Bruce Campbell ST-617 Homework 5

Thu Jul 28 12:21:19 2016

```
rm(list = ls())
set.seed(7)
setwd("C:/st-617/")
```

Chapter 10

Problem 11

On the book website, www.StatLearning.com, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

a)

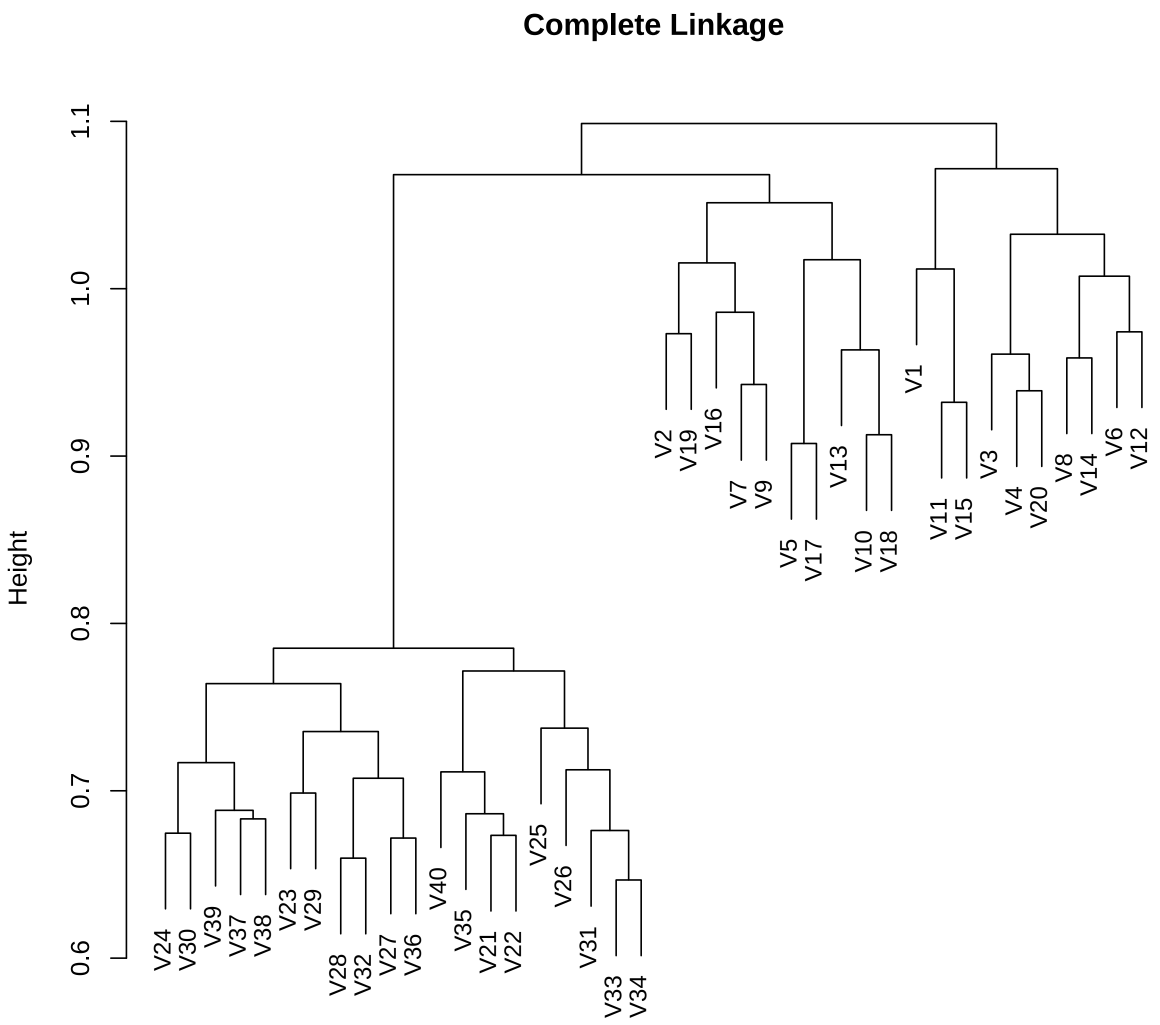
“Load in the data using read.csv(). You will need to select header=F“‘

```
DF <- read.csv("Ch10Ex11.csv", header = FALSE)
DF <- data.frame(t(DF))
```

