PSTAT131HW4

Tahsin Azad

2023-12-08

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

Clustering and dimension reduction for gene expression data

A.

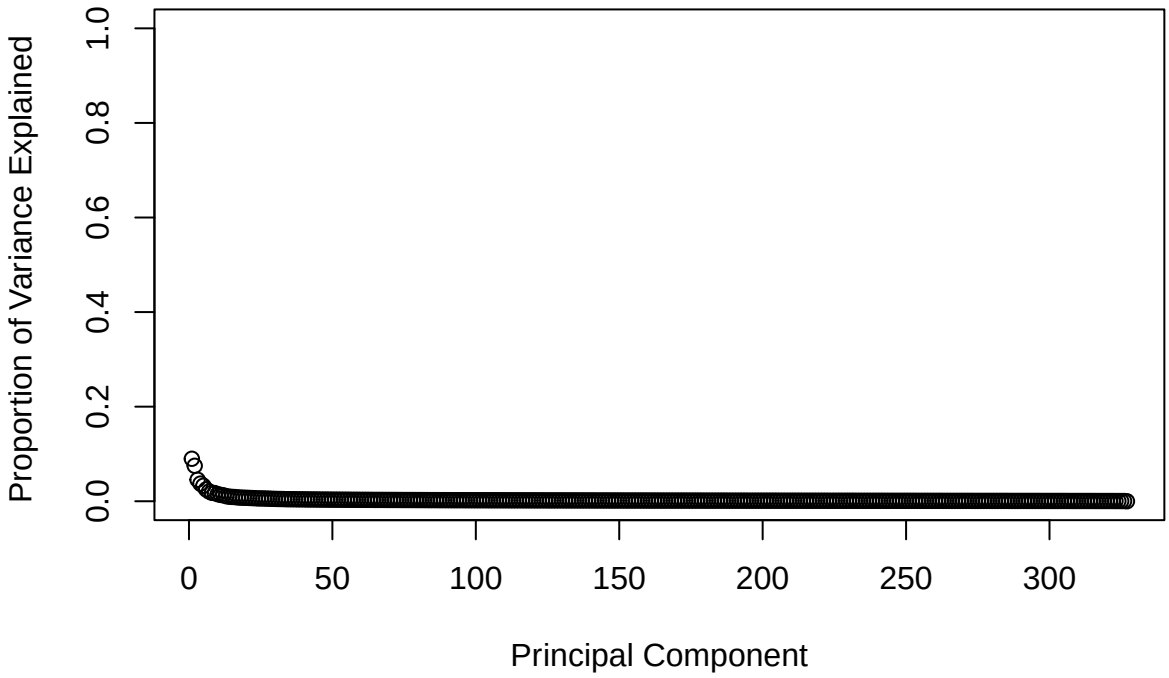
```
leukemia_data <- leukemia_data %>% mutate(Type = as.factor(Type))
table(leukemia_data$Type)
```

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

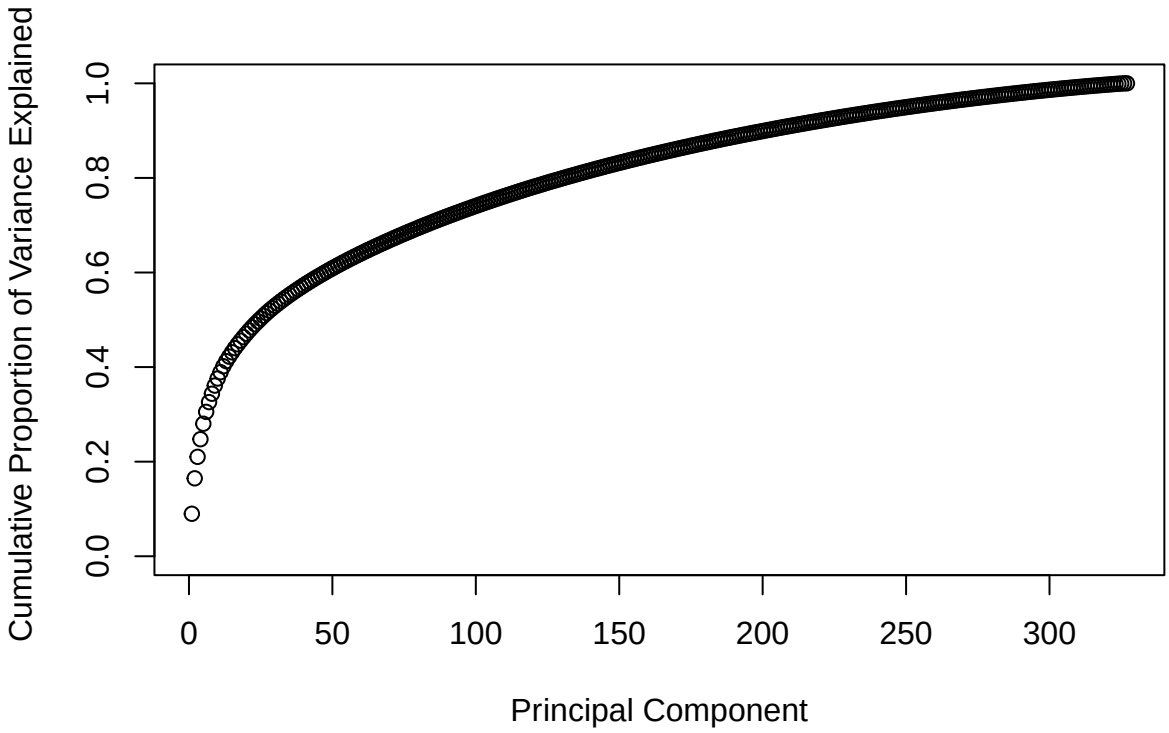
Type MLL occurs the least in this data.

