Homework 4

Amy Nguyen

Clustering and dimension reduction for gene expression data

```
library(tidyverse)
library(ROCR)
library(ggridges)
library(dendextend)
```

```
leukemia_data <- read_csv("leukemia_data.csv")
```

(a)

```
leukemia_data <- leukemia_data %>%
  mutate(Type = as.factor(Type))
table(leukemia_data$Type)
```

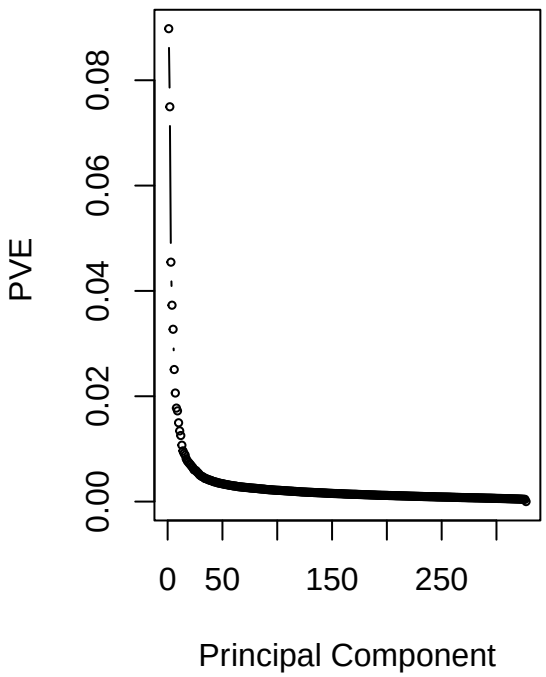
```
##
##    BCR-ABL   E2A-PBX1 Hyperdip50        MLL     OTHERS      T-ALL   TEL-AML1
##         15         27         64         20         79         43         79
```

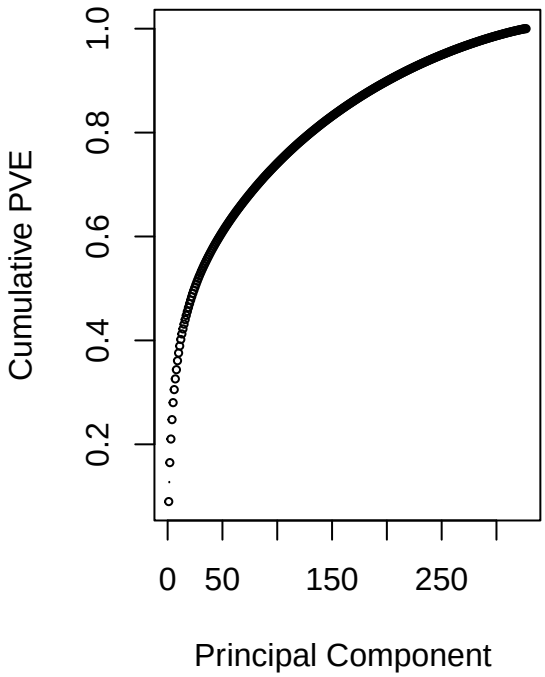
From the table, the leukemia subtype that occurs the least is BCR-ABL with 15 patients.

(b)

```
pr.out = prcomp(leukemia_data[,c(-1)], scale=TRUE, center=TRUE)
pr.var = pr.out$sdev^2
pve <- pr.var / sum(pr.var)

par(mfrow=c(1,2))
plot(pve, xlab="Principal Component",
     ylab="PVE ", type='b', cex=0.5)
plot(cumsum(pve), xlab="Principal Component ",
     ylab=" Cumulative PVE ",  type='b', cex=0.5)
```





