Uncovering Relationship in Gene Expression

maycd

Contents

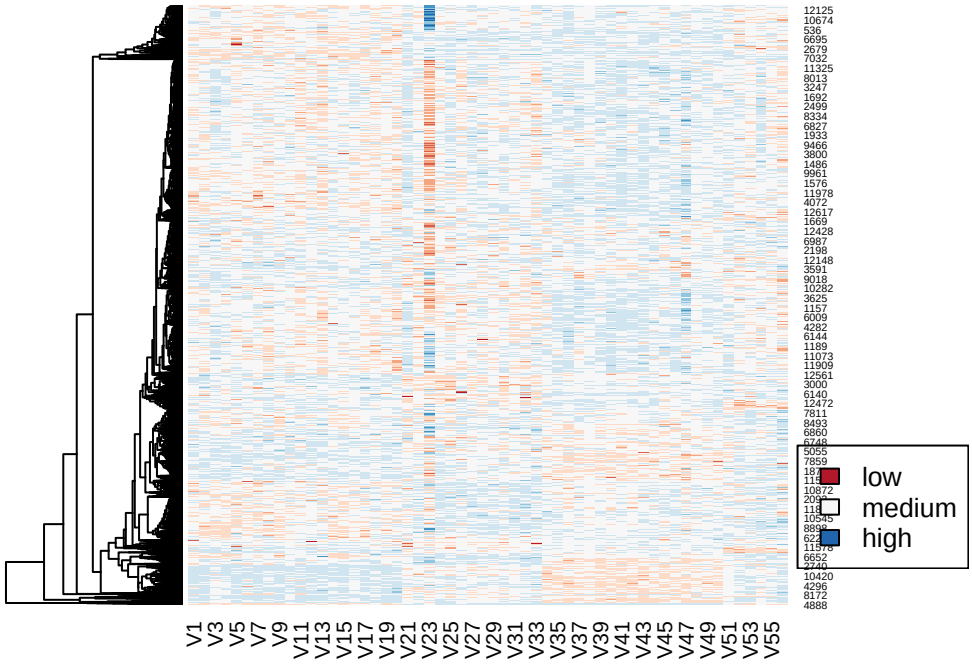
Lung Cancer Data

Lung Cancer Data

```
lung <- data.matrix(read.table("lungcancer.txt"))
```

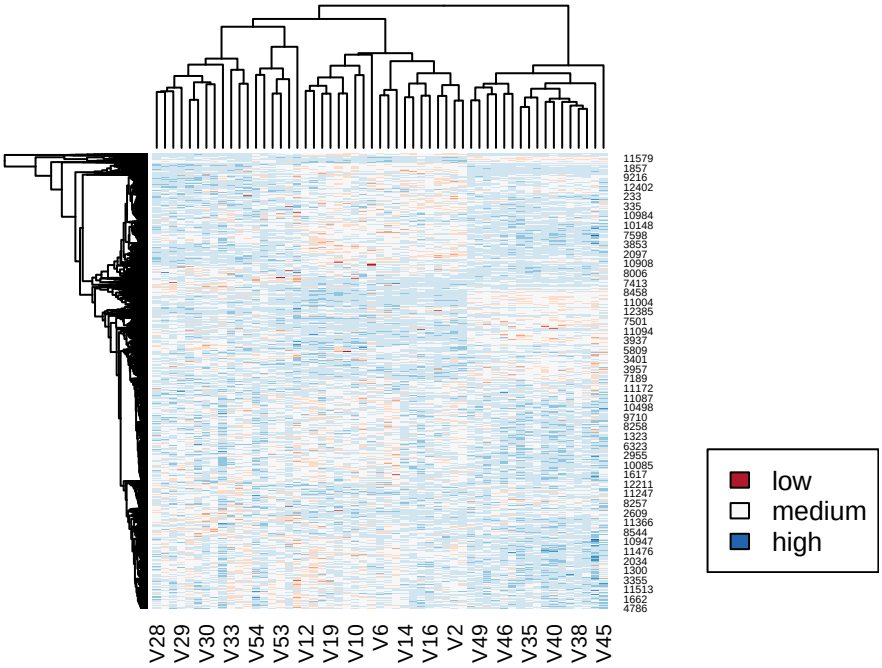
Heatmap of expression of gene vs sample

```
heatmap(lung, Colv=NA, scale="row",
        col=rev(brewer.pal(9,"RdBu")))
legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8, fill = colorRampPalette(brewer.pal(9,"RdBu"))(3))
```



Heatmap without outliers

```
heatmap(lung[,-c(23)], scale="row",
        col=rev(brewer.pal(9,"RdBu")))
legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8, fill = colorRampPalette(brewer.pal(9,"RdBu"))(3))
```



Transpose and normalize the data by subtracting the mean

```
lung <- scale(t(lung), scale = F)
```