B.

```
leukemia_data_no_type = leukemia_data[,-which(names(leukemia_data) == 'Type')]
pr = prcomp(leukemia_data_no_type, scale = TRUE, center = TRUE)
prpve = pr$sdev^2/ sum(pr$sdev^2)
plot(prpve, xlab="Principal Component",
ylab="Proportion of Variance Explained ", ylim=c(0,1),type='b')
```




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



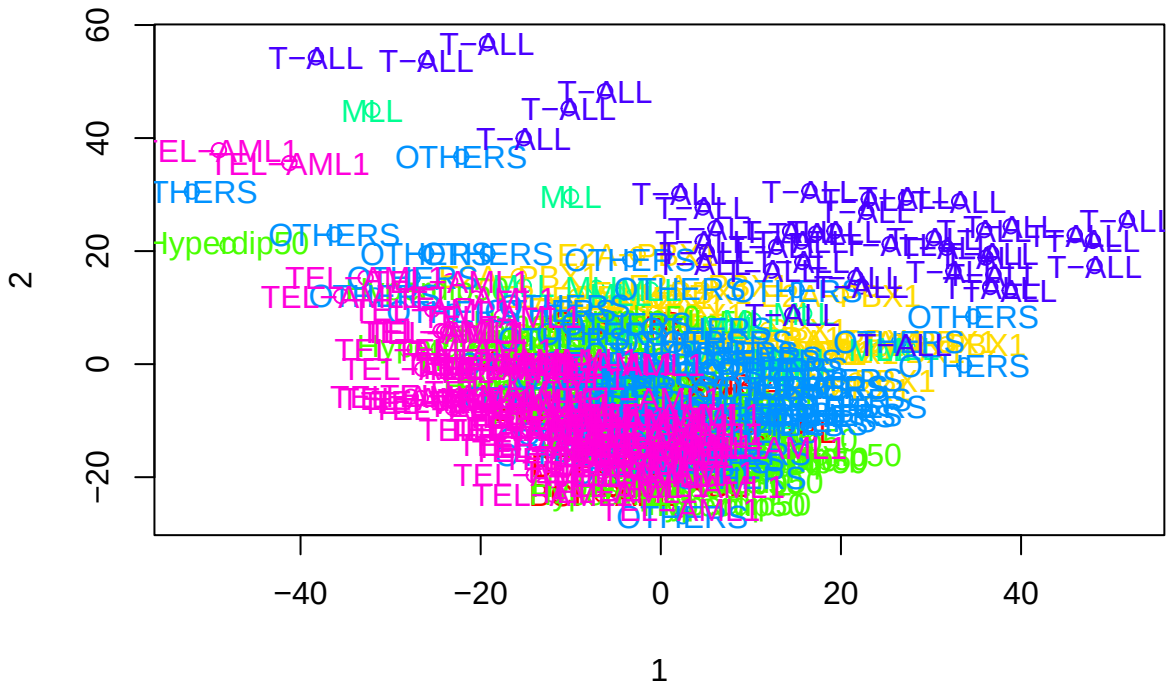

```
which(cumsum(prpve) >= 0.90)[1]
```

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

201 PCs to explain 90% of variation.

C.

```
rainbow_colors <- rainbow(7)
plot_colors <- rainbow_colors[as.factor(leukemia_data$Type)]
first_two <- pr$x[, 1:2]
label = as.character(leukemia_data$Type)
plot(first_two, col = plot_colors, xlab = "1", ylab = "2")
text(first_two, labels = label, col = plot_colors)
```



