Homework 4

Gabriel Fei

11/30/2021

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

Clustering and dimension reduction for gene expression data

```
leukemia_data <- read_csv("leukemia_data.csv")
class(leukemia_data$Type)
```

```
## [1] "character"
```

a)

```
leukemia_data = leukemia_data %>%
  mutate(Type = as.factor(Type))
class(leukemia_data$Type)
```

```
## [1] "factor"
```

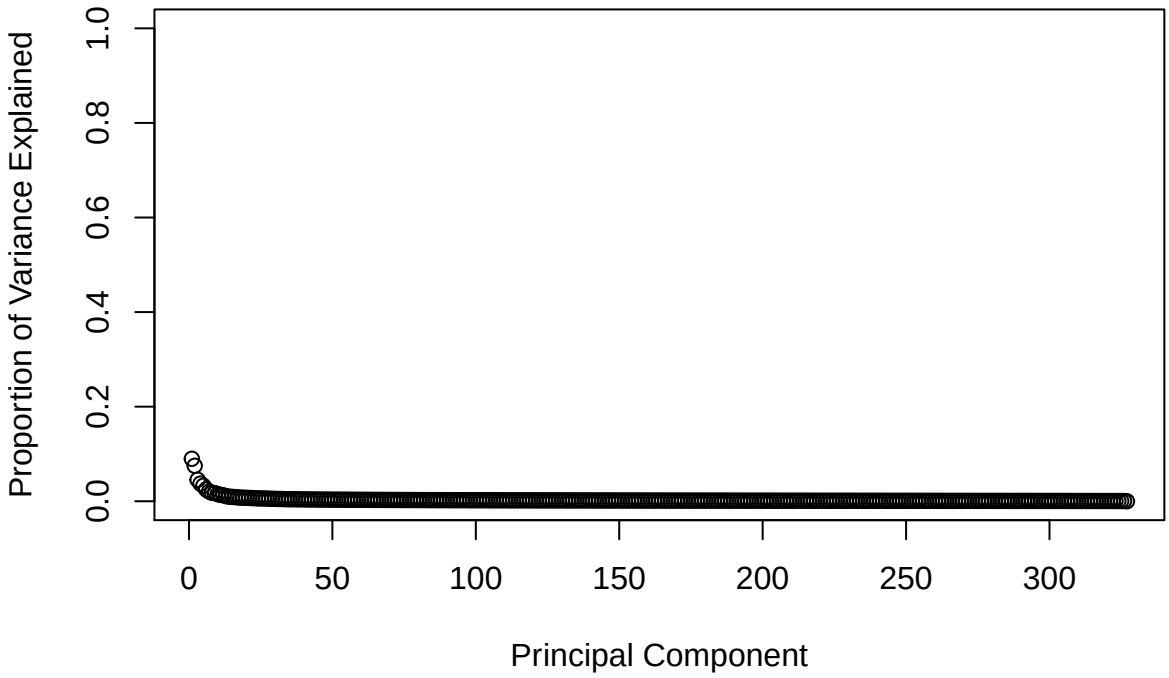
```
table(Type = leukemia_data$Type)
```

