Assignment 2 - Part 2B

Clustering

Fernando Cagua

```
library(ISLR)
nci.data <- t(NCI60$data)
nci_pr <- princomp(scale(nci.data))
```

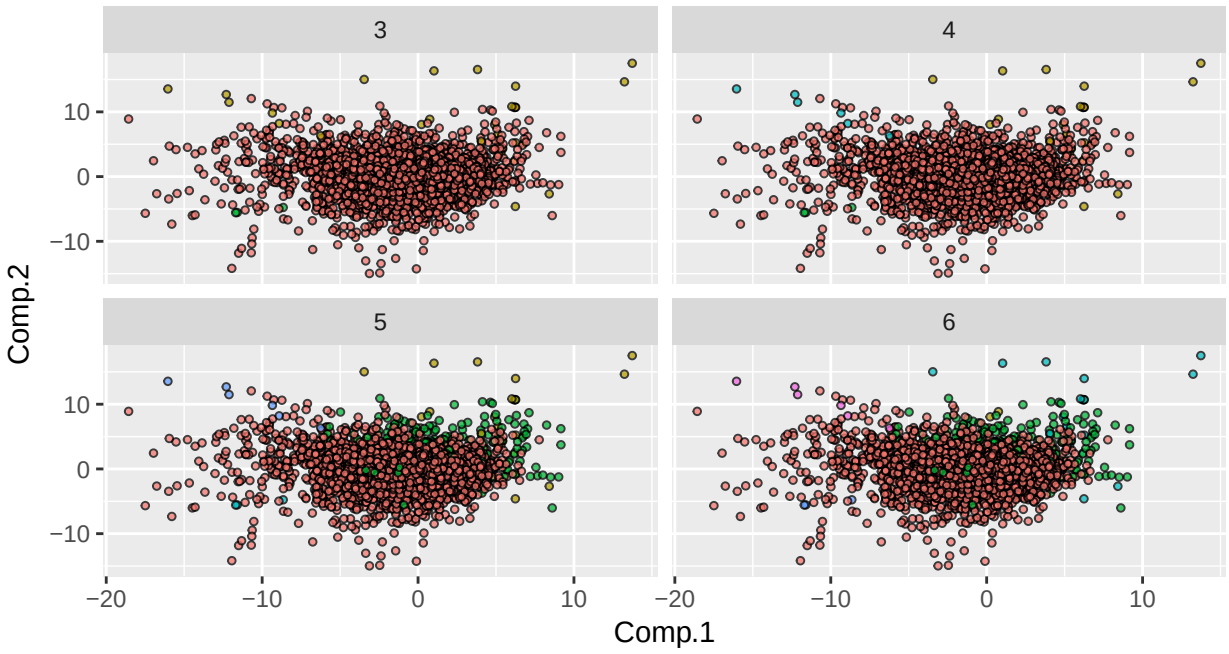
1 Hierarchical clustering

```
pr_data <- as.data.frame(predict(nci_pr))
e_clust <- pr_data %>% dist() %>% hclust()
# get clusters
euclidean <- 3:6 %>%
  plyr::ldply(function(x){
    cl <- e_clust %>% cutree(x)
    pr_data %>% mutate(k = x, cluster = cl)
  })
# table of membership
euclidean %>% group_by(k, cluster) %>% summarise(n = n()) %>%
  reshape2::dcast(k~cluster) %>% knitr::kable()
```

k	1	2	3	4	5	6
3	6792	35	3	NA	NA	NA
4	6792	29	3	6	NA	NA
5	6559	29	233	3	6	NA
6	6559	7	233	22	3	6

The table shows the number of genes in each of the clusters (columns) for different number of clusters (rows). Based on this table and the plot of the clusters below, it is possible to infer that there is a very large cluster that contains most of the genes and some smaller ones that might be very particular to specific cell lines. It is also possible to observe that indeed the clusters are nested. For exampl cluster 2 and 4 in k = 6 are contained by cluster 2 in k = 5. Or cluster 1 and 3 in k = 5 are contained by cluster 1 in k = 4.

```
# plot clusterings
euclidean %>%
  ggplot(aes(x = Comp.1, y = Comp.2)) +
  geom_point(aes(fill = as.factor(cluster)), shape = 21, size = 1, alpha = 0.75,
             show.legend = FALSE) +
  facet_wrap(~k)
```