1 PCA

1.1 PCA and anomaly detection

```
pca_res0 <- prcomp(lung, scale=T)
summary(pca_res0)
```

```
## Importance of components:
##                             PC1     PC2      PC3     PC4      PC5      PC6
## Standard deviation     50.8052 38.0526 32.07527 26.6844 24.96534 20.89695
## Proportion of Variance  0.2044  0.1147  0.08149  0.0564  0.04937  0.03459
## Cumulative Proportion   0.2044  0.3191  0.40063  0.4570  0.50640  0.54099
##                             PC7      PC8      PC9     PC10     PC11     PC12
## Standard deviation     19.01180 17.44608 17.02844 16.81463 15.90180 14.90014
## Proportion of Variance  0.02863  0.02411  0.02297  0.02239  0.02003  0.01759
## Cumulative Proportion   0.56962  0.59373  0.61670  0.63909  0.65912  0.67671
##                            PC13     PC14     PC15     PC16     PC17     PC18
## Standard deviation     14.33671 13.68204 13.47583 12.73508 12.47303 12.31128
## Proportion of Variance  0.01628  0.01483  0.01438  0.01285  0.01232  0.01201
## Cumulative Proportion   0.69299  0.70781  0.72220  0.73504  0.74737  0.75937
##                            PC19    PC20     PC21     PC22     PC23     PC24
## Standard deviation     11.75366 11.5671 11.22493 10.84332 10.73431 10.56154
## Proportion of Variance  0.01094  0.0106  0.00998  0.00931  0.00913  0.00884
## Cumulative Proportion   0.77031  0.7809  0.79089  0.80020  0.80933  0.81817
##                            PC25     PC26     PC27     PC28    PC29    PC30
## Standard deviation     10.40424 10.25829 10.15815 10.00291 9.83224 9.72541
```

```
## Proportion of Variance  0.00857  0.00834  0.00817  0.00793 0.00766 0.00749
## Cumulative Proportion   0.82674  0.83508  0.84325  0.85118 0.85883 0.86632
##                            PC31    PC32    PC33    PC34    PC35    PC36    PC37
## Standard deviation     9.43479 9.36562 9.20831 9.09100 9.02921 9.01001 8.84051
## Proportion of Variance 0.00705 0.00695 0.00672 0.00655 0.00646 0.00643 0.00619
## Cumulative Proportion  0.87337 0.88032 0.88704 0.89359 0.90004 0.90647 0.91266
##                            PC38    PC39    PC40    PC41    PC42    PC43    PC44
## Standard deviation     8.67532 8.59152 8.54591 8.43000 8.37379 8.28347 8.21472
## Proportion of Variance 0.00596 0.00585 0.00578 0.00563 0.00555 0.00543 0.00535
## Cumulative Proportion  0.91862 0.92447 0.93026 0.93588 0.94144 0.94687 0.95222
##                            PC45    PC46    PC47    PC48    PC49    PC50    PC51
## Standard deviation     8.0217 7.90761 7.79718 7.73543 7.55494 7.46896 7.39256
## Proportion of Variance 0.0051 0.00495 0.00482 0.00474 0.00452 0.00442 0.00433
## Cumulative Proportion  0.9573 0.96227 0.96708 0.97182 0.97634 0.98076 0.98509
##                            PC52    PC53    PC54    PC55     PC56
## Standard deviation     7.24408 7.08031 6.69711 6.38381 4.89e-14
## Proportion of Variance 0.00416 0.00397 0.00355 0.00323 0.00e+00
## Cumulative Proportion  0.98925 0.99322 0.99677 1.00000 1.00e+00
```

Add sample labels as factor to the data

```
lung_df0 <- data.frame(lung)
lung_df0$Label <- c(rep("Carcinoid",20), rep("Colon", 13), rep("Normal", 17), rep("SmallCell", 6))
lung_df0$Label <- factor(lung_df0$Label)
dim(lung_df0)
```

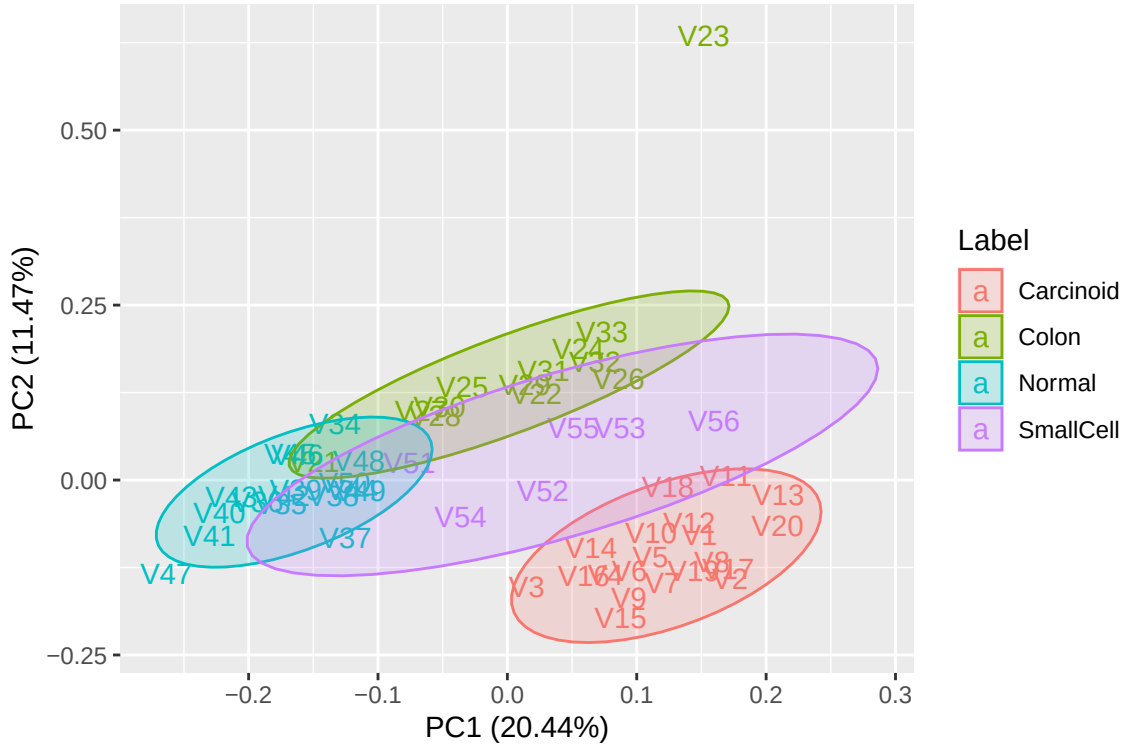
```
## [1]    56 12626
```

```
lung_df0[1:3,12623:12626]
```

```
##         X12623     X12624       X12625     Label
## V1 -0.05674101 -0.3349708  0.008610441 Carcinoid
## V2 -0.08102412 -0.3774571  0.019483222 Carcinoid
## V3 -0.19249953 -0.6011643 -0.144426416 Carcinoid
```