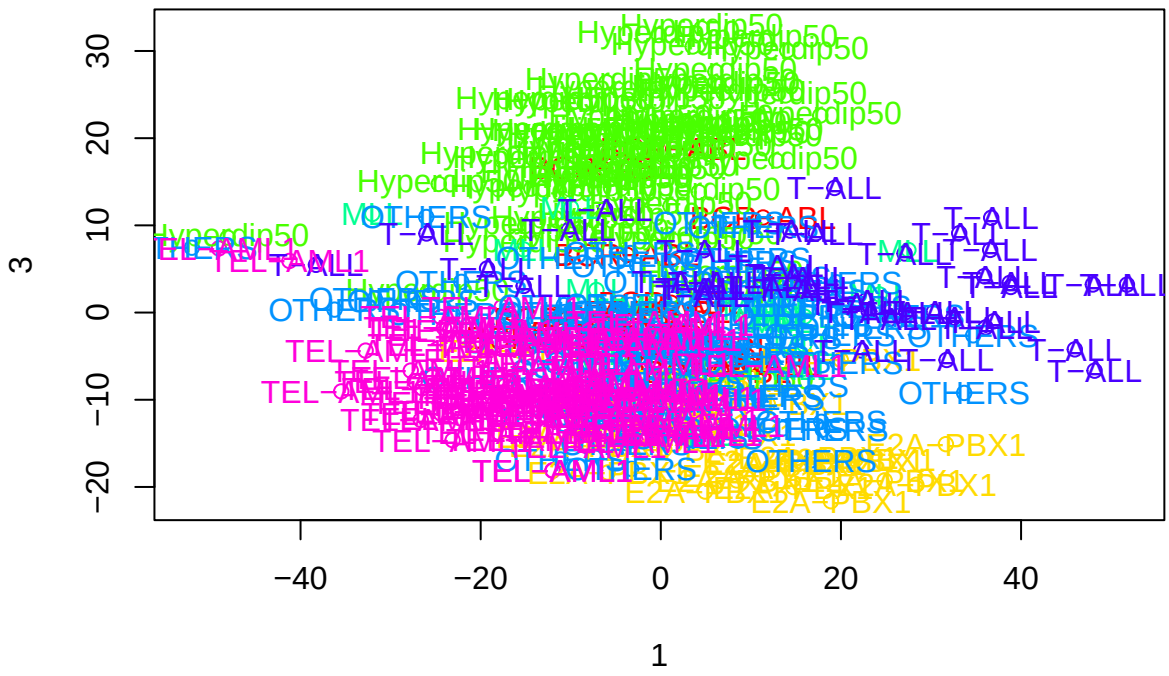
The T-ALL group is clearly seperated from the others along the second score.

```
sorted_vector <- sort(abs(pr$rotation[, 1]), decreasing = TRUE)
head(sorted_vector,6)
```

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

D.

```
first_hird <- pr$x[, c(1, 3)]
plot(first_hird, col = plot_colors, xlab = "1", ylab = "3")
text(first_hird, labels = label, col = plot_colors)
```

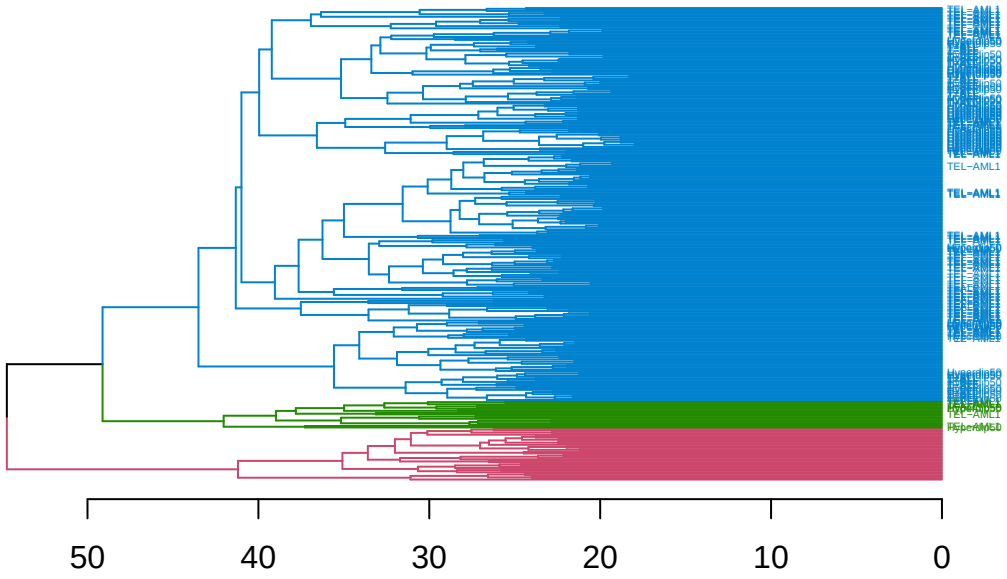


This third one does, there are areas which are specifically only for that type and it seperates it in this way, compared to the second one which did not really do this.

E.

```
leukemia_filtered <- leukemia_data %>% filter(Type %in% c("T-ALL", "TEL-AML1", "Hyperdip50"))
euclidean_dist <- dist(leukemia_data_no_type)
leukemia.hclust <- hclust(euclidean_dist, method = "complete")
```

```
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_filtered$Type[order.dendrogram(dend1)])
plot(dend1, horiz=T, main="Dendrogram Colored by Three Clusters")
```




```
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_filtered$Type[order.dendrogram(dend2)])
plot(dend2, horiz=T, main="Dendrogram Colored by Five Clusters")
```

Dendrogram Colored by Five Clusters