Homework Assignment 4

Parker Reedy

November 28, 2023

```
## Warning: package 'dendextend' was built under R version 4.3.2
```

1.

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

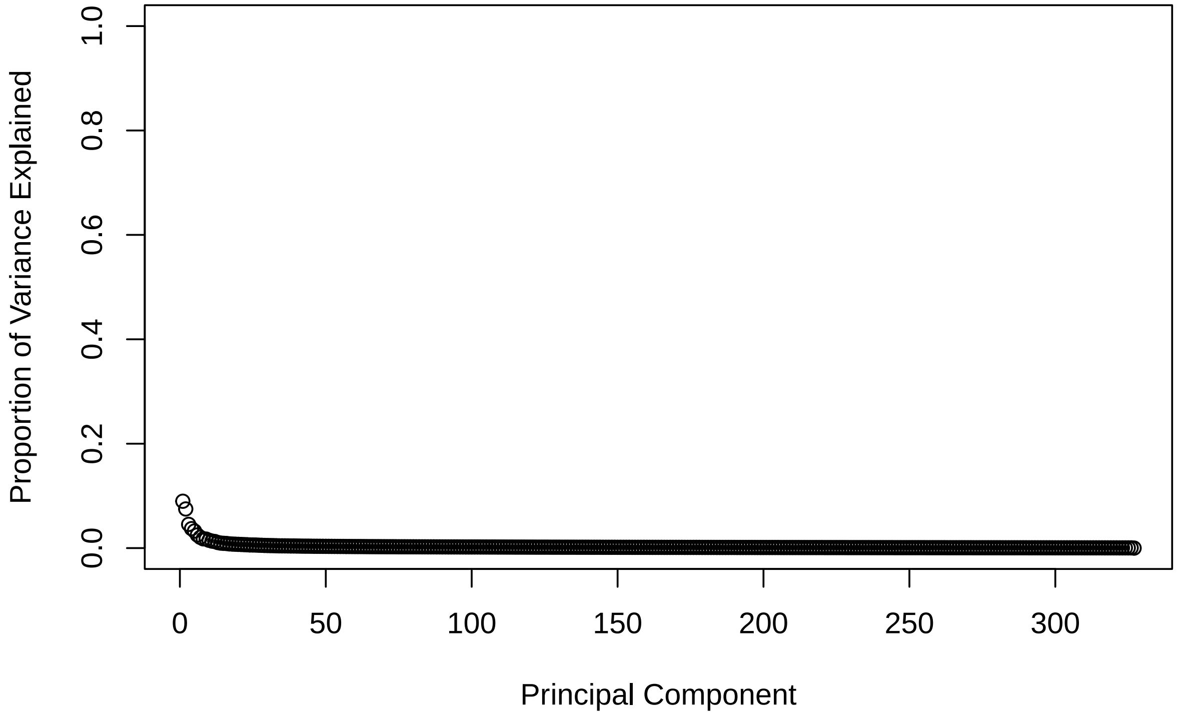
a. From the table, we can see that the least common subtype in the data is "BCR-ABL."