Plot the first two principal components

```
autoplot(pca_res0, data = lung_df0, shape = F, colour = "Label", frame = T, frame.type = "t")
```



The 23rd sample is an outlier to be removed. It is far away from other samples in the group of Colon.

```
lung_df1 <- lung_df0[-c(23),]
dim(lung_df1)
```

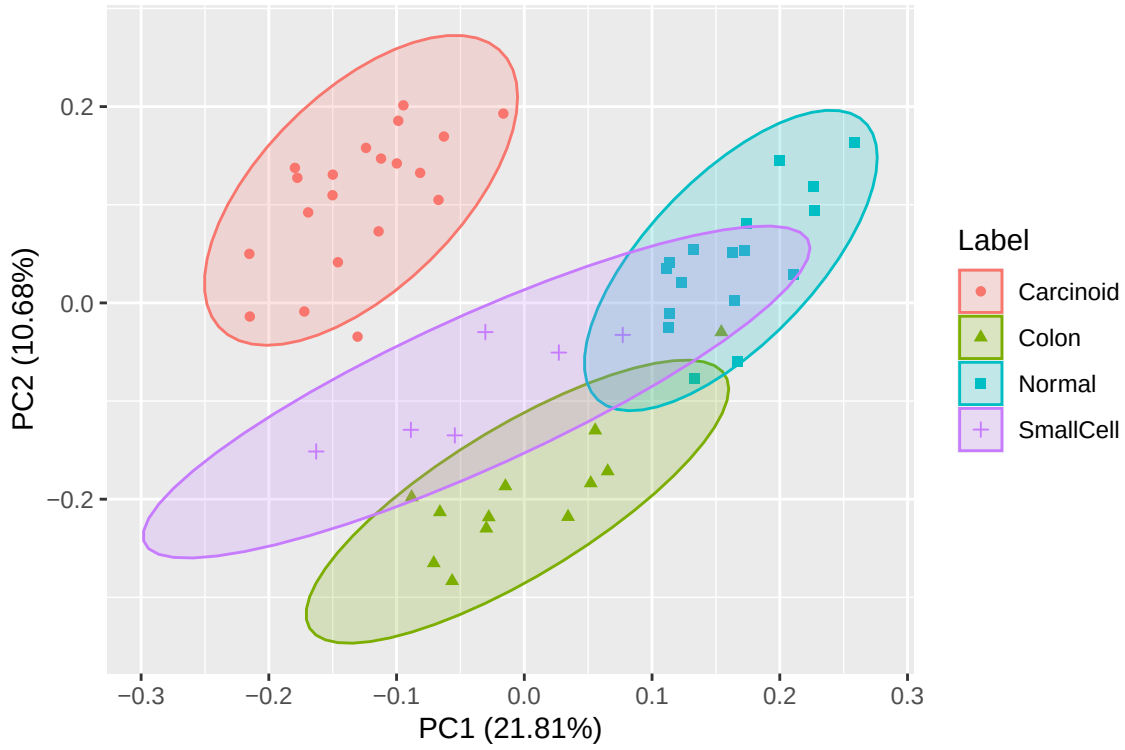
```
## [1]    55 12626
```

```
pca_res1 <- prcomp(lung_df1[,-c(12626)], scale=T)
summary(pca_res1)
```

```
## Importance of components:
##                             PC1     PC2      PC3      PC4      PC5      PC6
## Standard deviation     52.4753 36.7182 27.88480 26.06401 22.42749 20.12299
## Proportion of Variance  0.2181  0.1068  0.06159  0.05381  0.03984  0.03207
## Cumulative Proportion   0.2181  0.3249  0.38649  0.44030  0.48014  0.51221
##                             PC7      PC8      PC9     PC10     PC11     PC12
## Standard deviation     19.71447 18.21333 17.66801 16.83474 15.80742 15.07840
## Proportion of Variance  0.03078  0.02628  0.02473  0.02245  0.01979  0.01801
## Cumulative Proportion   0.54300  0.56927  0.59400  0.61645  0.63624  0.65425
##                            PC13    PC14     PC15     PC16     PC17     PC18
## Standard deviation     14.36515 14.2137 13.58011 13.21515 13.12813 12.48127
## Proportion of Variance  0.01635  0.0160  0.01461  0.01383  0.01365  0.01234
## Cumulative Proportion   0.67059  0.6866  0.70120  0.71504  0.72869  0.74103
##                            PC19    PC20     PC21     PC22     PC23     PC24
## Standard deviation     12.24277 11.8364 11.64913 11.32566 11.14800 10.97020
## Proportion of Variance  0.01187  0.0111  0.01075  0.01016  0.00984  0.00953
## Cumulative Proportion   0.75290  0.7640  0.77475  0.78491  0.79475  0.80428
##                            PC25     PC26     PC27    PC28     PC29     PC30
## Standard deviation     10.8362 10.73552 10.57006 10.4219 10.29102 10.14782
## Proportion of Variance  0.0093  0.00913  0.00885  0.0086  0.00839  0.00816
## Cumulative Proportion   0.8136  0.82271  0.83156  0.8402  0.84855  0.85671
```

```
##                            PC31    PC32    PC33    PC34    PC35    PC36    PC37
## Standard deviation     9.95289 9.74672 9.65557 9.60595 9.53929 9.34771 9.23670
## Proportion of Variance 0.00785 0.00752 0.00738 0.00731 0.00721 0.00692 0.00676
## Cumulative Proportion  0.86456 0.87208 0.87946 0.88677 0.89398 0.90090 0.90766
##                            PC38    PC39    PC40    PC41    PC42    PC43    PC44
## Standard deviation     9.10760 9.05061 8.96241 8.88134 8.84274 8.72712 8.62657
## Proportion of Variance 0.00657 0.00649 0.00636 0.00625 0.00619 0.00603 0.00589
## Cumulative Proportion  0.91423 0.92072 0.92708 0.93333 0.93952 0.94555 0.95145
##                            PC45    PC46    PC47    PC48    PC49    PC50   PC51
## Standard deviation     8.46080 8.31939 8.22816 8.06926 7.95398 7.90153 7.7037
## Proportion of Variance 0.00567 0.00548 0.00536 0.00516 0.00501 0.00495 0.0047
## Cumulative Proportion  0.95712 0.96260 0.96796 0.97312 0.97813 0.98308 0.9878
##                            PC52    PC53    PC54      PC55
## Standard deviation     7.50830 7.16943 6.82036 4.359e-14
## Proportion of Variance 0.00447 0.00407 0.00368 0.000e+00
## Cumulative Proportion  0.99224 0.99632 1.00000 1.000e+00
```

```
autoplot(pca_res1, data = lung_df1, shape = "Label", colour = "Label", frame = T, frame.type = "t")
```