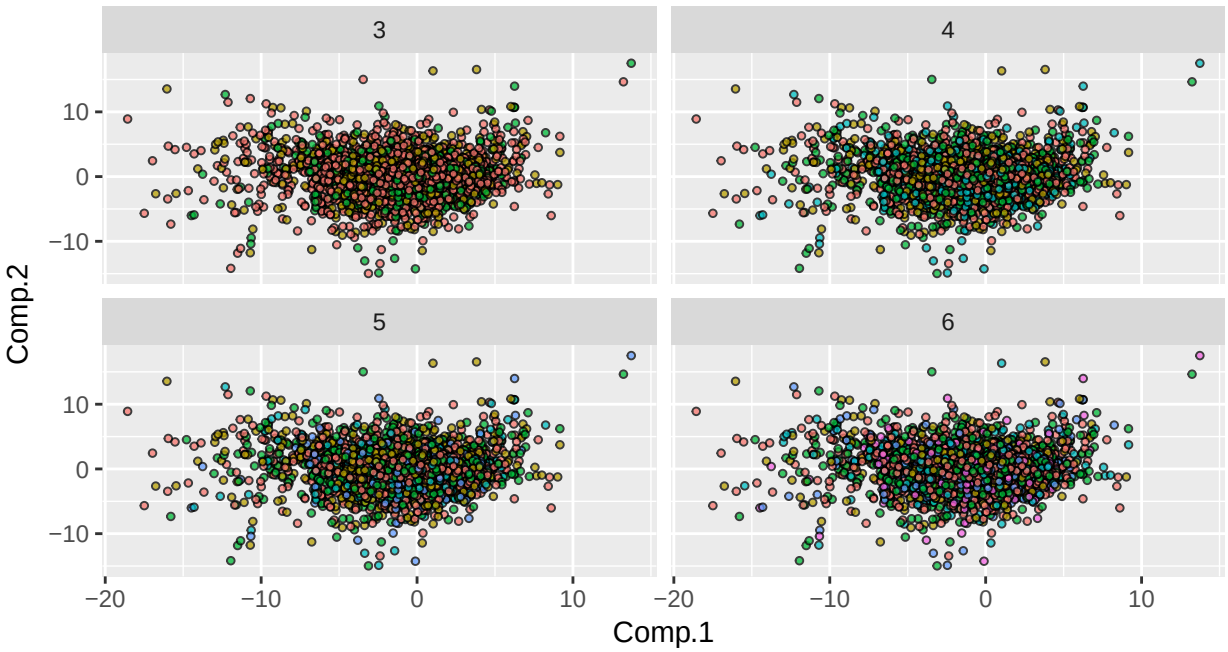


2 Correlation based distance

```
c_clust <- nci.data %>% scale() %>% t() %>% cor() %>% as.dist() %>% hclust()
correlation <- 3:6 %>%
  plyr::ldply(function(x){
    cl <- c_clust %>% cutree(x)
    pr_data %>% mutate(k = x, cluster = cl)
  })
# table of membership
correlation %>% group_by(k, cluster) %>% summarise(n = n()) %>%
  reshape2::dcast(k~cluster) %>% knitr::kable()
```

k	1	2	3	4	5	6
3	3855	1665	1310	NA	NA	NA
4	2215	1665	1640	1310	NA	NA
5	2215	1665	1640	763	547	NA
6	2215	920	1640	745	763	547

```
# plot clusterings
correlation %>%
  ggplot(aes(x = Comp.1, y = Comp.2)) +
  geom_point(aes(fill = as.factor(cluster)), shape = 21, size = 1, alpha = 0.75,
             show.legend = FALSE) +
  facet_wrap(~k)
```



The clusters found using the correlation-based distance are completelly different to those found using the ecuclidean distance. Not only the number of genes in each cluster is a bit more balanced but it allows to differentiate genes within the large blob of points in the PC1-2 space.