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

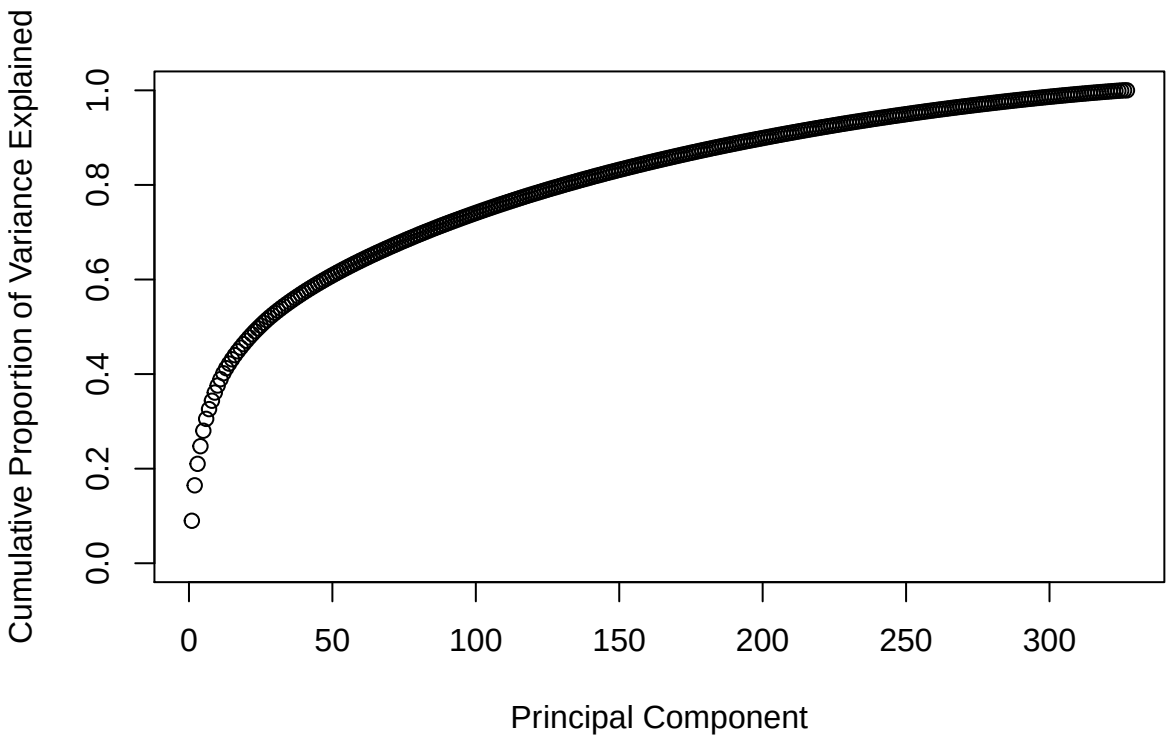
The BCR-ABL leukemia subtype occurs the least in this data with 15 occurances.

