Clustering: Hierarchical Clustering

Algorithm 10.2

1. Begin with n observations and a dissimilarity matrix. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

    - Examine all pairwise inter-cluster dissimilarities and identify the pair that are most similar. Fuse these clusters at the dendrogram height corresponding to their dissimilarity.
    - Compute pairwise dissimilarities among remaining $i-1$ clusters.

Linkage types

Complete: Maximal intercluster dissimilarity.

Single: Minimal intercluster dissimilarity.

Average: Mean intercluster dissimiliarity.

Centroid: Dissimilarity between cluster centroids.

Example: Cancer Genomics

How do you characterize the gene expression in cancerous cells?

`NCI60`: Cancer cell mircoarray data with 6,830 expression measurements on 64 cancer cell lines.

```r
library(ISLR)
NCI <- NCI60$data
dim(NCI)
```

```
## [1]   64 6830
```

```r
NCI[1:3, 1:5]
```

```
##            1          2          3          4          5
## V1 0.300000  1.180000  0.550000  1.140000 -0.265000
## V2 0.679961  1.289961  0.169961  0.379961  0.464961
## V3 0.940000 -0.040000 -0.170000 -0.040000 -0.605000
```

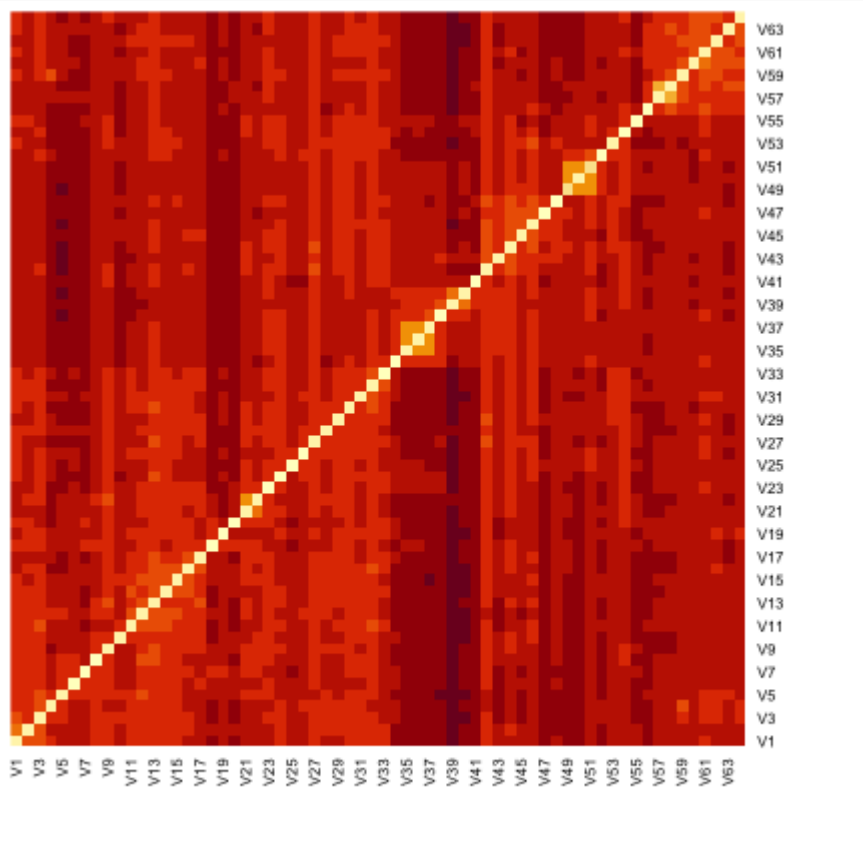
```r
NCI_scaled <- scale(NCI)
NCI_dist <- dist(NCI_scaled)
dissmatrix <- as.matrix(NCI_dist)
dissmatrix[1:5, 1:5]
```

```
##            V1        V2        V3        V4       V5
## V1   0.00000  77.04594  87.30561 103.18322 113.7230
## V2  77.04594   0.00000  88.89531 106.64318 116.1610
## V3  87.30561  88.89531   0.00000  95.79984 101.0443
## V4 103.18322 106.64318  95.79984   0.00000 107.0625
## V5 113.72295 116.16097 101.04429 107.06253   0.0000
```

This distance matrix can be visualized as a *heatmap*.

