.height=12)
pheatmap(cor(ex),
         labels_row = gs,
         labels_col = gs,
         border_color = NA,)
```



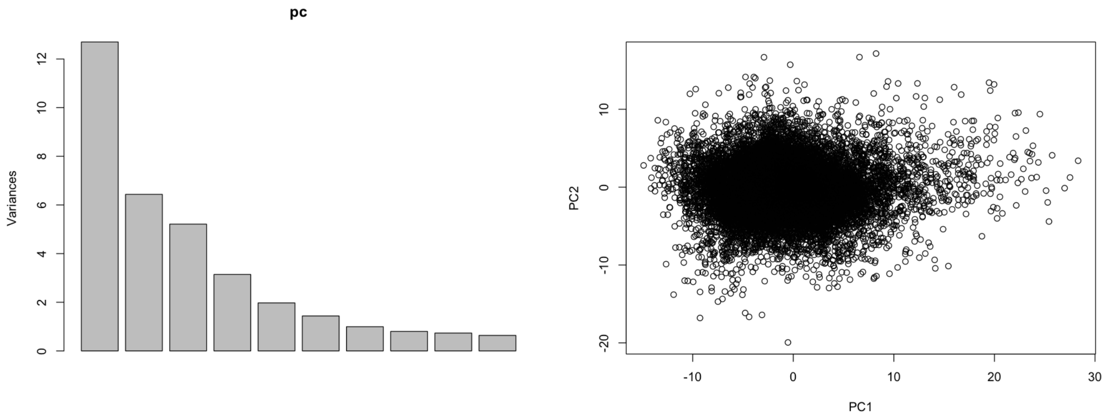
[8]:
```
pc <- prcomp(ex)
options(repr.plot.width=15, repr.plot.height=6)
par(mfrow = c(1, 2))
p1 <- plot(pc)
p2 <- plot(pc$x[, 1:2], las = 2)
```



[9]:

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

options(repr.plot.width=15, repr.plot.height=6)
par(mfrow = c(1, 2))
plot(pc)
plot(pc$x[, 1:2]) # x column are genes (so each point in this plot represents a gene)
```



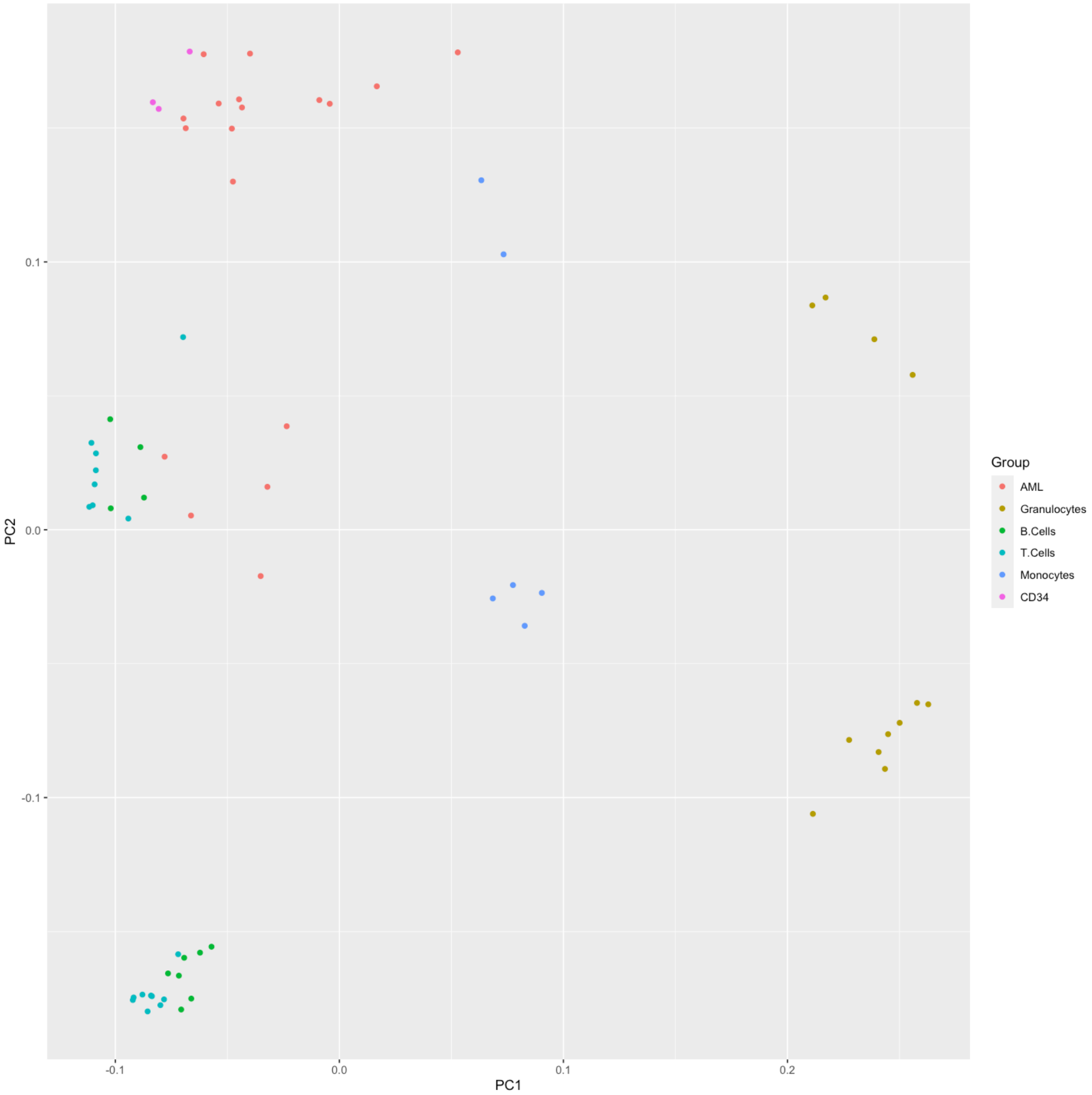
Question 3)