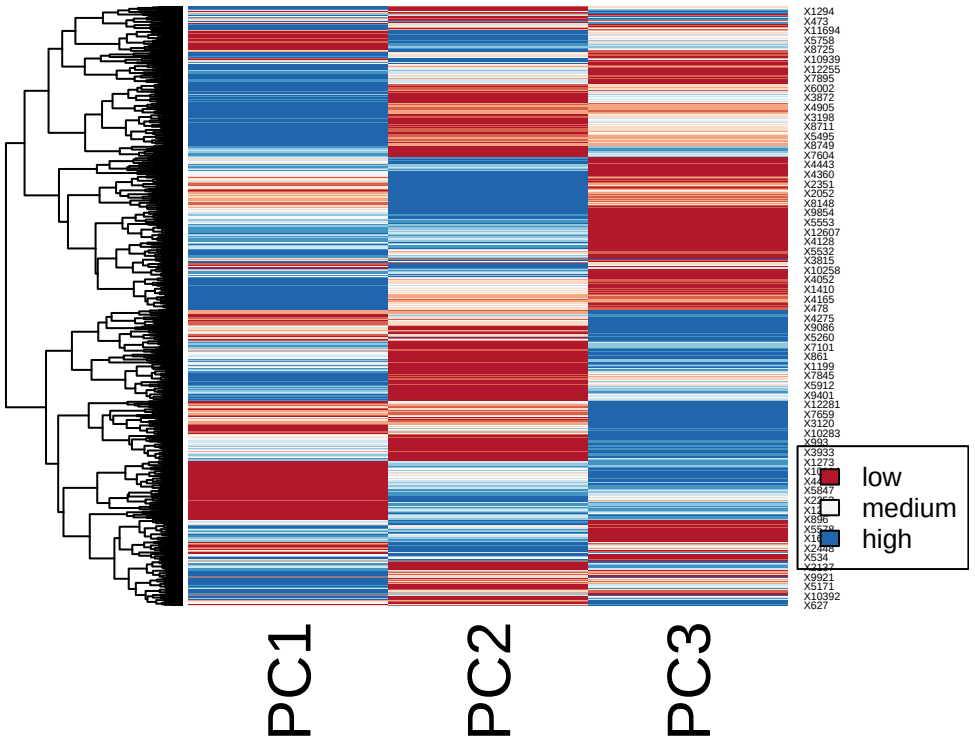
The loadings of the three principal components on some genes

```
pca_res1_loading <- data.frame(round(pca_res1$rotation[,1:3], 4))
head(pca_res1_loading)
```

```
##        PC1     PC2     PC3
## X1  0.0006  0.0041  0.0041
## X2  0.0145  0.0047 -0.0036
## X3 -0.0061 -0.0135 -0.0040
## X4 -0.0081 -0.0115 -0.0062
## X5 -0.0087 -0.0133 -0.0063
## X6  0.0131  0.0060 -0.0017
```

Heatmap of loading on gene vs PC

```
heatmap(pca_res1$rotation[,1:3], Colv=NA, scale="row",
        col=rev(brewer.pal(9,"RdBu")))
legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8, fill = colorRampPalette(brewer.pal(9,"RdBu"))(3))
```



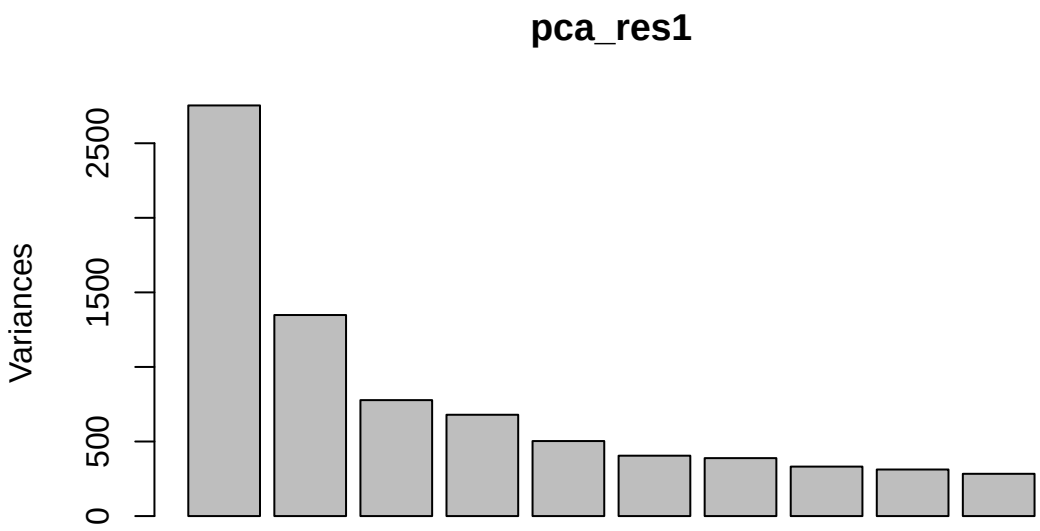

```
head(pca_res1_loading[order(-pca_res1_loading$PC1, -pca_res1_loading$PC2, -pca_res1_loading$PC3), ])
```

```
##            PC1     PC2     PC3
## X8188   0.0180 -0.0023 -0.0004
## X10242 0.0179 -0.0022 -0.0038
## X11613 0.0179 -0.0040 -0.0012
## X7474   0.0178 -0.0007 -0.0030
## X3359   0.0177  0.0044 -0.0034
## X9269   0.0177 -0.0026 -0.0041
```