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

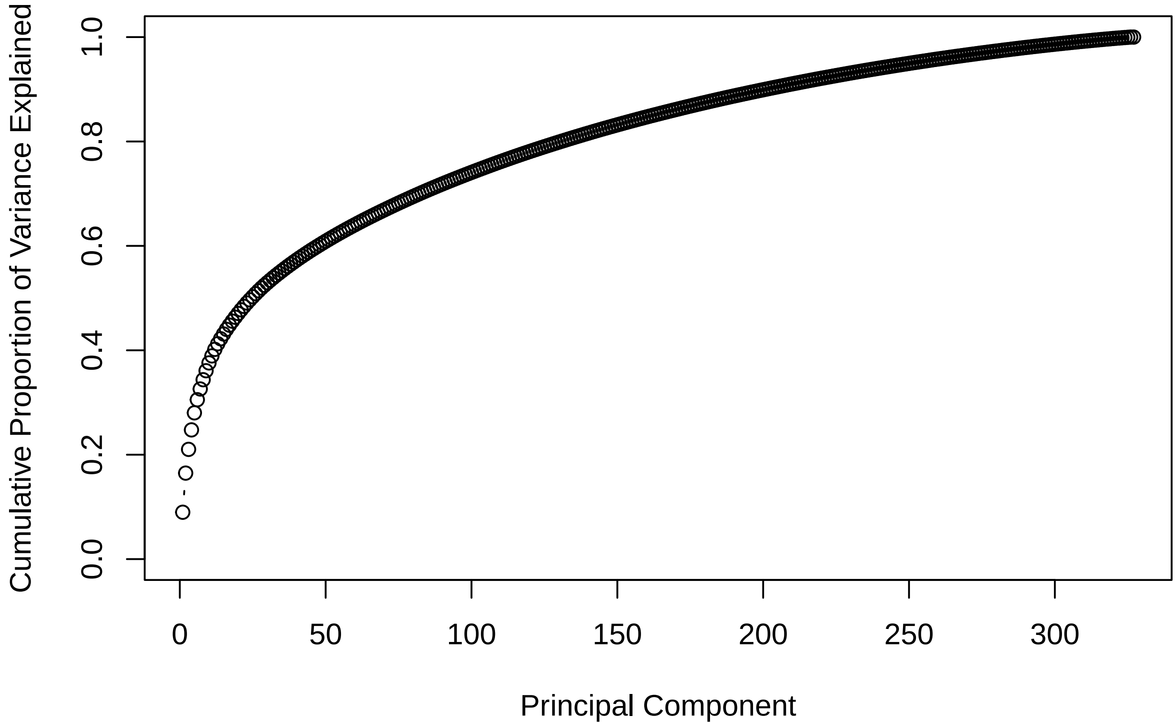
```
##
##    BCR-ABL   E2A-PBX1 Hyperdip50        MLL     OTHERS      T-ALL   TEL-AML1
##         15         27         64         20         79         43         79
```

b. you would need 201 Principal Components to explain 90% of the total variation in the data.

```
pca_leukemia <- prcomp(select(leukemia_data, -Type), scale=TRUE, center=TRUE)

pr.var = pca_leukemia$sdev^2

pve = pr.var/sum(pr.var)

plot(pve, xlab='Principal Component', ylab='Proportion of Variance Explained', ylim=c(0,1),type='b')
```



```
plot(cumsum(pve), xlab='Principal Component', ylab='Cumulative Proportion of Variance Explained', ylim=c
```