b)

```
pr.out = prcomp(leukemia_data[,-c(1)], scale = TRUE, center = TRUE)
pr.var = pr.out$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(0, 1), type = 'b')
```




```
cumsum(pve)
```

```
##   [1] 0.08979245 0.16475316 0.21020956 0.24749596 0.28021500 0.30528743
##   [7] 0.32589670 0.34364880 0.36086742 0.37582671 0.38925117 0.40179078
##  [13] 0.41249429 0.42212222 0.43126702 0.44007352 0.44817716 0.45587498
##  [19] 0.46332304 0.47052503 0.47749391 0.48415799 0.49061751 0.49665743
##  [25] 0.50259676 0.50846104 0.51402683 0.51943000 0.52454233 0.52948493
##  [31] 0.53427437 0.53898902 0.54356760 0.54796229 0.55231010 0.55658687
##  [37] 0.56077287 0.56484225 0.56889939 0.57286143 0.57674317 0.58052278
##  [43] 0.58425568 0.58797182 0.59159107 0.59513092 0.59865454 0.60212762
##  [49] 0.60554539 0.60892888 0.61225083 0.61555540 0.61879406 0.62200551
##  [55] 0.62517545 0.62830733 0.63138438 0.63444633 0.63746784 0.64043363
##  [61] 0.64339736 0.64633592 0.64923340 0.65208053 0.65489494 0.65770032
##  [67] 0.66048081 0.66322973 0.66597061 0.66867846 0.67136962 0.67403560
##  [73] 0.67669393 0.67932247 0.68192320 0.68449823 0.68706387 0.68961143
##  [79] 0.69214460 0.69465318 0.69715113 0.69962146 0.70206347 0.70449043
##  [85] 0.70689470 0.70927720 0.71165083 0.71400820 0.71634026 0.71863939
##  [91] 0.72091903 0.72318189 0.72542585 0.72766212 0.72988904 0.73208917
##  [97] 0.73427295 0.73644543 0.73860102 0.74074420 0.74288174 0.74501320
## [103] 0.74712584 0.74922169 0.75130583 0.75337552 0.75542565 0.75746898
## [109] 0.75949813 0.76150900 0.76350600 0.76548803 0.76745830 0.76942515
## [115] 0.77136010 0.77328757 0.77520285 0.77709962 0.77898915 0.78086407
## [121] 0.78272609 0.78456401 0.78639454 0.78821184 0.79002514 0.79182319
## [127] 0.79361243 0.79538150 0.79714721 0.79890966 0.80065698 0.80239484
## [133] 0.80411891 0.80583086 0.80753909 0.80923311 0.81091861 0.81259793
## [139] 0.81426541 0.81592451 0.81757609 0.81920613 0.82083363 0.82244978
```

```
## [145] 0.82404763 0.82563799 0.82721776 0.82878482 0.83034520 0.83189836
## [151] 0.83343520 0.83496678 0.83648894 0.83800954 0.83951744 0.84101718
## [157] 0.84251350 0.84400335 0.84548685 0.84695375 0.84840578 0.84985127
## [163] 0.85128368 0.85270963 0.85412802 0.85553567 0.85693779 0.85833600
## [169] 0.85972253 0.86110645 0.86248661 0.86386299 0.86522923 0.86657982
## [175] 0.86792473 0.86926695 0.87059809 0.87192143 0.87323441 0.87453894
## [181] 0.87583707 0.87713158 0.87841926 0.87969152 0.88096015 0.88222474
## [187] 0.88347544 0.88472275 0.88596726 0.88719380 0.88840932 0.88961736
## [193] 0.89082086 0.89201677 0.89320691 0.89439224 0.89557306 0.89675010
## [199] 0.89792232 0.89908796 0.90024904 0.90139766 0.90254464 0.90368982
## [205] 0.90482846 0.90595853 0.90708115 0.90819565 0.90930399 0.91040562
## [211] 0.91150511 0.91260410 0.91368793 0.91476647 0.91584315 0.91691208
## [217] 0.91797259 0.91902722 0.92007578 0.92111784 0.92215096 0.92317825
## [223] 0.92420029 0.92521599 0.92622800 0.92723184 0.92822915 0.92922525
## [229] 0.93020917 0.93118973 0.93216178 0.93313292 0.93409314 0.93504714
## [235] 0.93599922 0.93694249 0.93788356 0.93881823 0.93974492 0.94066798
## [241] 0.94158910 0.94250597 0.94341443 0.94432023 0.94521167 0.94609804
## [247] 0.94697855 0.94785587 0.94872709 0.94959388 0.95045892 0.95131627
## [253] 0.95217192 0.95302409 0.95387315 0.95471712 0.95555662 0.95639137
## [259] 0.95721599 0.95803124 0.95884527 0.95965212 0.96044695 0.96123777
## [265] 0.96202694 0.96280574 0.96358212 0.96435186 0.96511655 0.96587753
## [271] 0.96663321 0.96738457 0.96813156 0.96887114 0.96960478 0.97033741
## [277] 0.97106646 0.97179120 0.97251125 0.97322091 0.97392962 0.97463361
## [283] 0.97533670 0.97602924 0.97671582 0.97739669 0.97807634 0.97874941
## [289] 0.97941423 0.98007714 0.98073095 0.98138317 0.98203207 0.98267692
## [295] 0.98331709 0.98394716 0.98457448 0.98519378 0.98580280 0.98640611
## [301] 0.98700276 0.98759576 0.98818598 0.98877066 0.98935322 0.98992918
## [307] 0.99050202 0.99106705 0.99162561 0.99217134 0.99271414 0.99325214
## [313] 0.99378284 0.99430365 0.99482153 0.99533165 0.99583548 0.99633010
## [319] 0.99682233 0.99730540 0.99777858 0.99824363 0.99870048 0.99914571
## [325] 0.99957685 1.00000000 1.00000000
```