It appears that the contrast of some genes vs. other genes constituting PC1 influences the response the most.

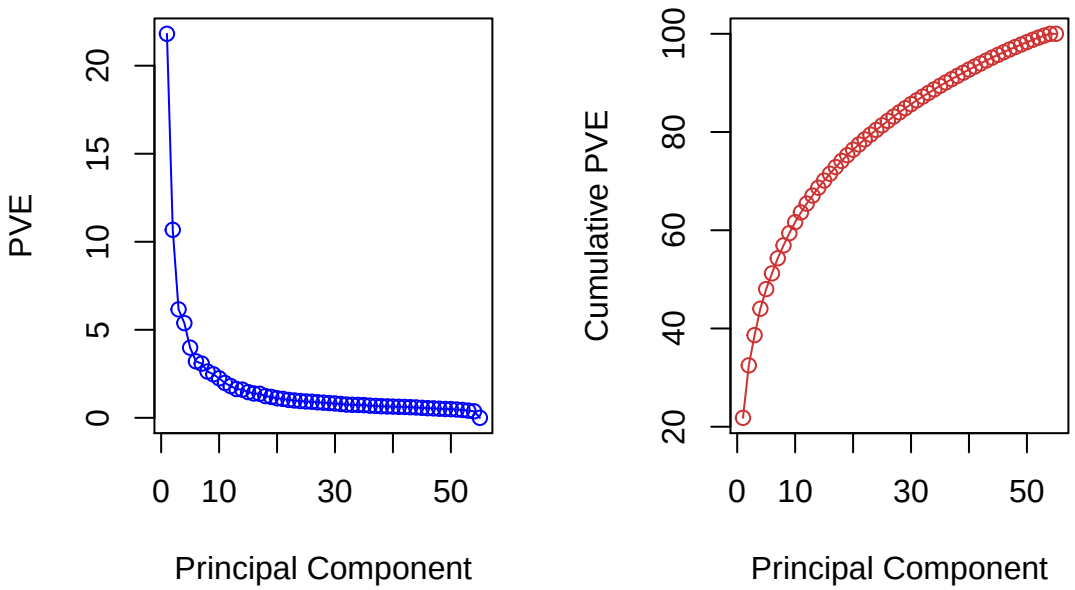
1.2 Scree plot

```
screeplot(pca_res1)
```



There is a marked decrease in the variance explained by further principal components.

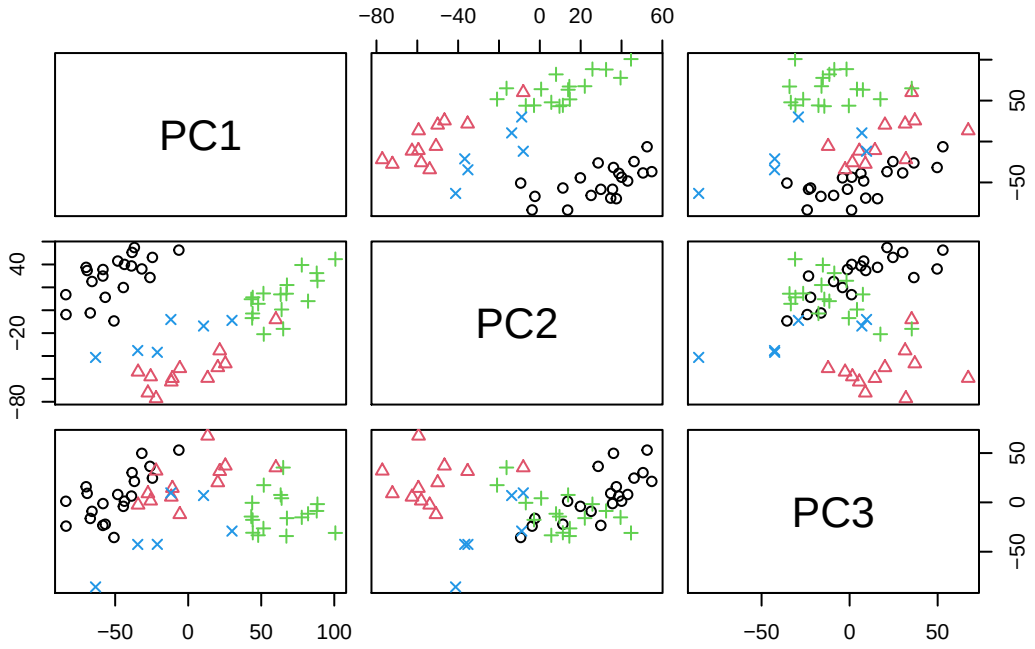
```
pve <- 100 * pca_res1$sdev^2 / sum(pca_res1$sdev^2)
par(mfrow = c(1, 2))
plot(pve, type = "o", ylab = "PVE", xlab = "Principal Component", col = "blue")
plot(cumsum(pve), type = "o", ylab = "Cumulative PVE",
xlab = "Principal Component", col = "brown3")
```



The first three components explain about 40% of the variation in data. However, there is an elbow in the plot after approximately the third principal component in the scree plot. Thus, the three components are sufficient.