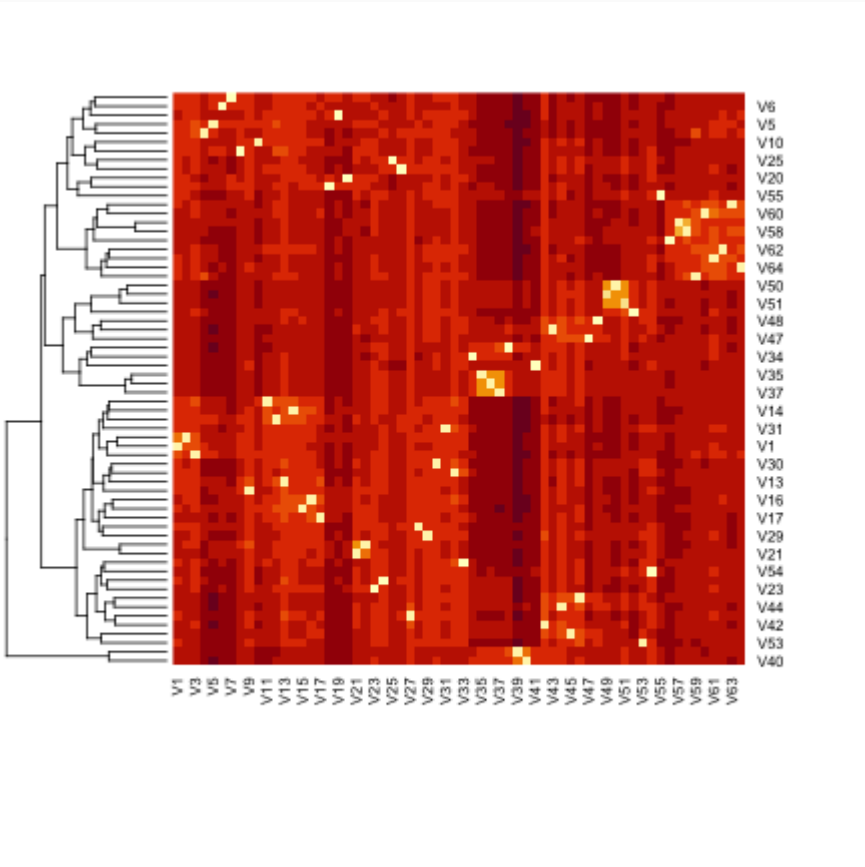
Heatmap

```
heatmap(dissmatrix, Rowv = NA, Colv = NA)
```