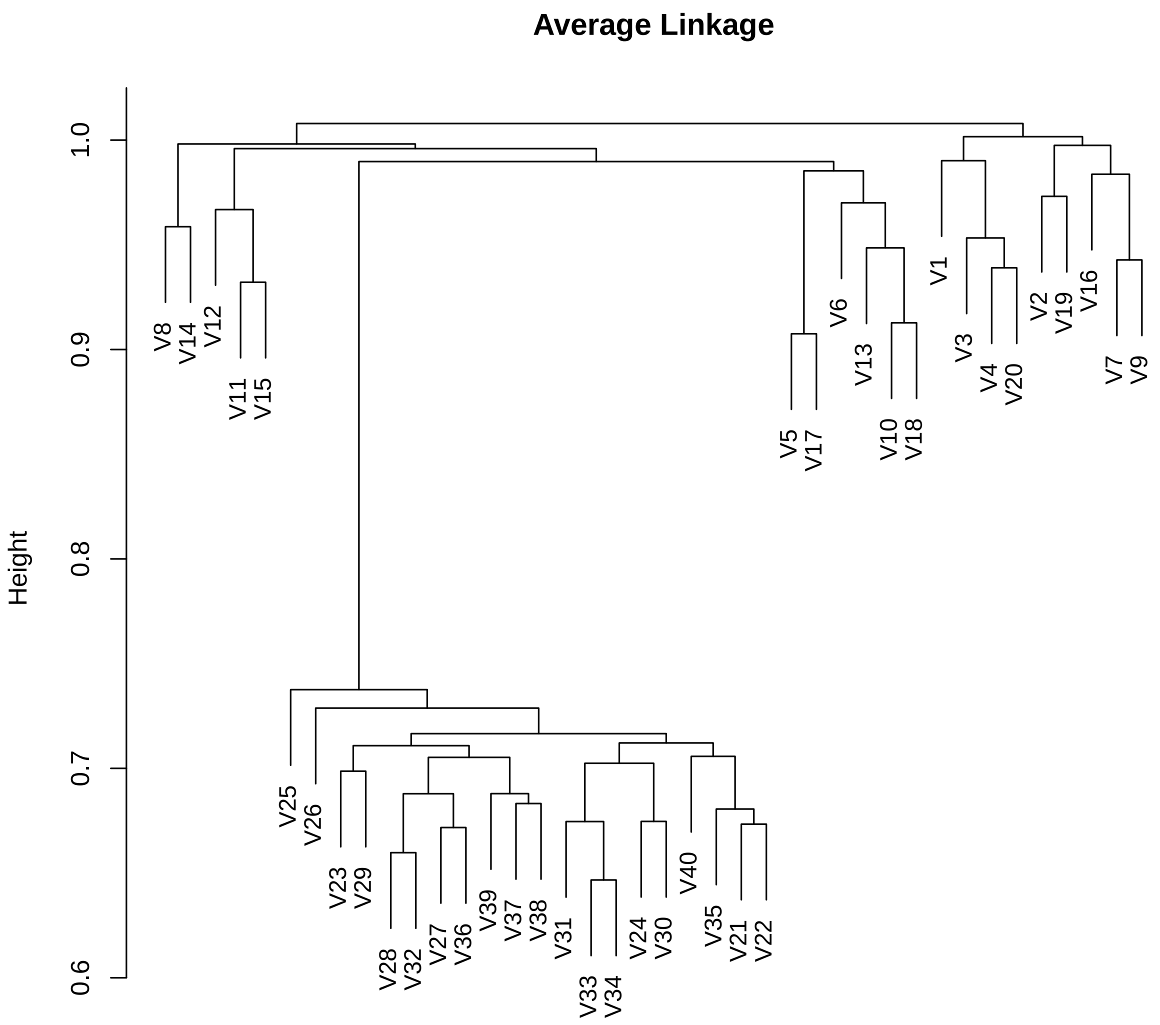
b)

Apply hierarchical clustering to the samples using correlation based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