1.3 Pair-wise scatterplots

```
pairs(pca_res1$x[,1:3], col=lung_df1$Label, pch=as.numeric(lung_df1$Label))
```



They are pairwise uncorrelated.

2 Nominal Logistic Regression, LDA and SVM

```
lung_df2 <- data.frame(pca_res1$x[,1:3])
lung_df2$Label <- lung_df1$Label
dim(lung_df2)
```

```
## [1] 55  4
```

2.1 Nominal logistic

```
lung.mult <- multinom(Label ~ ., data = lung_df2)
```

```
## # weights:  20 (12 variable)
## initial  value 76.246190
## iter  10 value 16.222444
## iter  20 value 3.999215
## iter  30 value 2.942475
## iter  40 value 1.481411
## iter  50 value 0.002899
## iter  60 value 0.001055
## iter  70 value 0.000781
## iter  80 value 0.000352
## iter  90 value 0.000314
## iter 100 value 0.000309
## final  value 0.000309
## stopped after 100 iterations
```

```r
summary(lung.mult)
```

```
## Call:
## multinom(formula = Label ~ ., data = lung_df2)
##
## Coefficients:
##           (Intercept)       PC1        PC2       PC3
## Colon       -60.42133  9.112873 -16.431392 12.930505
## Normal     -104.76296 18.575661 -13.007192 -1.399957
## SmallCell   179.11517  9.449174  -7.722415 -3.008478
##
## Std. Errors:
##           (Intercept)       PC1      PC2      PC3
## Colon        324.6174 12.003227 6.207454 14.40315
## Normal       529.3444  9.241709 8.695097 11.50943
## SmallCell    204.7323 13.942131 6.412376 16.45732
##
## Residual Deviance: 0.0006186194
## AIC: 24.00062
```

With one-unit increase in PC1, the probability of Colon to the probability of Carcinoid will increase by a multiplicative factor of $e^{9.112873}$.

```r
Anova(lung.mult)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Label
##     LR Chisq Df Pr(>Chisq)
## PC1   62.169  3  2.022e-13 ***
## PC2   26.137  3  8.928e-06 ***
## PC3   12.407  3   0.006112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values are less than 0.05. The variables are significant.

2.2 LDA

```r
lung.lda <- linDA(lung_df2[,1:3], lung_df2$Label)
lung.lda$functions
```

```
##            Carcinoid       Colon     Normal   SmallCell
## constant -14.5596327 -9.87534588 -9.3977403 -3.89101340
## PC1       -0.3309840  0.14807608  0.2644855  0.05775209
## PC2        0.3401735 -0.30402911 -0.1502391 -0.10017597
## PC3        0.0982544  0.03079339 -0.1163329 -0.05949161
```

$$\hat{c}(Carcinoid|PC1, PC2, PC3) = -14.5596 - 0.3310 * PC1 + 0.3402 * PC2 + 0.0983 * PC3$$
$$\hat{c}(Colon|PC1, PC2, PC3) = -9.8753 + 0.1481 * PC1 - 0.3040 * PC2 + 0.0308 * PC3$$
$$\dots$$

```r
confusionMatrix(lung.lda$classification, lung_df2$Label)
```

```
## Confusion Matrix and Statistics
```

```
##
##            Reference
## Prediction  Carcinoid Colon Normal SmallCell
##   Carcinoid        19     0      0         0
##   Colon             0    11      0         0
##   Normal            0     1     17         1
##   SmallCell         1     0      0         5
##
## Overall Statistics
##
##                Accuracy : 0.9455
##                  95% CI : (0.8488, 0.9886)
##     No Information Rate : 0.3636
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9234
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Carcinoid Class: Colon Class: Normal
## Sensitivity                    0.9500       0.9167        1.0000
## Specificity                    1.0000       1.0000        0.9474
## Pos Pred Value                 1.0000       1.0000        0.8947
## Neg Pred Value                 0.9722       0.9773        1.0000
## Prevalence                     0.3636       0.2182        0.3091
## Detection Rate                 0.3455       0.2000        0.3091
## Detection Prevalence           0.3455       0.2000        0.3455
## Balanced Accuracy              0.9750       0.9583        0.9737
##                      Class: SmallCell
## Sensitivity                   0.83333
## Specificity                   0.97959
## Pos Pred Value                0.83333
## Neg Pred Value                0.97959
## Prevalence                    0.10909
## Detection Rate                0.09091
## Detection Prevalence          0.10909
## Balanced Accuracy             0.90646
```

The vast majority of data points are classified correctly.

2.3 SVM

From the plot in 2.1.1, the points are not linearly separable.

```
lung.svm1 <- svm(Label ~ ., data = lung_df2, kernel = "linear",
cost = 10, scale = FALSE)
summary(lung.svm1)
```

```
##
## Call:
## svm(formula = Label ~ ., data = lung_df2, kernel = "linear", cost = 10,
##     scale = FALSE)
##
```

```
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  10
##
## Number of Support Vectors:  16
##
##  ( 3 4 5 4 )
##
##
## Number of Classes:  4
##
## Levels:
##  Carcinoid Colon Normal SmallCell
```

The indeces of support vectors are