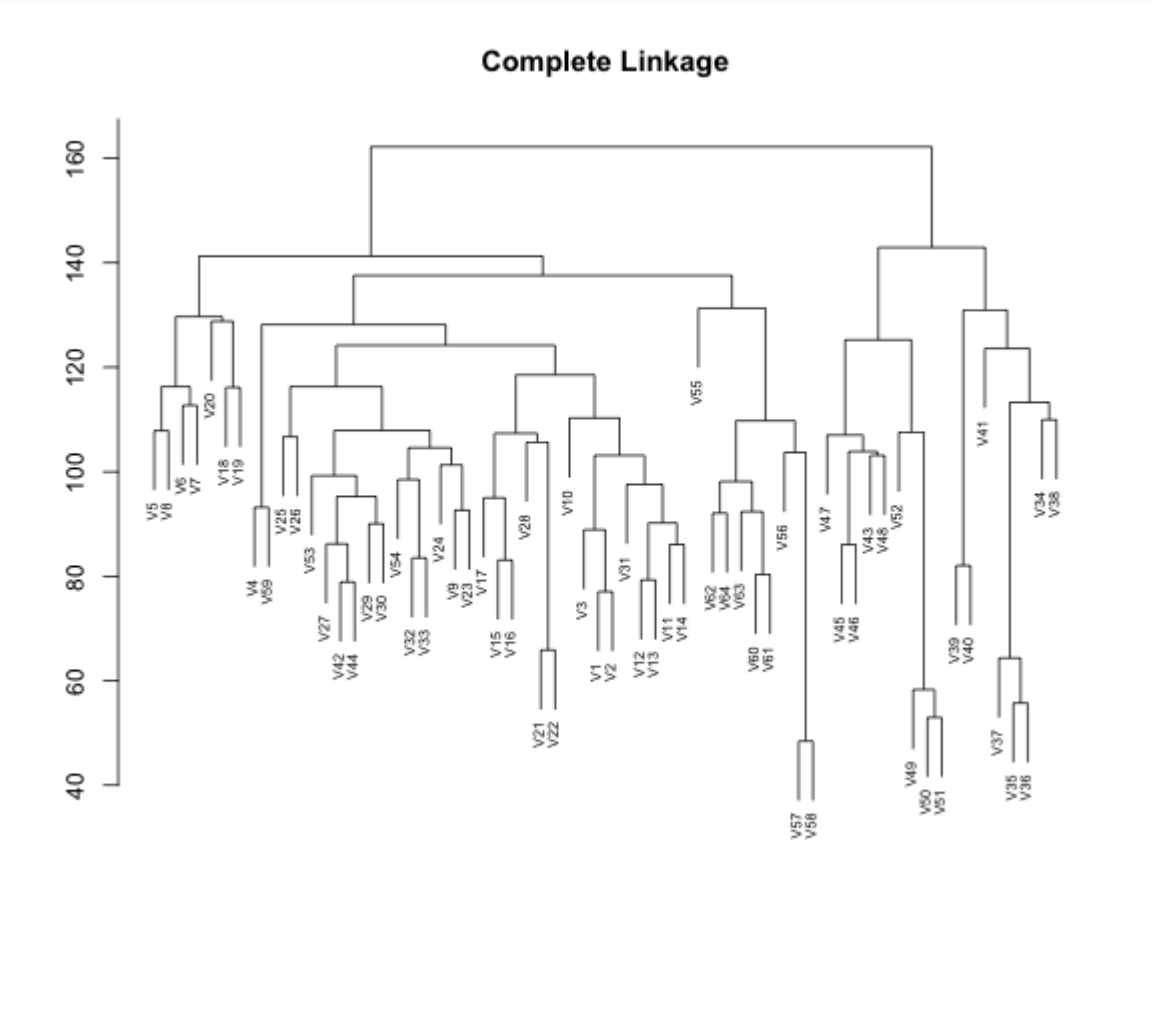


Reordered Heatmap

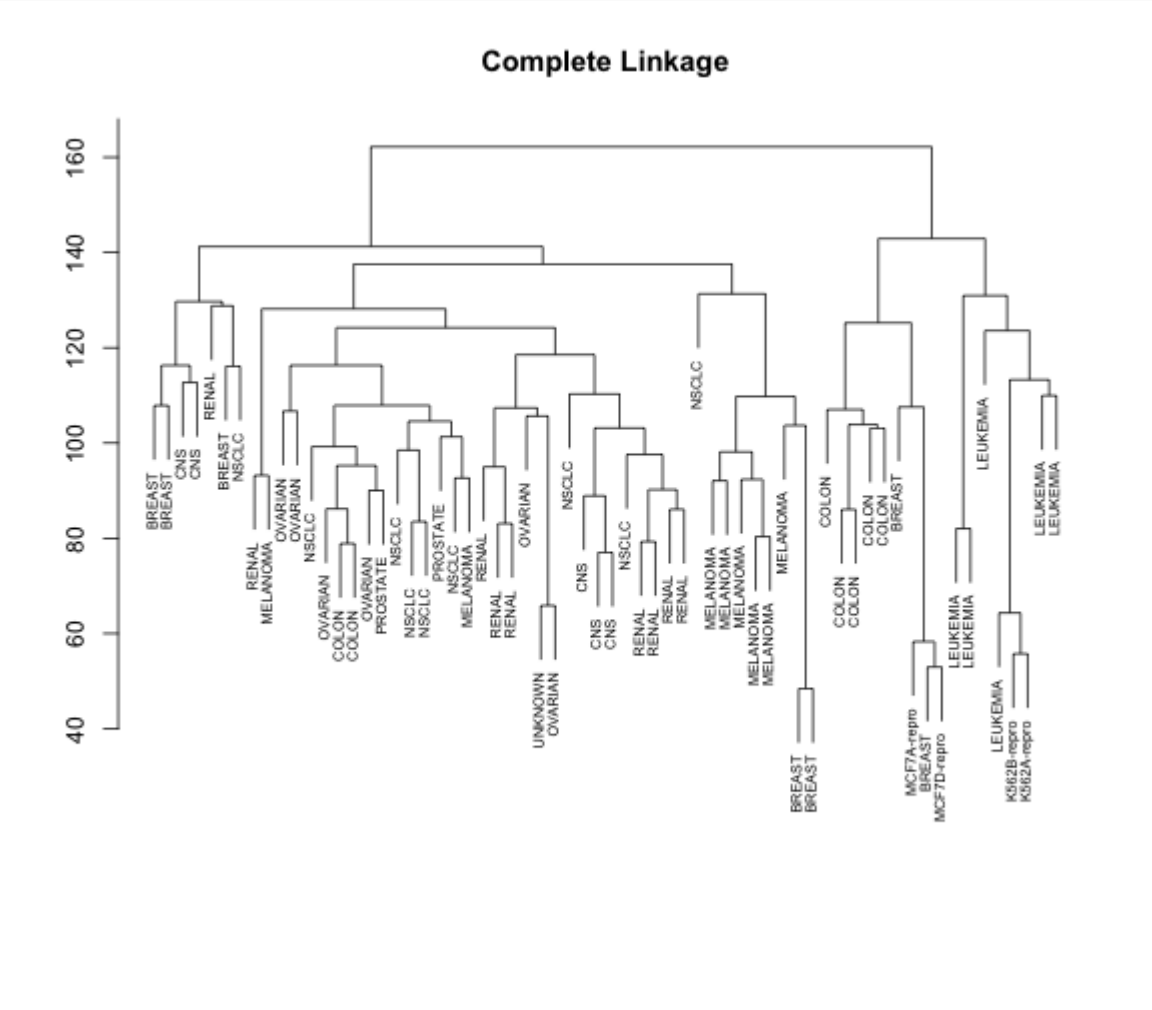
```
heatmap(dissmatrix, Colv = NA)
```



