HW4 -Stefan Zdraljevic

```
library(scales)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

load yeast cell cycle data

```
yeast <- read.table("~/Dropbox/AndersenLab/LabFolders/Stefan/Courses/Stats_for_Bioinformati
cs/Homework/HW4/yst_cell_cycle.txt",
                    fill=T,header=T, sep="\t")
colnames(yeast) <- c("gene",colnames(yeast[,2:83]))
```

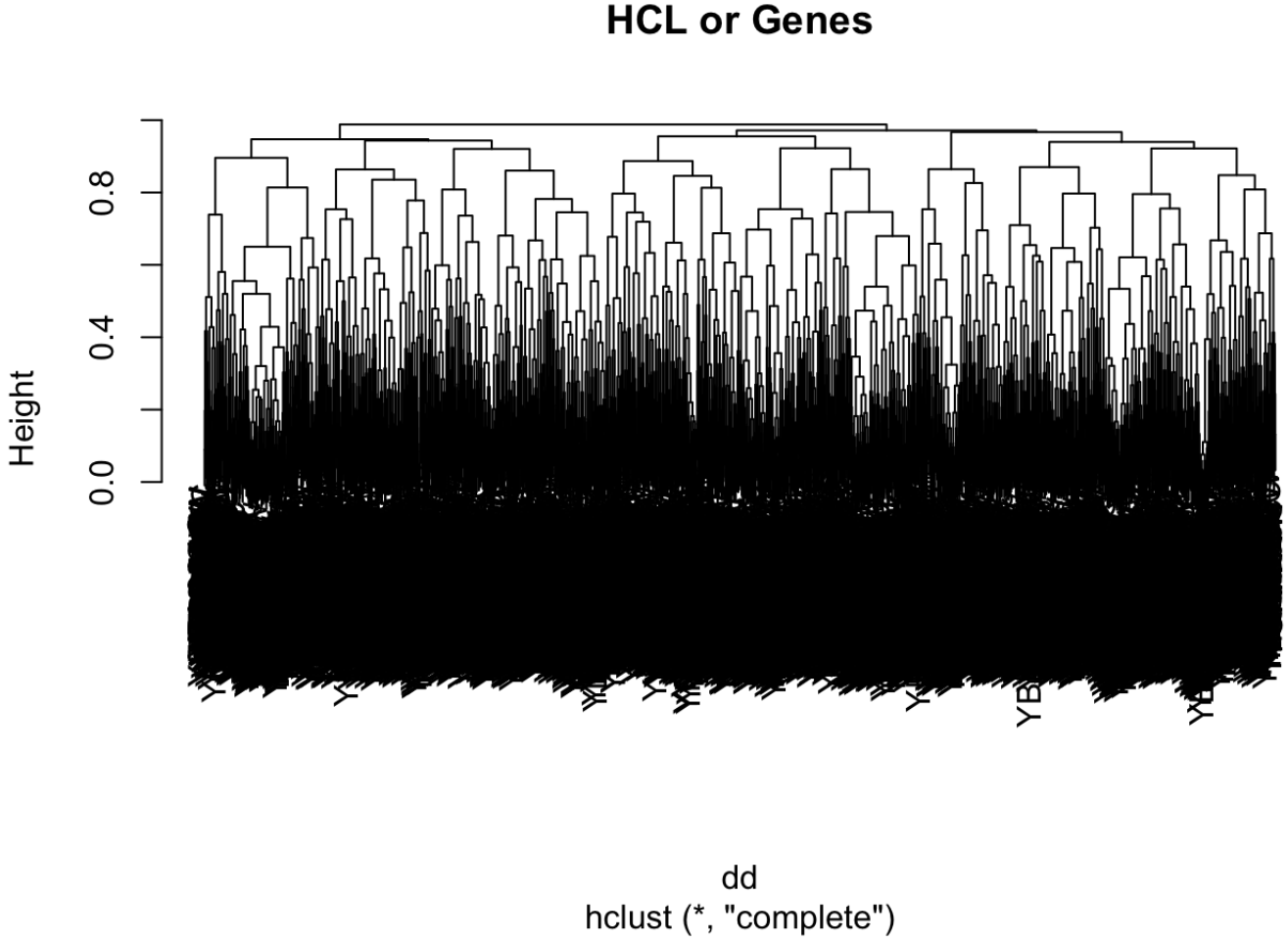
subset data to only contain cdc28 arrested experiments and remove NA genes

remove cdc28_140 timepoint because it eliminates a lot of data

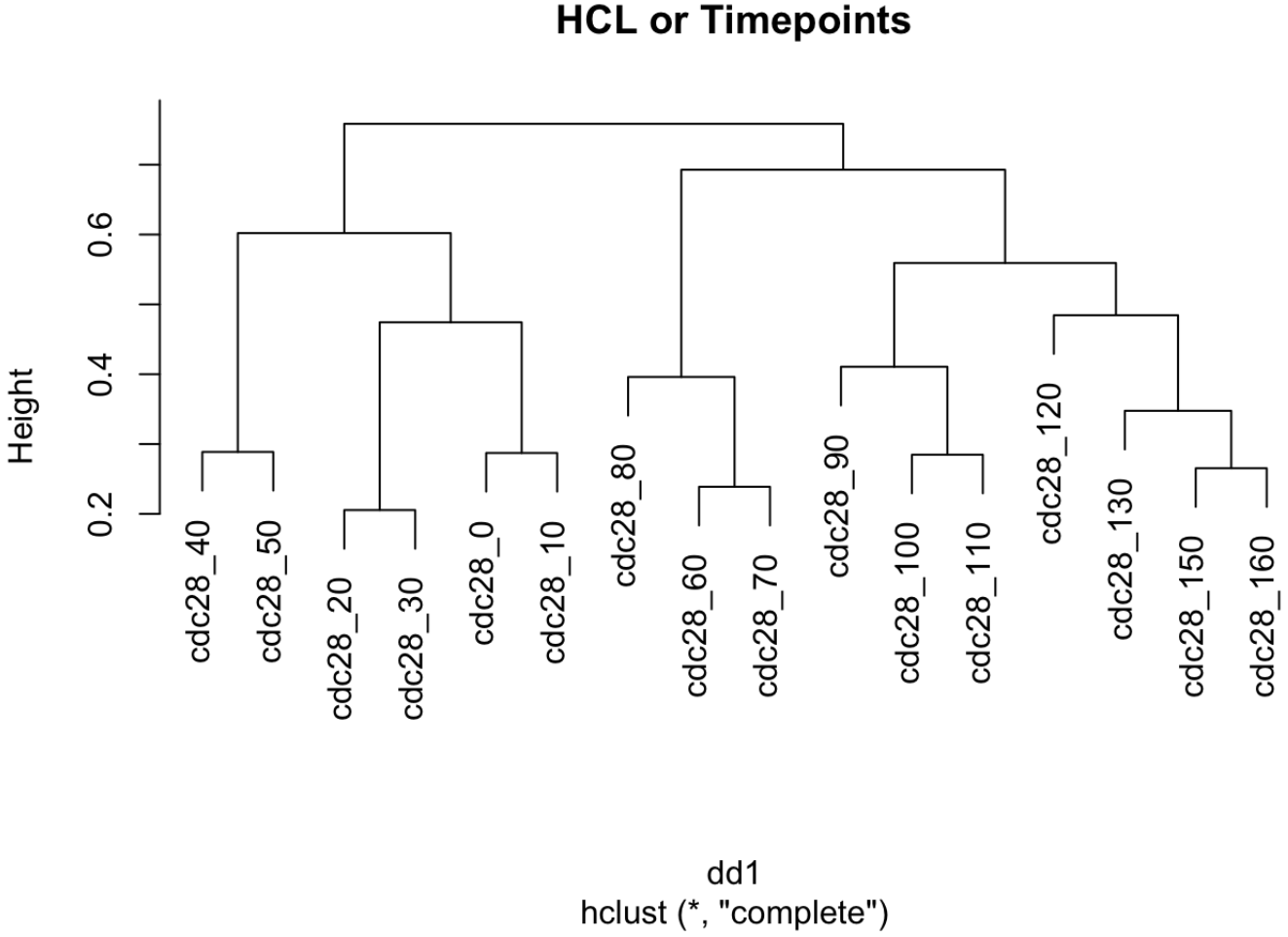
cluster based on correlation distance

Little is gained from the gene clustering, but you can see that that certain timepoints cluster together due to proximity