```
print(lung.svm1$index)
```

```
##  [1]  3 14 18 21 25 27 28 43 44 47 48 49 50 51 53 55
```

```
set.seed(1)
tune.out <- tune(svm, Label ~ ., data = lung_df2, kernel = "linear",
                 ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)))
summary(tune.out)
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##     1
##
## - best performance: 0.07666667
##
## - Detailed performance results:
##    cost      error dispersion
## 1 1e-03 0.64666667 0.18803730
## 2 1e-02 0.61000000 0.23467866
## 3 1e-01 0.13333333 0.13240417
## 4 1e+00 0.07666667 0.09944289
## 5 5e+00 0.09333333 0.13680661
## 6 1e+01 0.09333333 0.13680661
## 7 1e+02 0.09333333 0.13680661
```

We see that cost = 1 results in the lowest cross-validation error rate. The best SVM model is

```
lung.svm.bestmod <- tune.out$best.model
summary(lung.svm.bestmod)
```

```
##
## Call:
## best.tune(method = svm, train.x = Label ~ ., data = lung_df2, ranges = list(cost = c(0.001,
##     0.01, 0.1, 1, 5, 10, 100)), kernel = "linear")
```

```
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  20
##
##  ( 5 5 5 5 )
##
##
## Number of Classes:  4
##
## Levels:
##  Carcinoid Colon Normal SmallCell
```

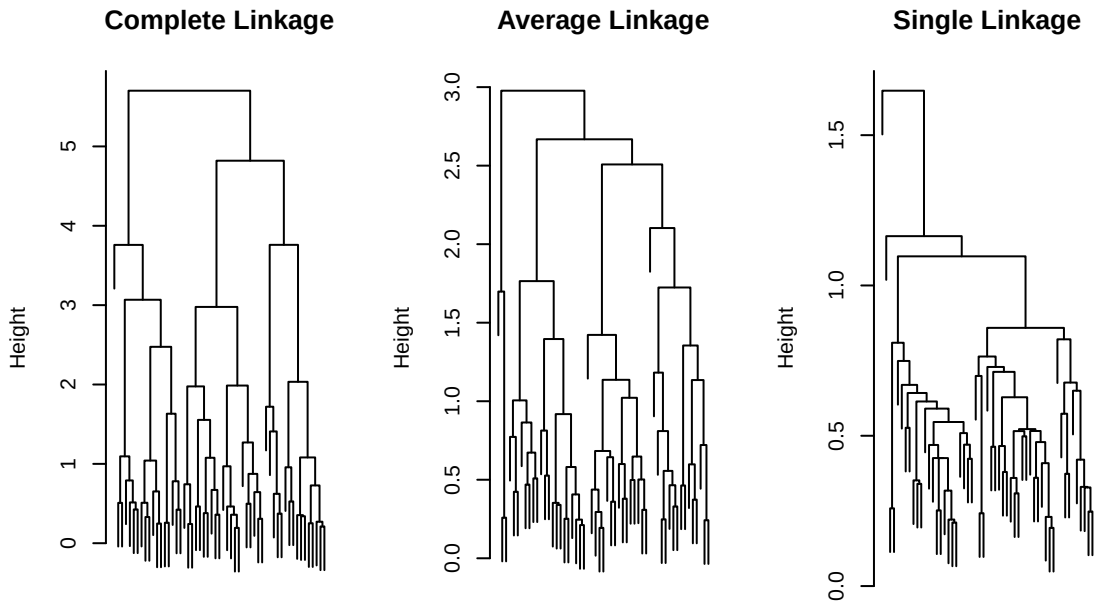
3 Clustering

Clustering with K = 4 using the first 3 PC's and data without outliers.

3.1 Hierarchical clustering

```
sd.data <- scale(lung_df2[,1:3])
hc.complete <- hclust(dist(sd.data), method = "complete")
hc.average <- hclust(dist(sd.data), method = "average")
hc.single <- hclust(dist(sd.data), method = "single")
```

```
par(mfrow = c(1,3))
plot(hc.complete, main = "Complete Linkage", xlab = "", sub = "", cex = .9, labels = F)
plot(hc.average, main = "Average Linkage", xlab = "", sub = "", cex = .9, labels = F)
plot(hc.single, main = "Single Linkage", xlab = "", sub = "", cex = .9, labels = F)
```




```
cutree(hc.single, 4)
```

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## V21 V22 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41
##   2   2   2   3   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56
##   2   2   2   2   2   2   2   2   2   2   2   1   2   1   4
```

Only one sample is classified as group 3 and 4, so single linkage does not suit the data well.

```
table(cutree(hc.complete, 4), lung_df2$Label)
```

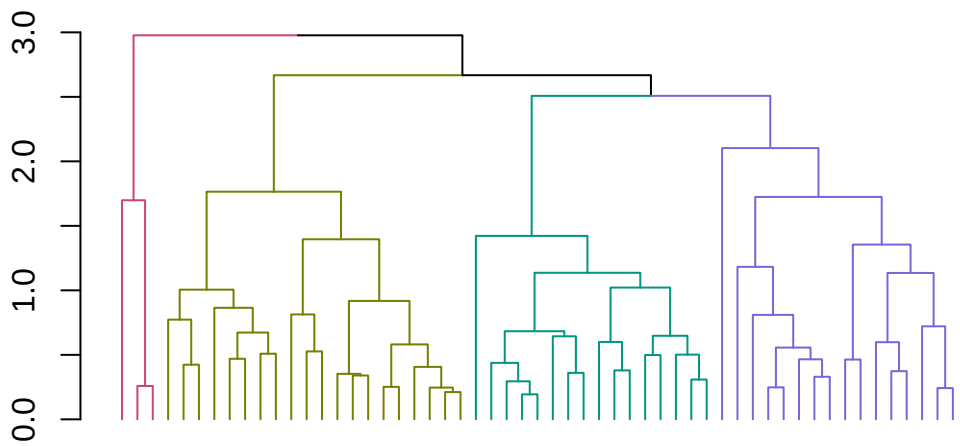
```
##
##     Carcinoid Colon Normal SmallCell
##   1         9     6      0         3
##   2        11     0      0         0
##   3         0     1     17         3
##   4         0     5      0         0
```

```
table(cutree(hc.average, 4), lung_df2$Label)
```

```
##
##     Carcinoid Colon Normal SmallCell
##   1        20     0      0         0
##   2         0    12      2         2
##   3         0     0     15         1
##   4         0     0      0         3
```

```
avg_dend_obj <- as.dendrogram(hc.average)
avg_col_dend <- color_branches(avg_dend_obj, k = 4)
labels_colors(avg_col_dend) <- "white"
```