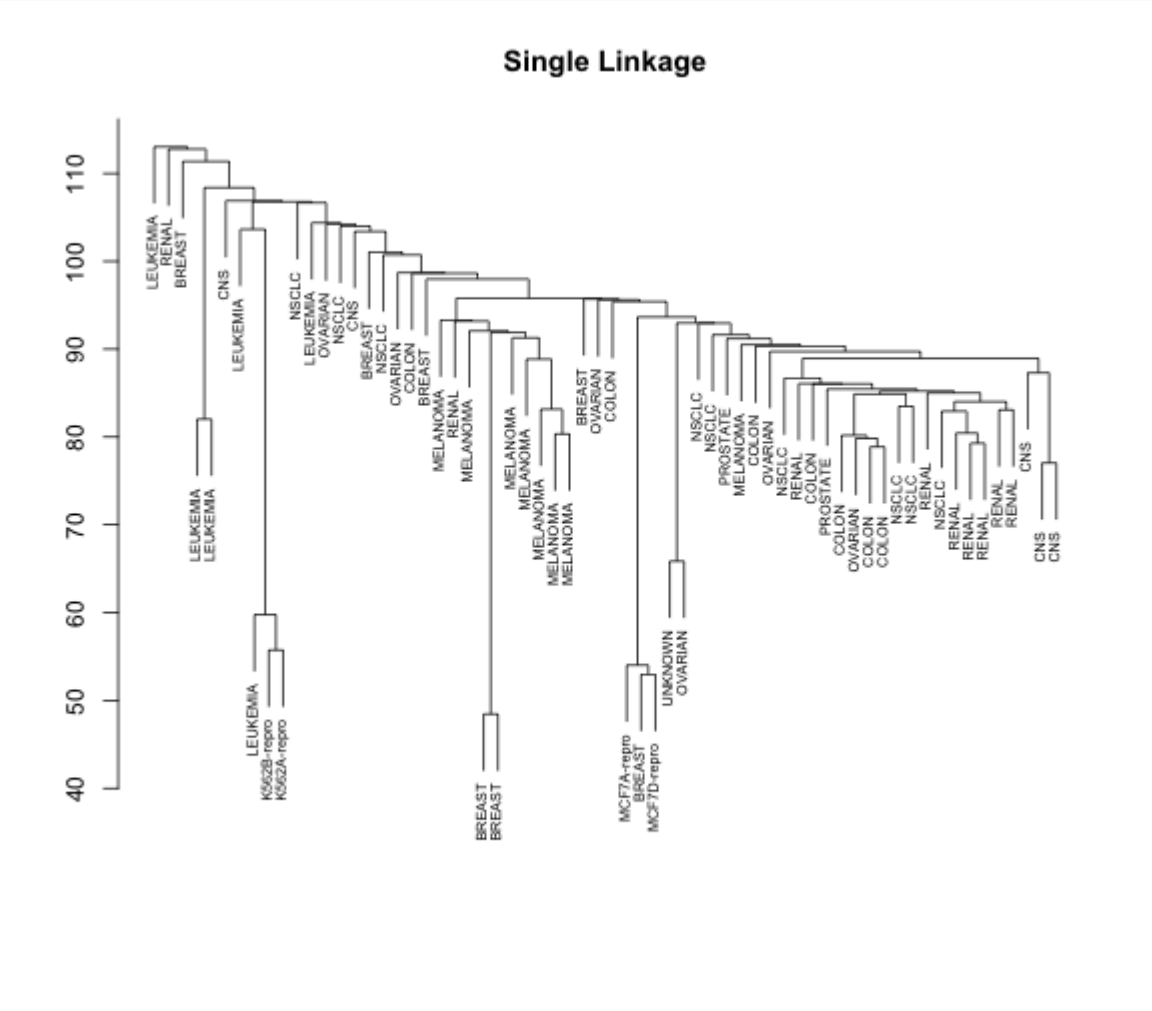
Dendrogram (complete linkage)



Dendrogram (complete linkage)



Dendrogram (single linkage)



Cutting the Dendro

```
hc <- hclust(NCI_dist)
hc_cut <- cutree(hc, k = 4)
length(hc_cut)
```

```
## [1] 64
```

```
head(hc_cut)
```

```
## V1 V2 V3 V4 V5 V6
##  1  1  1  1  2  2
```