Dimension reduction can visualize data so we can determine if we performed well on experiment or not.

Best dimension reduction method is tSNE which has most discriminative clustering of samples.

[10]:

```r
pcr <- data.frame(pc$rotation[, 1:3], Group=gs)
options(repr.plot.width=12, repr.plot.height=12)
ggplot(pcr, aes(x = PC1, y = PC2, color = Group)) + geom_point() + theme_gray()
```

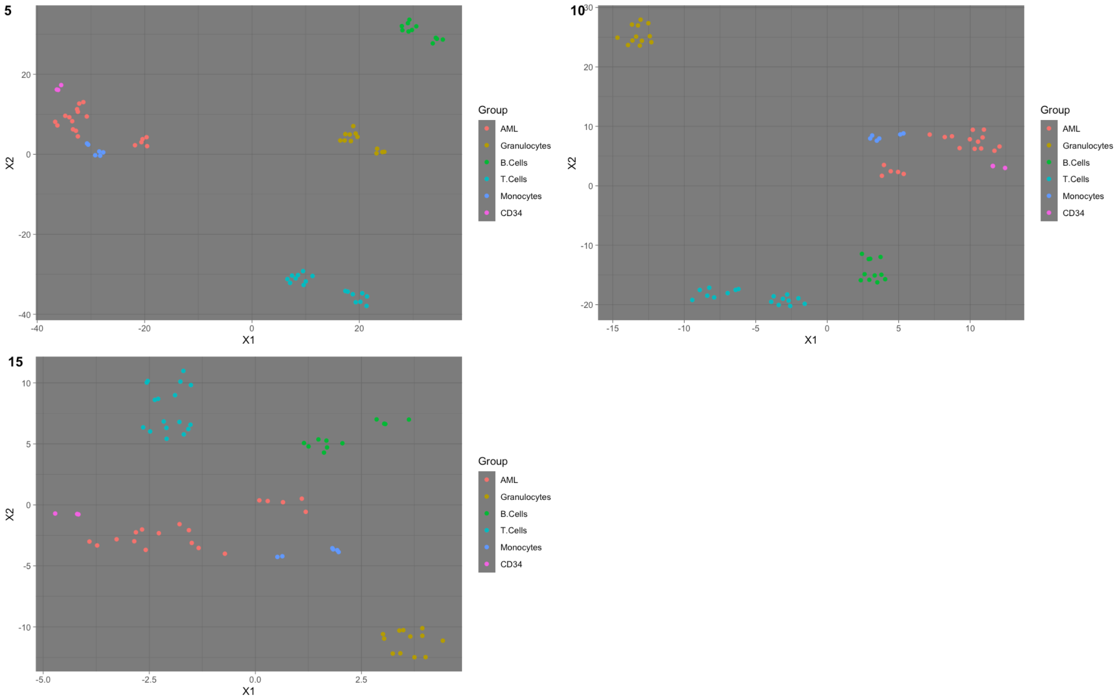


[11]:

```r
tsne_results <- list(Rtsne(t(ex), perplexity=5, check_duplicates = FALSE),
                     Rtsne(t(ex), perplexity=10, check_duplicates = FALSE),
                     Rtsne(t(ex), perplexity=15, check_duplicates = FALSE))

options(repr.plot.width=16, repr.plot.height=10)
plots.list <- list()

for(i in seq_along(tsne_results)) {
  tsne <- data.frame(tsne_results[[i]]$Y[, 1:2], Group=gs)
  plots.list[[i]] <- ggplot(tsne, aes(X1, X2, color = Group)) + geom_point() + theme_dark()
}
ggarrange(plotlist = plots.list,
          ncol = 2,
          nrow = 2,
          labels = c(5, 10, 15))
```


X2
X1
Group
AML
Granulocytes
B.Cells
T.Cells
Monocytes
CD34
X2
X1
Group
AML
Granulocytes
B.Cells
T.Cells
Monocytes
CD34
X2
X1
Group
AML
Granulocytes
B.Cells
T.Cells
Monocytes
CD34