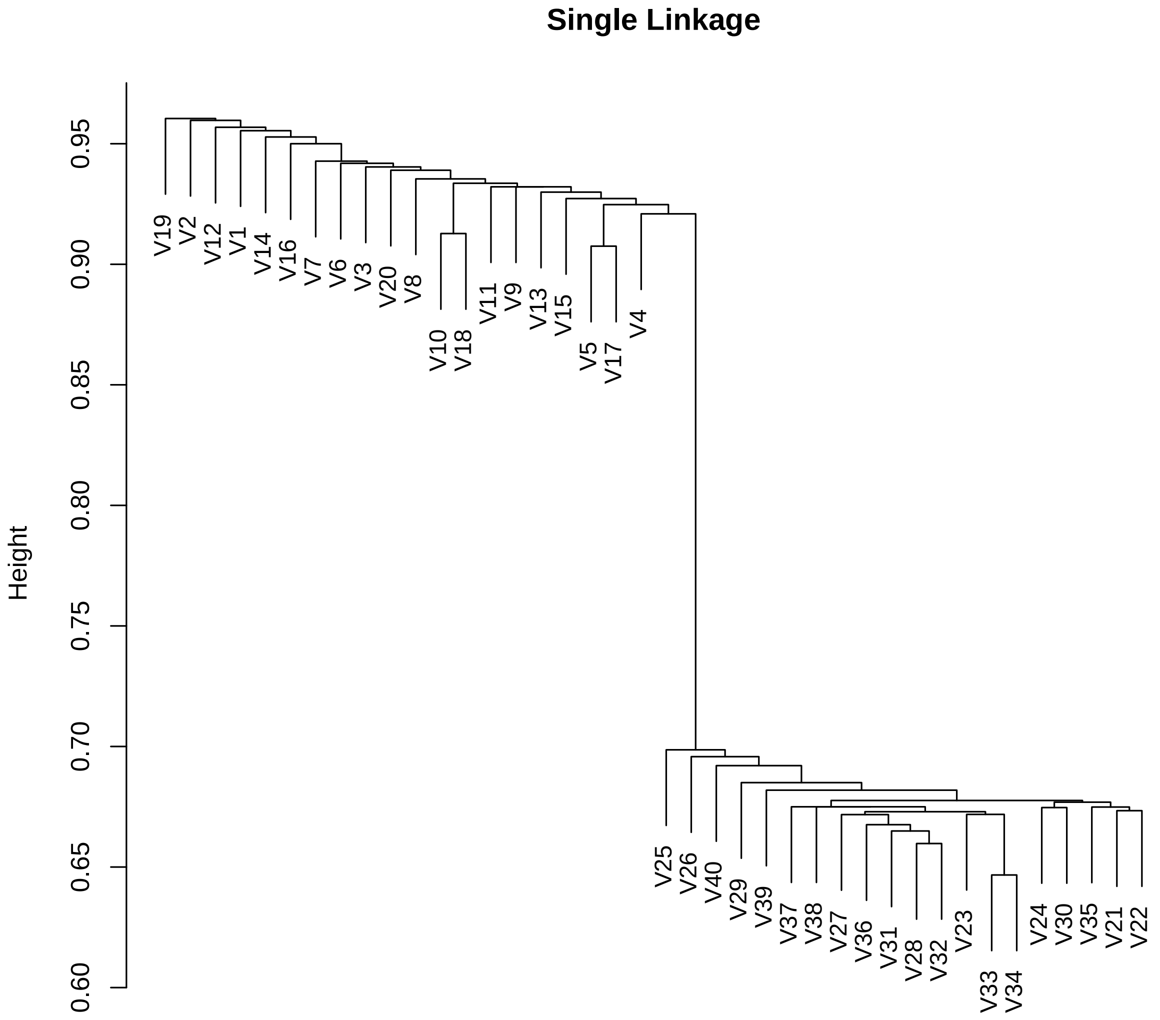
```
dd <- as.dist(1 - cor(t(DF)))

hc.complete = hclust(dd, method = "complete")
plot(hc.complete, main = " Complete Linkage ", xlab = "", sub = "", cex = 0.9)
```



```r
hc.average = hclust(dd, method = "average")
plot(hc.average, main = " Average Linkage ", xlab = "", sub = "", cex = 0.9)
```



```
hc.single = hclust(dd, method = "single")
plot(hc.single, main = " Single Linkage ", xlab = "", sub = "", cex = 0.9)
```



We have to look carefully but we can see that the genes generally separate into two groups. The quality of the cut does depend on the linkage used with complete pr0viding the best spearation.

c)

`Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.`

We would cut the tree at a level that separated the groups. And look at the variablility of the genese between groups.