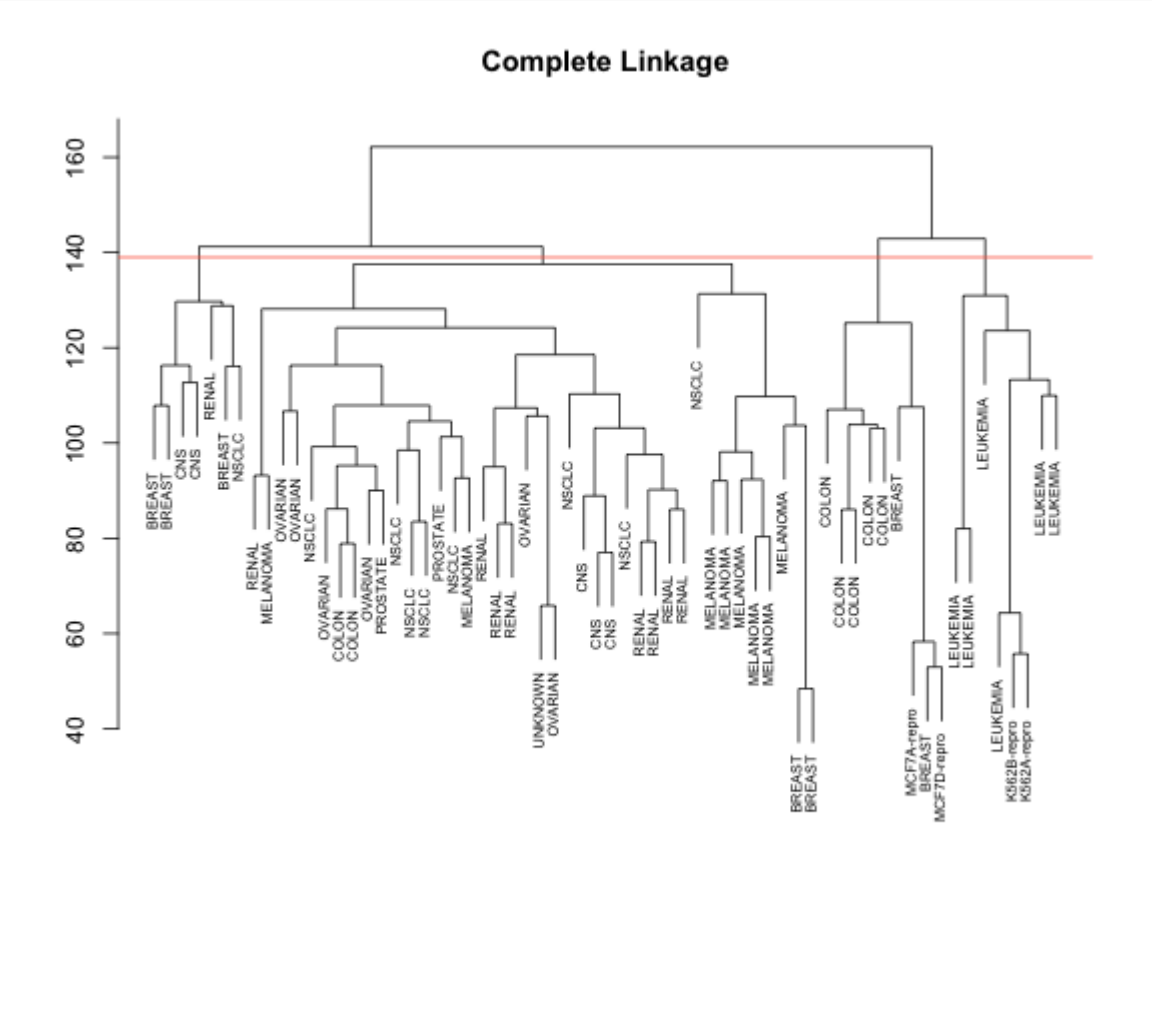
Cutting the Dendro

```
table(hc_cut, NCI60$labs)
```

```
##
## hc_cut BREAST CNS COLON K562A-repro K562B-repro LEUKEM
##      1      2   3     2           0           0
##      2      3   2     0           0           0
##      3      0   0     0           1           1
##      4      2   0     5           0           0
##
## hc_cut MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE REN
##      1           0        8     8       6        2
##      2           0        0     1       0        0
##      3           0        0     0       0        0
##      4           1        0     0       0        0
```

- Leukemia cell lines in cluster 3.
- Breast cancer cell lines spread across 1, 2, and 4.

Dendrogram


Complete Linkage
160
140
120
100
80
60
40
BREAST
BREAST
CNS
CNS
RENAL
BREAST
NSCLC
RENAL
MELANOMA
OVARIAN
OVARIAN
NSCLC
OVARIAN
COLON
COLON
OVARIAN
PROSTATE
NSCLC
NSCLC
NSCLC
PROSTATE
NSCLC
MELANOMA
RENAL
RENAL
RENAL
OVARIAN
UNKNOWN
OVARIAN
NSCLC
CNS
CNS
NSCLC
RENAL
RENAL
RENAL
RENAL
NSCLC
MELANOMA
MELANOMA
MELANOMA
MELANOMA
MELANOMA
MELANOMA
BREAST
BREAST
COLON
COLON
COLON
COLON
COLON
BREAST
MCF7A-repro
BREAST
MCF7D-repro
LEUKEMIA
LEUKEMIA
LEUKEMIA
LEUKEMIA
LEUKEMIA
K562B-repro
K562A-repro
LEUKEMIA
LEUKEMIA
LEUKEMIA