```
dd <- as.dist((1 - cor(t(cdc28),use = "pairwise.complete.obs"))/2)
hc=hclust(dd)
plot(hc, main = "HCL or Genes")
```

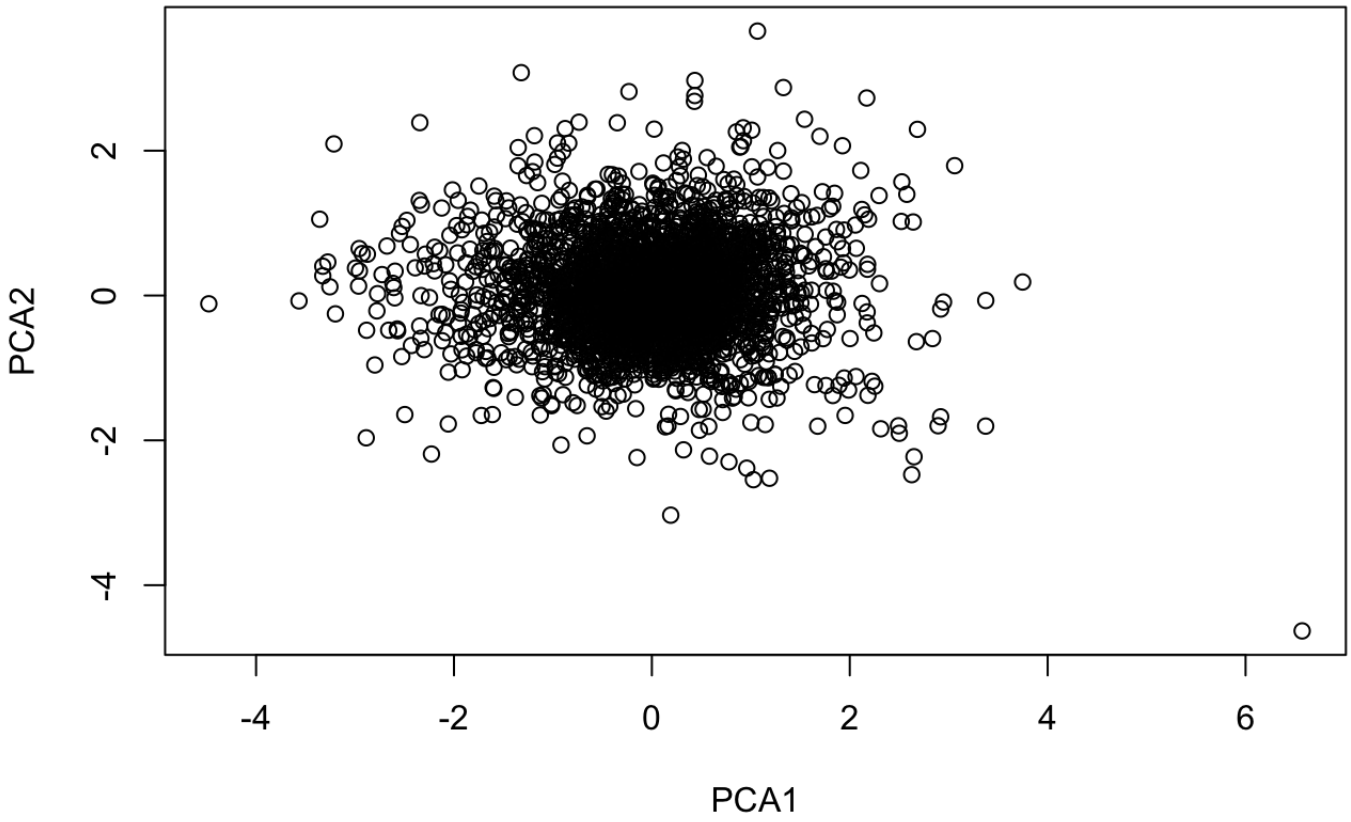