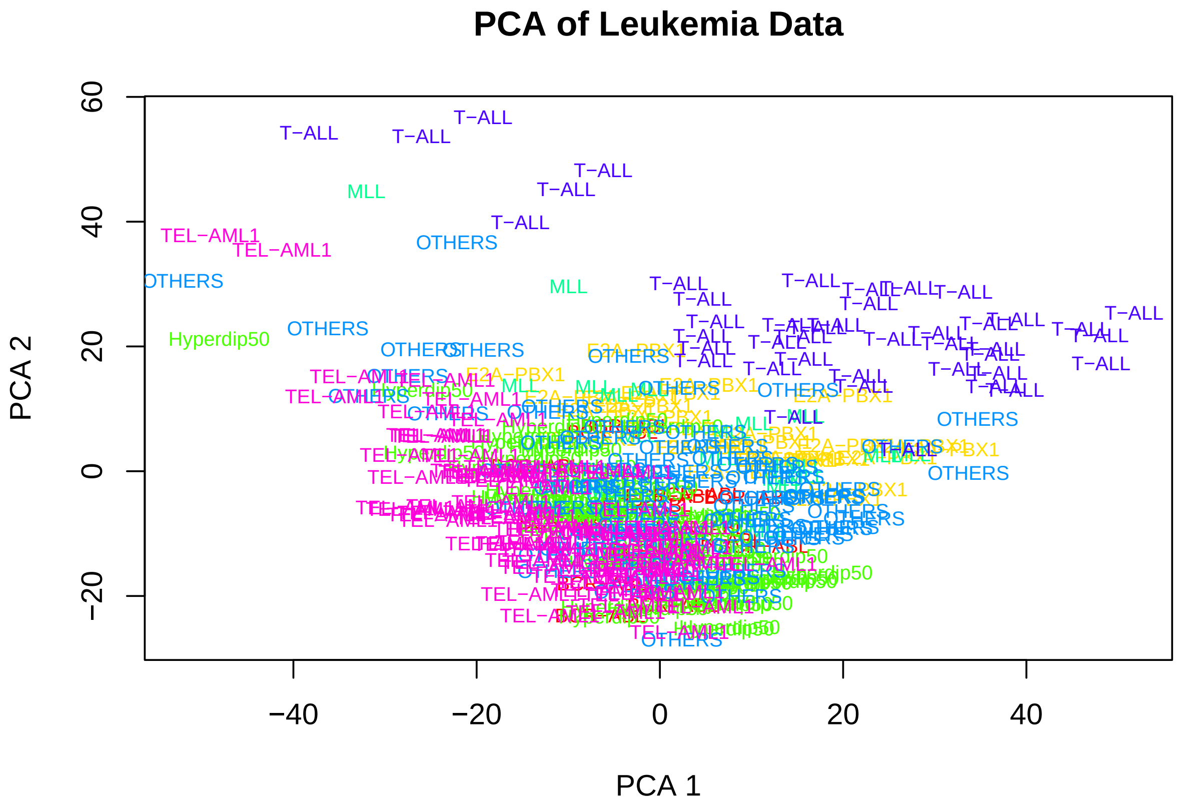


```r
which(cumsum(pve) > .90)[1]
```

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

c. The T-ALL group is most clearly separated from the others along the PC2 axis. The 6 genes with the largest absolute weights are SEMA3F, CCT2, LDHB, COX6C, SNRPD2, and ELK3.

```r
rainbow_colors <- rainbow(7)
plot_colors <- rainbow_colors[as.factor(leukemia_data$Type)]

plot(pca_leukemia$x[, 1], pca_leukemia$x[, 2], col = plot_colors, cex = 0, main = 'PCA of Leukemia Data
text(pca_leukemia$x[, 1], pca_leukemia$x[, 2], labels = leukemia_data$Type, col = plot_colors, cex = .7
```

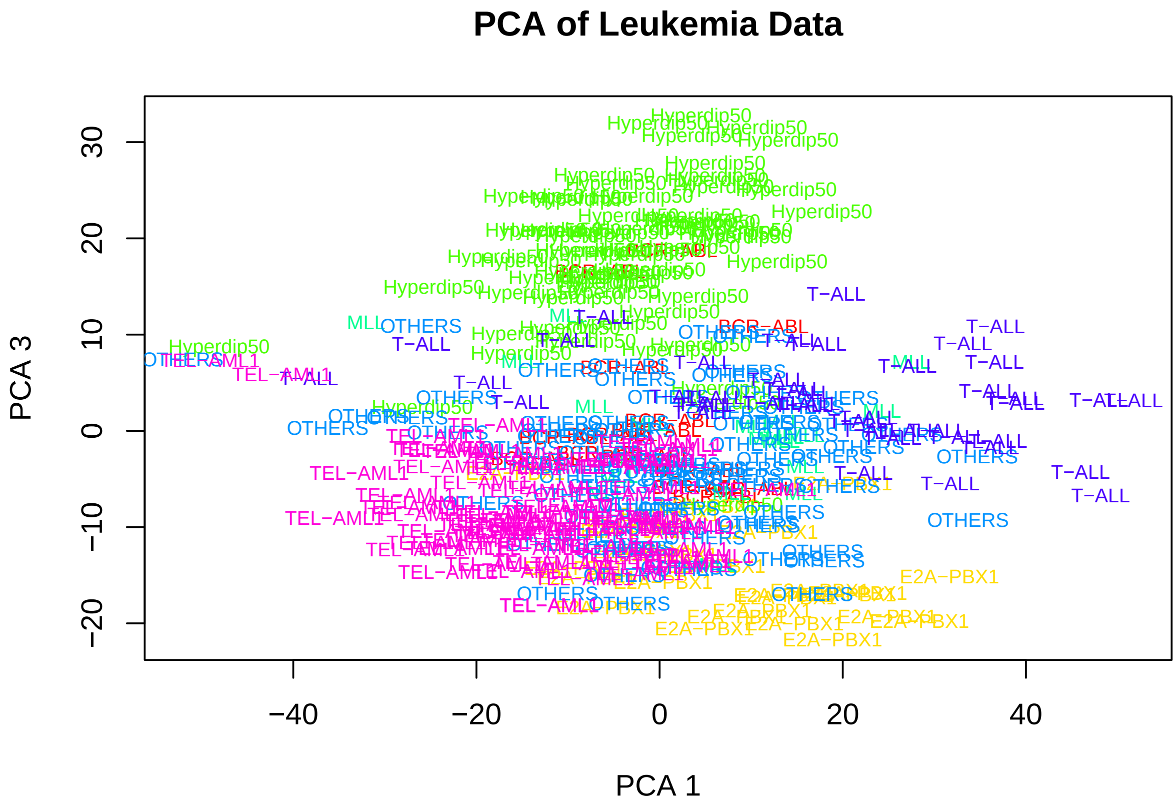


```r
head(names(sort(abs(pca_leukemia$rotation[, 1]), decreasing=TRUE)))
```

```
## [1] "SEMA3F" "CCT2"   "LDHB"   "COX6C"  "SNRPD2" "ELK3"
```

d. The third PC does seem to be better at discriminating between leukemia types because in this plot, the leukemia types are more grouped together between their individual types.