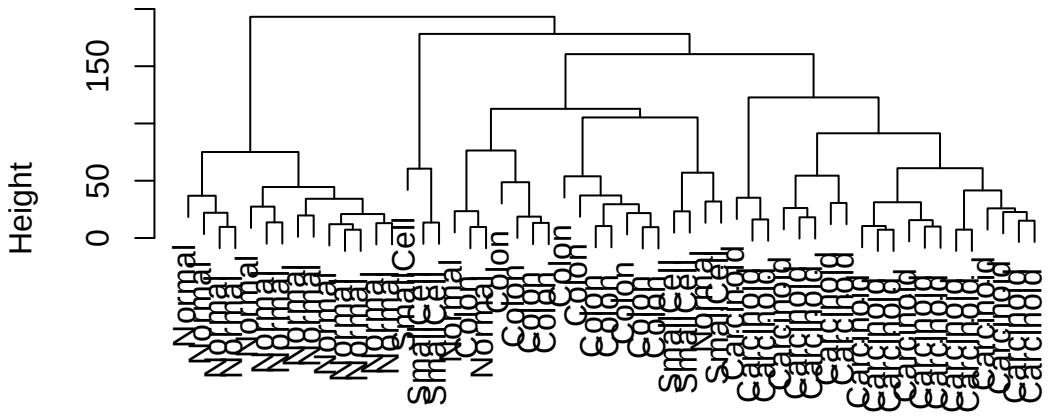
```r
plot(avg_col_dend)
```



3.2 K-means

```r
set.seed(2)
km.out <- kmeans(sd.data, 4, nstart = 20)
km.clusters <- km.out$cluster
```

```r
hc.out <- hclust(dist(pca_res1$x[, 1:3]))
plot(hc.out, labels = lung_df2$Label, main = "Hier. Clust. on First Three Score Vectors")
```



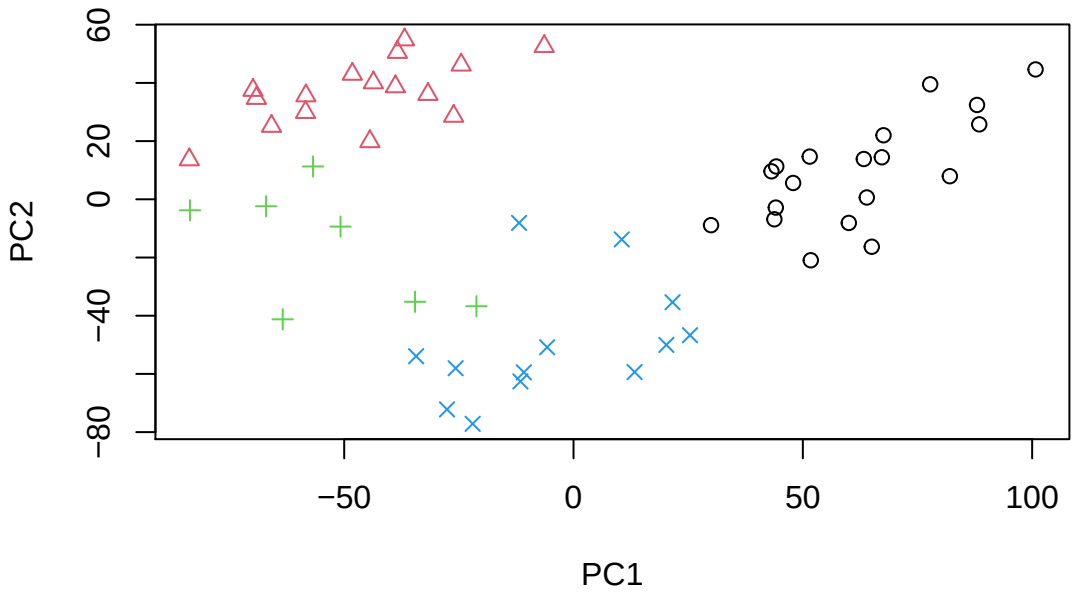

```
hc.clusters <- cutree(hc.out, 4)
table(hc.clusters, lung_df2$Label)
```

```
##
## hc.clusters Carcinoid Colon Normal SmallCell
##           1        20     0      0         0
##           2         0    12      3         3
##           3         0     0     14         0
##           4         0     0      0         3
```

```
set.seed(2)
km.out <- kmeans(pca_res1$x[, 1:3], 4, nstart = 20)
km.clusters <- km.out$cluster
table(km.clusters, lung_df2$Label)
```

```
##
## km.clusters Carcinoid Colon Normal SmallCell
##           1         0     1     17         1
##           2        16     0      0         0
##           3         4     0      0         3
##           4         0    11      0         2
```

```
plot(pca_res1$x[, 1:3], col = km.out$cluster, pch = km.out$cluster)
```




```
table(km.clusters, hc.clusters)
```

```
##            hc.clusters
## km.clusters  1  2  3  4
##           1  0  5 14  0
##           2 16  0  0  0
##           3  4  0  0  3
##           4  0 13  0  0
```

Cluster 3 and 4 in hierarchical clustering are identical to Cluster 1 and 3 in K-means clustering. However, Cluster 1 in hierarchical clustering distributes into Cluster 2 and 3 in K-means clustering. Cluster 2 in hierarchical clustering distributes into Cluster 1 and 4 in K-means clustering.