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

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

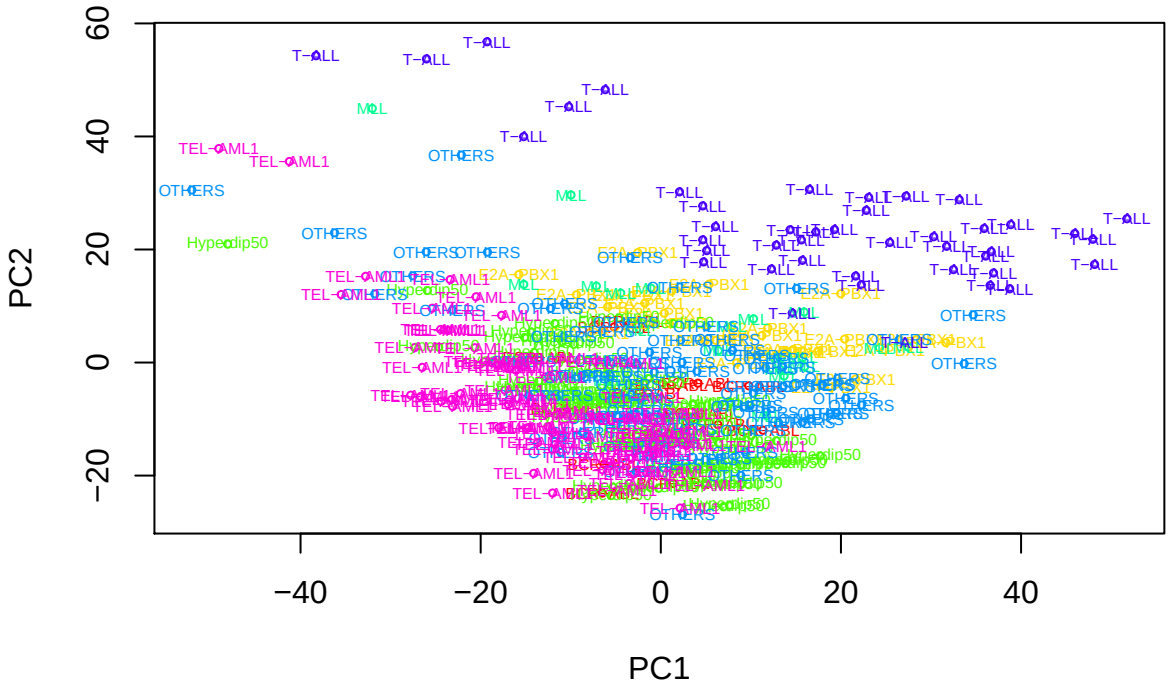
We need at least 201 Principal Compenents in order to explain 90% of the total variation in the data.

c)

```r
rainbow_colors <- rainbow(7)
plot_colors <- rainbow_colors[leukemia_data$Type]
head(pr.out$x[,c(1,2)]) # firct couple values of the first two principal components
```

```
##             PC1        PC2
## [1,] -10.414898  -8.107584
## [2,]  -1.377304  -5.386586
## [3,]  -3.720294   7.290351
## [4,]   1.159456  -3.953322
## [5,]  -5.177178   6.313023
## [6,]  11.346689 -11.979690
```