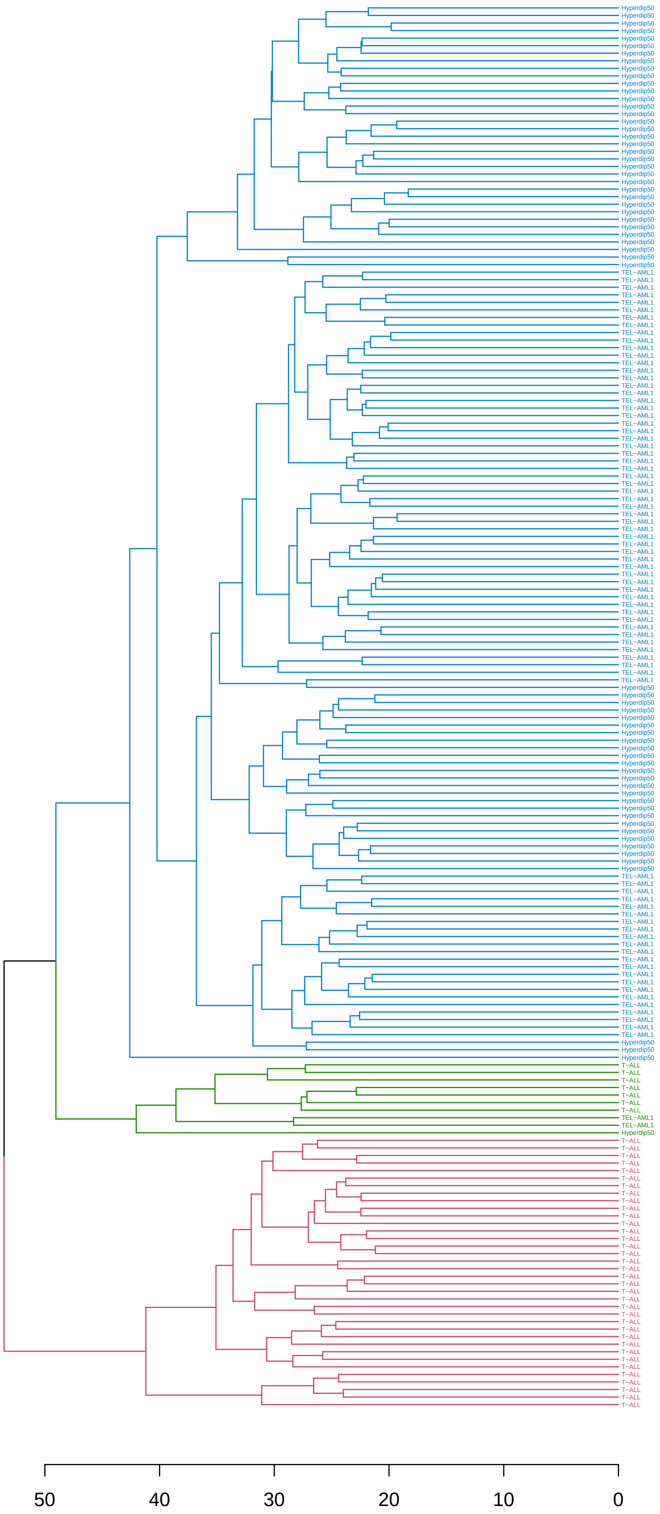
```r
plot(pca_leukemia$x[, 1], pca_leukemia$x[, 3], col = plot_colors, cex = 0, main = 'PCA of Leukemia Data
text(pca_leukemia$x[, 1], pca_leukemia$x[, 3], labels = leukemia_data$Type, col = plot_colors, cex = .7
```