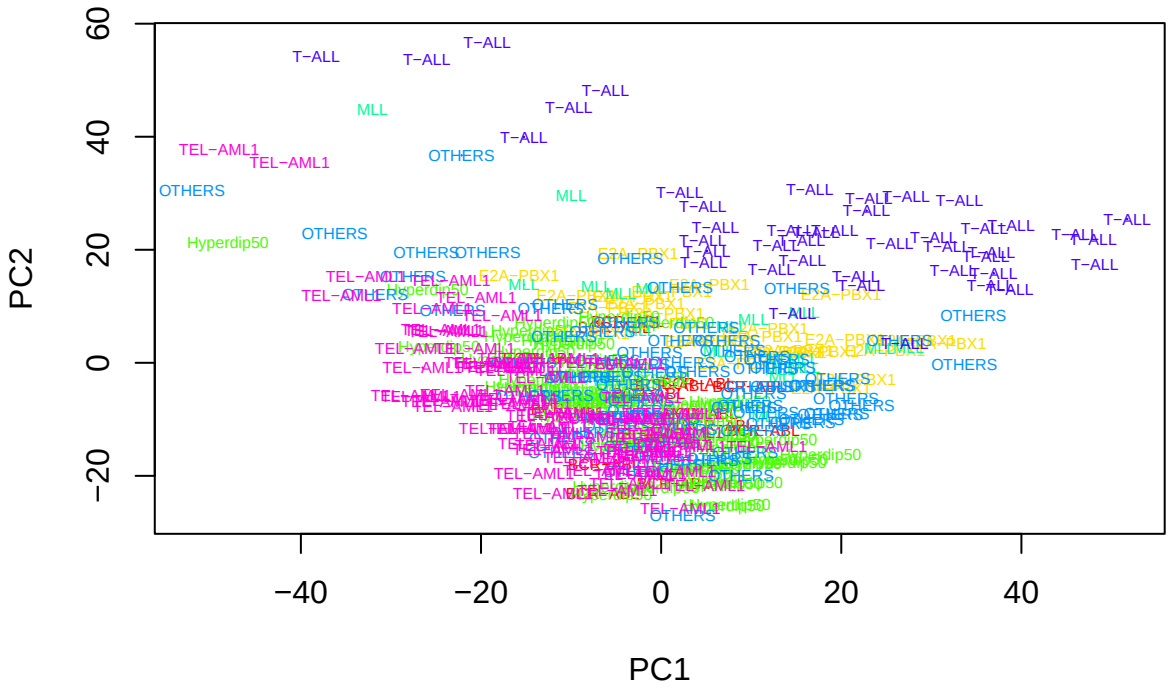

```
which(cumsum(pve) >= 0.9)[1]
```

```
## [1] 201
```

At least 201 PCs are needed in order to explain 90% of the total variation in the data.

(c)

```
rainbow_colors <- rainbow(7)
plot_colors <- rainbow_colors[leukemia_data$Type]
plot(pr.out$x[,c(1,2)], col = plot_colors, cex = 0.0001)
text(pr.out$x[,c(1,2)], col = plot_colors, labels = leukemia_data$Type, cex = 0.5)
```



The subtype T-ALL is most clearly separated from the other groups along the PC2 axis.

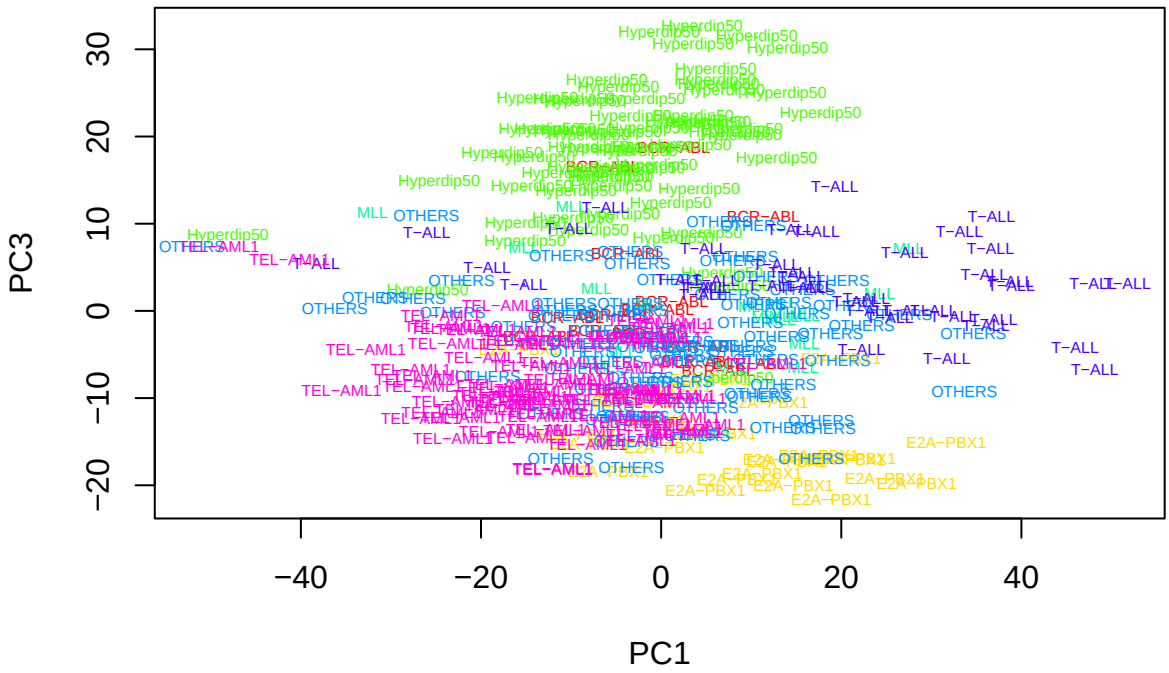
```
pr.out$rotation[, 1] %>%
  abs %>% sort(decreasing = TRUE) %>% head(6)
```

```
##     SEMA3F       CCT2       LDHB      COX6C     SNRPD2       ELK3
## 0.04517148 0.04323818 0.04231619 0.04183480 0.04179822 0.04155821
```