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



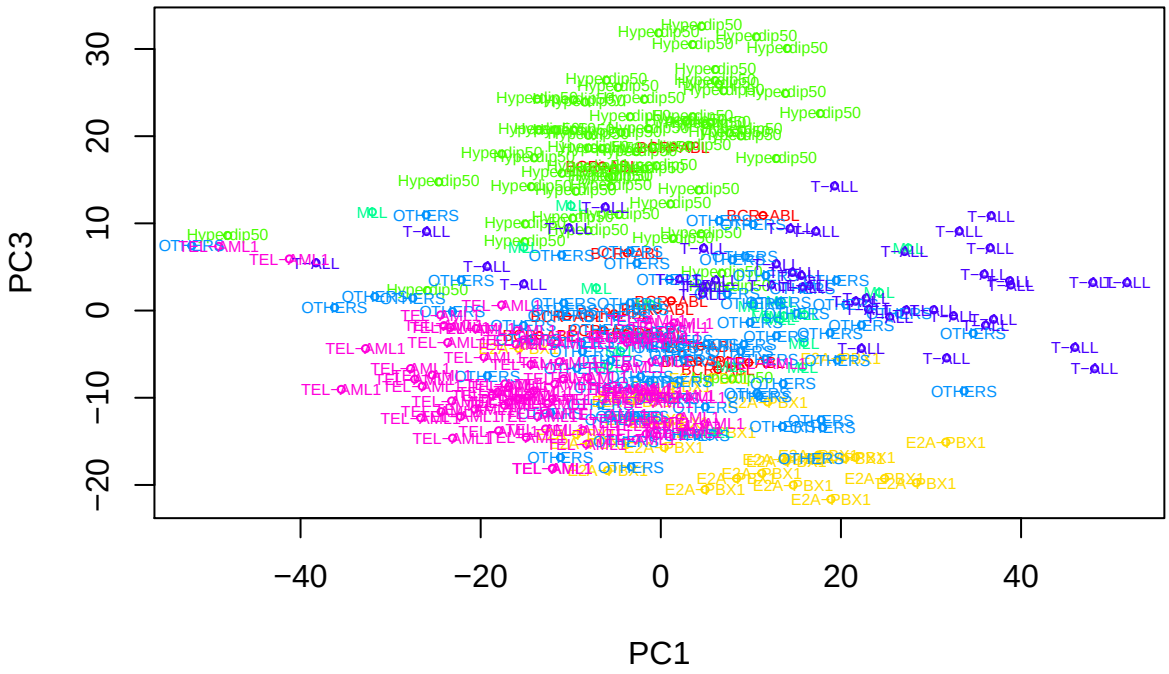

```
ordered_PC1 = sort(abs(pr.out$rotation[, 1]), decreasing = TRUE)
ordered_ind = match(c(head(ordered_PC1, 6)), abs(pr.out$rotation[, 1]))
head(ordered_PC1, 6)
```

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

The T-ALL group is most clearly separated from the others along the PC2 axis. The top 6 highest absolute loadings for PC1 are SEMA3F, CCT2, LDHB, COX6C, SNRPD2, and ELK3.

d)

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



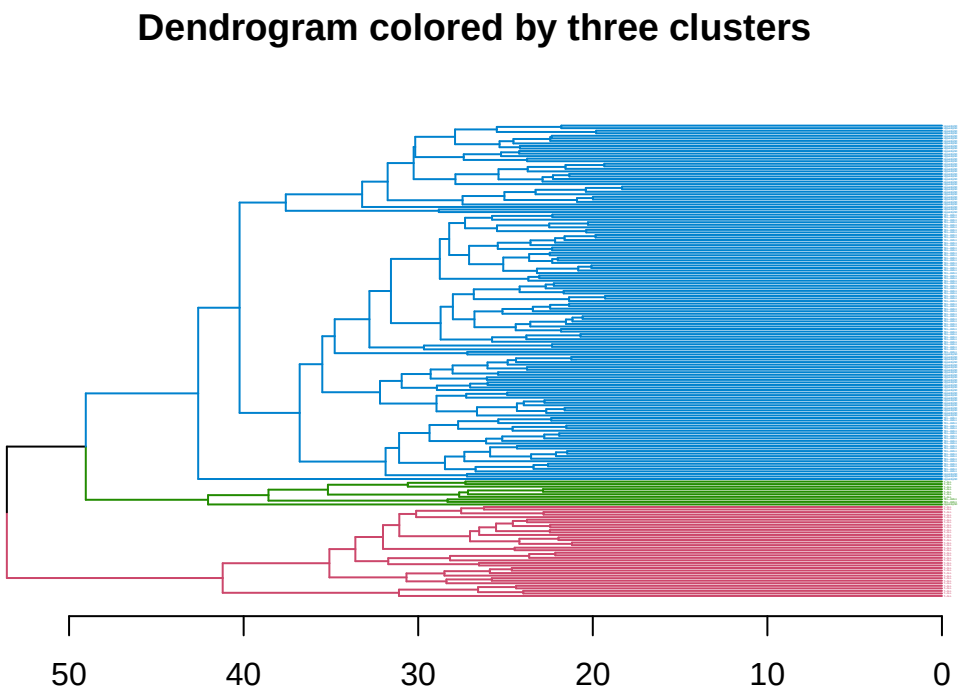
The 3rd PCA does seem to be better at discriminating between leukemia types in comparison to the 2nd PCA as the plot for the 3rd has data that's a bit more separated in comparison to the plot for the 2nd.