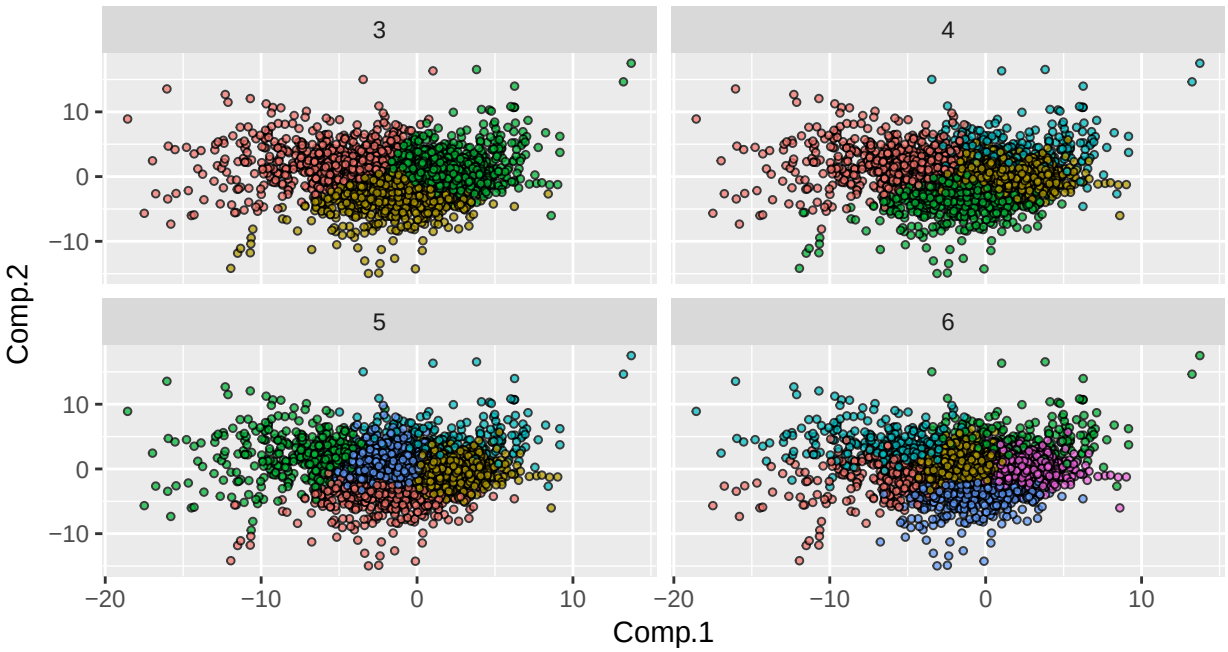
3 K-means

I'ts not completely clear from the assignment which matrix should be used as the input for `kmeans`. Here I assume that the matrix is the scaled gene x cell-line matrix.

```
kmeans_raw <- 3:6 %>%
  plyr::ldply(function(x){
    cl <- nci.data %>% scale() %>% kmeans(x)
    pr_data %>% mutate(k = x, cluster = cl$cluster)
  })
# table of membership
kmeans_raw %>% group_by(k, cluster) %>% summarise(n = n()) %>%
  reshape2::dcast(k~cluster) %>% knitr::kable()
```

k	1	2	3	4	5	6
3	750	1523	4557	NA	NA	NA
4	718	4749	1033	330	NA	NA
5	542	3787	391	277	1833	NA
6	346	2008	278	265	595	3338

```
# plot clusterings
kmeans_raw %>%
  ggplot(aes(x = Comp.1, y = Comp.2)) +
  geom_point(aes(fill = as.factor(cluster)), shape = 21, size = 1, alpha = 0.75,
             show.legend = FALSE) +
  facet_wrap(~k)
```



The k-means clustering is different to both the hierarchical clustering using euclidean distances on the principal components or correlation distances. However it resembles more the one that uses euclidean distances in the sense that the clusters are distinguishable in the PC space.