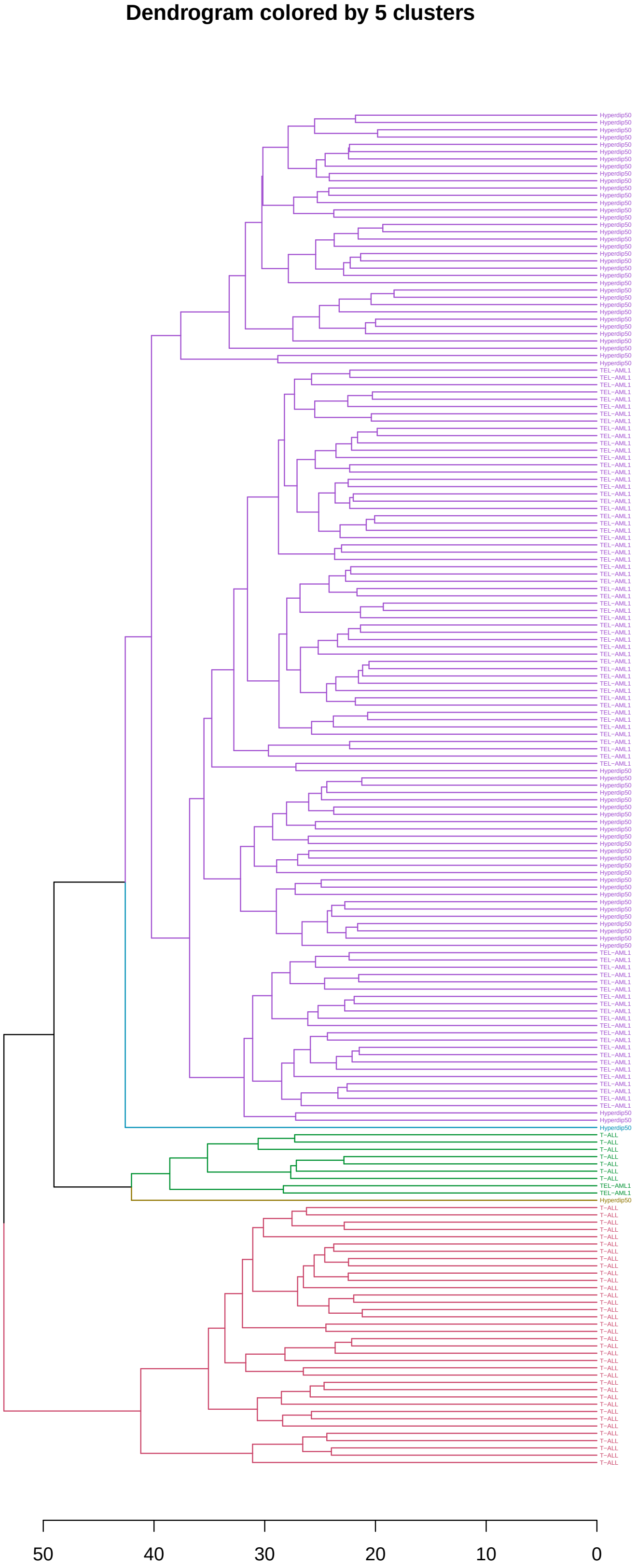
e.

```
leukemia_subset <- filter(leukemia_data, Type %in% c('T-ALL','Hyperdip50','TEL-AML1'))

distance_matrix <- dist(select(leukemia_subset, -Type), method = 'euclidean')
leukemia.hclust <-  hclust(distance_matrix)

dend1 <- as.dendrogram(leukemia.hclust)
dend1 = color_branches(dend1, k=3)
dend1 = color_labels(dend1, k=3)
dend1 = set(dend1, 'labels_cex', 0.3)
dend1 = set_labels(dend1, labels=leukemia_subset$Type[order.dendrogram(dend1)])

plot(dend1, horiz=T, main = "Dendrogram colored by 3 clusters", cex = 0.5)
```

Dendrogram colored by 3 clusters



```r
dend2 <- as.dendrogram(leukemia.hclust)
dend2 = color_branches(dend2, k=5)
dend2 = color_labels(dend2, k=5)
dend2 = set(dend2, 'labels_cex', 0.3)
dend2 = set_labels(dend2, labels=leukemia_subset$Type[order.dendrogram(dend2)])

plot(dend2, horiz=T, main = "Dendrogram colored by 5 clusters", cex = 0.5)
```



Dendrogram colored by 5 clusters