The top 6 genes with the highest absolute loadings for PC1 are SEMA3F, CCT2, LDHB, COX6C, SNRPD2, and ELK3.

(d)

```
plot(pr.out$x[,c(1,3)], col = plot_colors, cex = 0.0001)
text(pr.out$x[,c(1,3)], col = plot_colors, labels = leukemia_data$Type, cex = 0.5)
```



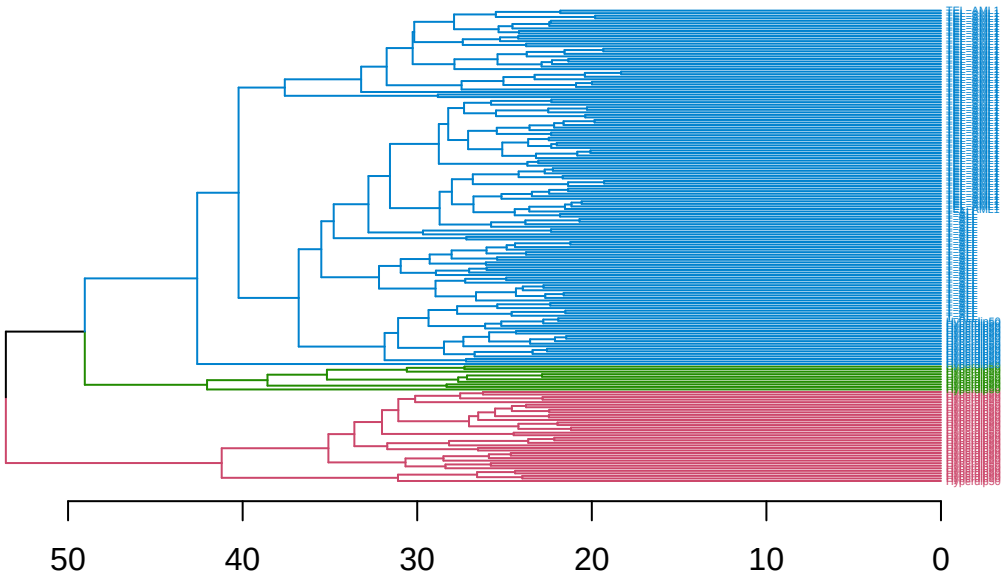
Based on the plot, PC3 does seem to do a slightly better job at disciminating between leukemia types than PC2. It is clear in this plot that leukemia of the same subtypes tend to have pretty similar gene expression levels.

(e)

```
leukemia_subset <- leukemia_data %>%
  filter(leukemia_data$Type == 'T-ALL' | leukemia_data$Type == 'TEL-AML1' |
           leukemia_data$Type == 'Hyperdip50')
set.seed(1)
hc = hclust(dist(leukemia_subset), method = 'complete')

dend1 <- as.dendrogram(hc)
dend1 %>%
  color_branches(k=3) %>%
  color_labels(k=3) %>%
  set("labels_cex", 0.3) %>%
  set_labels(labels=leukemia_subset$Type) %>%
  plot(horiz=TRUE, main="Dendrogram of Three Leukemia Subtypes")
```

Dendrogram of Three Leukemia Subtypes




```
dend2 <- as.dendrogram(hc)
dend2 %>%
  color_branches(k=5) %>%
  color_labels(k=5) %>%
  set("labels_cex", 0.3) %>%
  set_labels(labels=leukemia_subset$Type) %>%
  plot(horiz=TRUE, main="Dendrogram of Five Leukemia Subtypes")
```

Dendrogram of Five Leukemia Subtypes