```
dd1 <- as.dist((1 - cor(cdc28,use = "pairwise.complete.obs"))/2)
hc1=hclust(dd1)
plot(hc1,main = "HCL or Timepoints")
```

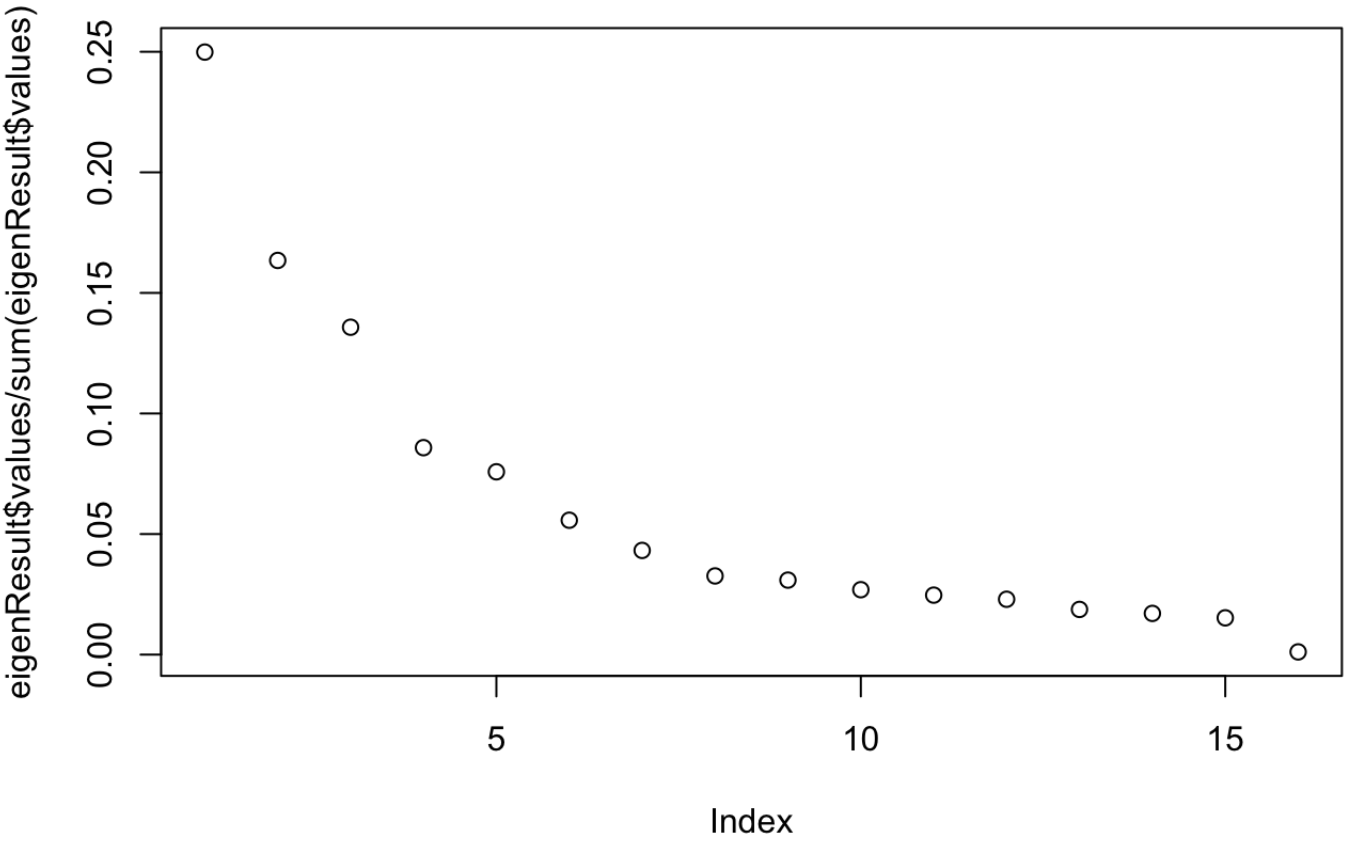


principal component analysis and k-means clustering