e)

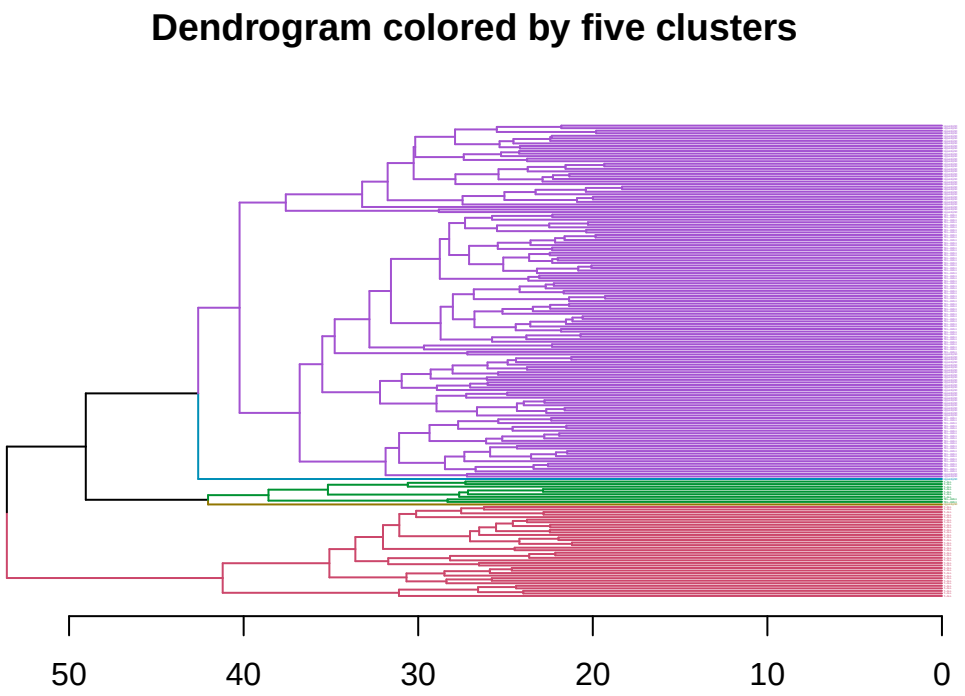
```
set.seed(131)
leukemia_subset = filter(leukemia_data, leukemia_data$Type == "T-ALL" | leukemia_data$Type == "TEL-AML1
leukemia.dist = dist(leukemia_subset, method = "euclidean")
```

```
## Warning in dist(leukemia_subset, method = "euclidean"): NAs introduced by
## coercion
```

```
leukemia.hclust = hclust(leukemia.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", 0.1)
dend1 = set_labels(dend1, labels=leukemia_subset$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", 0.1)
dend2 = set_labels(dend2, labels=leukemia_subset$Type[order.dendrogram(dend2)])
plot(dend2, horiz = T, main = 'Dendrogram colored by five clusters')
```


Dendrogram colored by five clusters
50
40
30
20
10
0