```
##############
##principle component analysis
eigenResult <- eigen(var(cdc28))     ## Eigen analysis
eigenVector <- eigenResult$vectors   ##extract Eigen vectors
p1 <- eigenVector[,1]                ## first eigen vector
p2 <- eigenVector[,2]                ## second eigen vector

cdc28m <- as.matrix(cdc28)
PCA1 <- cdc28m%*%p1                  ## convert to principal component 1
PCA2 <- cdc28m%*%p2                 ## convert to principal component 2
plot(PCA1,PCA2)
```




```
plot(eigenResult$values/sum(eigenResult$values))
```

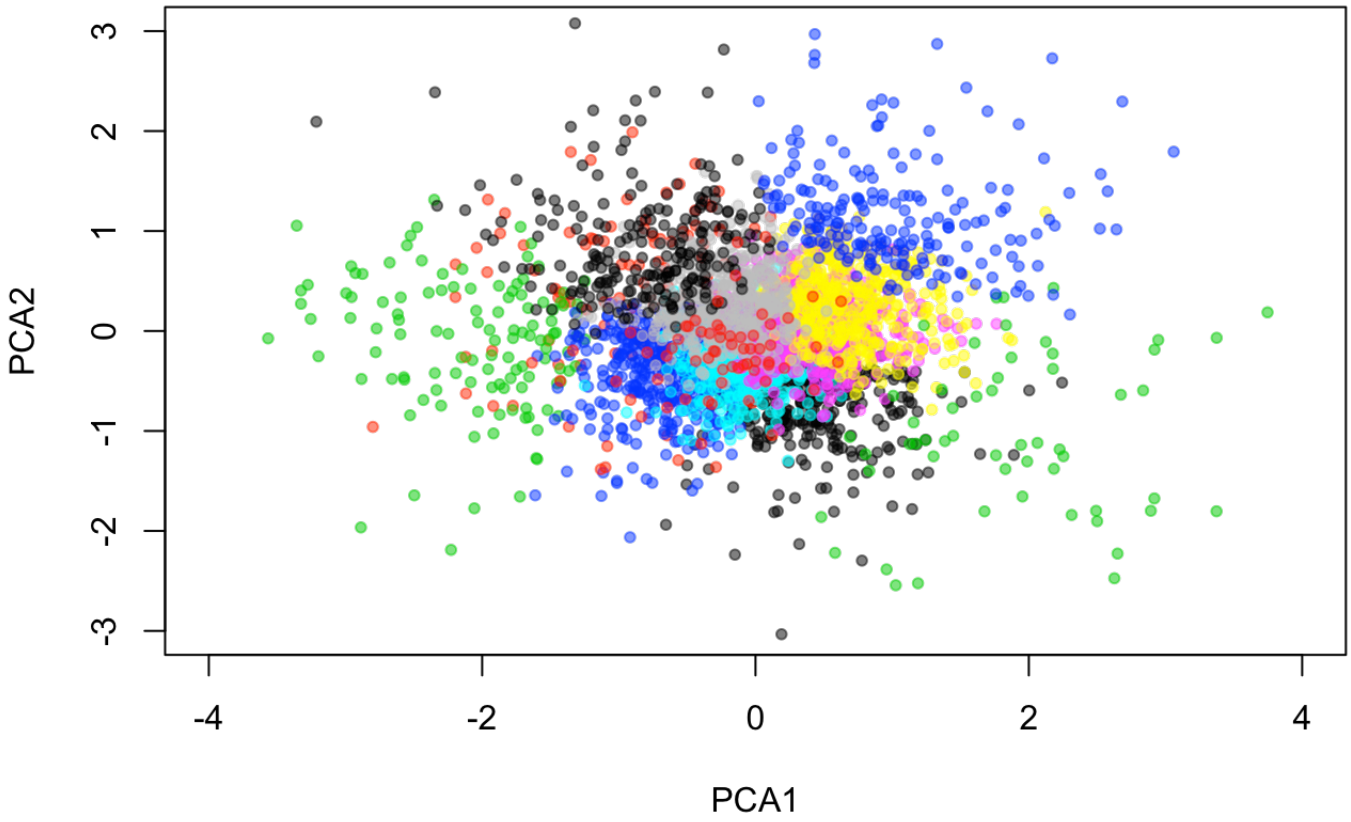



```
View(eigenResult$values/sum(eigenResult$values))

varcomp <- data.frame(varexp = eigenResult$values/sum(eigenResult$values))
varcomp$expl <- cumsum(varcomp$varexp)
print(varcomp)
```

```
##      varexp   expl
## 1  0.249864 0.2499
## 2  0.163466 0.4133
## 3  0.135771 0.5491
## 4  0.085823 0.6349
## 5  0.075816 0.7107
## 6  0.055735 0.7665
## 7  0.043212 0.8097
## 8  0.032635 0.8423
## 9  0.030899 0.8732
## 10 0.026922 0.9001
## 11 0.024671 0.9248
## 12 0.022963 0.9478
## 13 0.018745 0.9665
## 14 0.017096 0.9836
## 15 0.015255 0.9989
## 16 0.001128 1.0000
```

NOTE ABOVE 12 principal components are required to capture 95% variance in the data.

```
KmsYeast <- kmeans(cdc28,12)
##
Yeast2 <- cbind(cdc28,PCA1,PCA2,KmsYeast$cluster)
plot(PCA1,PCA2,type="n", ylim = c(-3,3), xlim = c(-4,4))

for (i in 1:12){
  points(PCA1[Yeast2[,19]==i],PCA2[Yeast2[,19]==i],col=alpha(i,.5),pch=20)
}
```



observe startification of k-means clusters on principal components space.