Cluster Analysis

Lavinia Carabet

Cluster Analysis

Clustering is an unsupervised analysis technique used to group similar objects (genes or samples) together

Builds structure to help explain the relationships that may exist between the objects

Hierarchical Clustering

Provides an informative display of ordered objects

Builts a tree structure dynamically (not model-based) using dissimilarities between objects being clustered

Preparation

Load the fibroEset library and dataset Obtain the classifications for the samples

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("fibroEset")
```

```
library(fibroEset)
data(fibroEset)

fibro.data <- exprs(fibroEset)

dim(fibro.data)
```

```
## [1] 12625    46
```

```
fibro.data[1:5,]
```

```
##              1   2   3    4   5   6   7   8   9  10   11  12  13  14  15  16
## 100_g_at   476 518 686  602 470 355 349 468 368 637  525 723 668 611 711 779
## 1000_at   1795 890 508 1113 708 629 484 795 941 857 1242 594 591 676 672 842
## 1001_at    100 119 100  100 100 115 100 100 100 100  145 100 100 100 108 100
## 1002_f_at  100 101 100  100 100 100 100 134 100 100  100 100 100 100 103 100
## 1003_s_at  100 100 100  100 100 100 100 100 100  90   90 100 100 100 100 100
##             17  18  19  20  21  22  23   24  25  26   27   28  29  30   31   32
## 100_g_at   575 741 560 575 616 585 612  662 484 484  537  553 599 461  489  562
## 1000_at    356 712 468 509 637 564 716 1124 897 897 1160 1154 861 957 1026 1035
## 1001_at   100 132 100 100 142 100 100  133 100 100  100  100 100 100  100  100
## 1002_f_at 100 100 100 100 100 100 100  100 100 100  100  100 100 103  100  100
## 1003_s_at 100 100 100 100 100 100 100  100 100 100  100  134 100 100  100  125
##             33   34   35  36   37  38  39   40  41   42   43   44  45  46
## 100_g_at   549  485  544 525  476 591 658  843 509  613  705  394 564 409
## 1000_at   1140 1233 1065 974 1183 813 919 1760 953 1076 1282 1000 980 828
## 1001_at    100  100  100 100  100 101 100  100 100  100  100  100 100 100
## 1002_f_at  100  100  100 100  100 100 100  100 100  100  100  100 100 100
## 1003_s_at  100  170  100 100  125 116 107  100 141  100  100  100 100 100
```

```
phenoData(fibroEset)$species
```

```
##  [1] b b b b b b b b b b b g g g g g g g g g g g g h h h h h h h h h h h h h h h
## [39] h h h h h h h h
## Levels: b g h
```

Select a random set of 50 genes from the data frame, and subset the data frame

```
rand.genes <- sample(row.names(fibro.data),50,replace=FALSE)
fibro.sample <- as.data.frame(fibro.data[rand.genes,])
dim(fibro.sample)
```

```
## [1] 50 46
```

```
row.names(fibro.sample)
```

```
##  [1] "39265_at"   "38591_at"   "40109_at"   "37111_g_at" "35933_f_at"
##  [6] "37971_at"   "1655_s_at"  "34188_at"   "1075_f_at"  "39134_at"
## [11] "31803_at"   "38491_at"   "33078_at"   "191_at"     "40706_at"
## [16] "35483_at"   "39349_at"   "38500_at"   "36378_at"   "32447_at"
## [21] "32078_at"   "1418_at"    "31521_f_at" "34260_at"   "33326_at"
## [26] "31607_at"   "32822_at"   "38389_at"   "31690_at"   "483_g_at"
## [31] "37133_at"   "40829_at"   "31626_i_at" "335_r_at"   "36478_at"
## [36] "36774_f_at" "1536_at"    "1475_s_at"  "38120_at"   "37441_at"
## [41] "574_s_at"   "34680_s_at" "31724_at"   "37505_at"   "36563_at"
## [46] "41668_r_at" "34502_g_at" "33619_at"   "34250_at"   "38490_r_at"
```

```
bs <-as.character(phenoData(fibroEset)$species)[as.character(phenoData(fibroEset)$species)=="b"]
gs <- as.character(phenoData(fibroEset)$species)[as.character(phenoData(fibroEset)$species)=="g"]
hs <- as.character(phenoData(fibroEset)$species)[as.character(phenoData(fibroEset)$species)=="h"]

length(bs); length(gs); length(hs)
```

```
## [1] 11
```

```
## [1] 12
```

```
## [1] 23
```

```
names(fibro.sample)<- c(paste(bs, '.', 1:length(bs), sep=''),
                         paste(gs, '.', 1:length(gs), sep=''), paste(hs, '.', 1:length(hs), sep=''))
names(fibro.sample)
```

```
##  [1] "b.1"  "b.2"  "b.3"  "b.4"  "b.5"  "b.6"  "b.7"  "b.8"  "b.9"  "b.10"
## [11] "b.11" "g.1"  "g.2"  "g.3"  "g.4"  "g.5"  "g.6"  "g.7"  "g.8"  "g.9"
## [21] "g.10" "g.11" "g.12" "h.1"  "h.2"  "h.3"  "h.4"  "h.5"  "h.6"  "h.7"
## [31] "h.8"  "h.9"  "h.10" "h.11" "h.12" "h.13" "h.14" "h.15" "h.16" "h.17"
## [41] "h.18" "h.19" "h.20" "h.21" "h.22" "h.23"
```

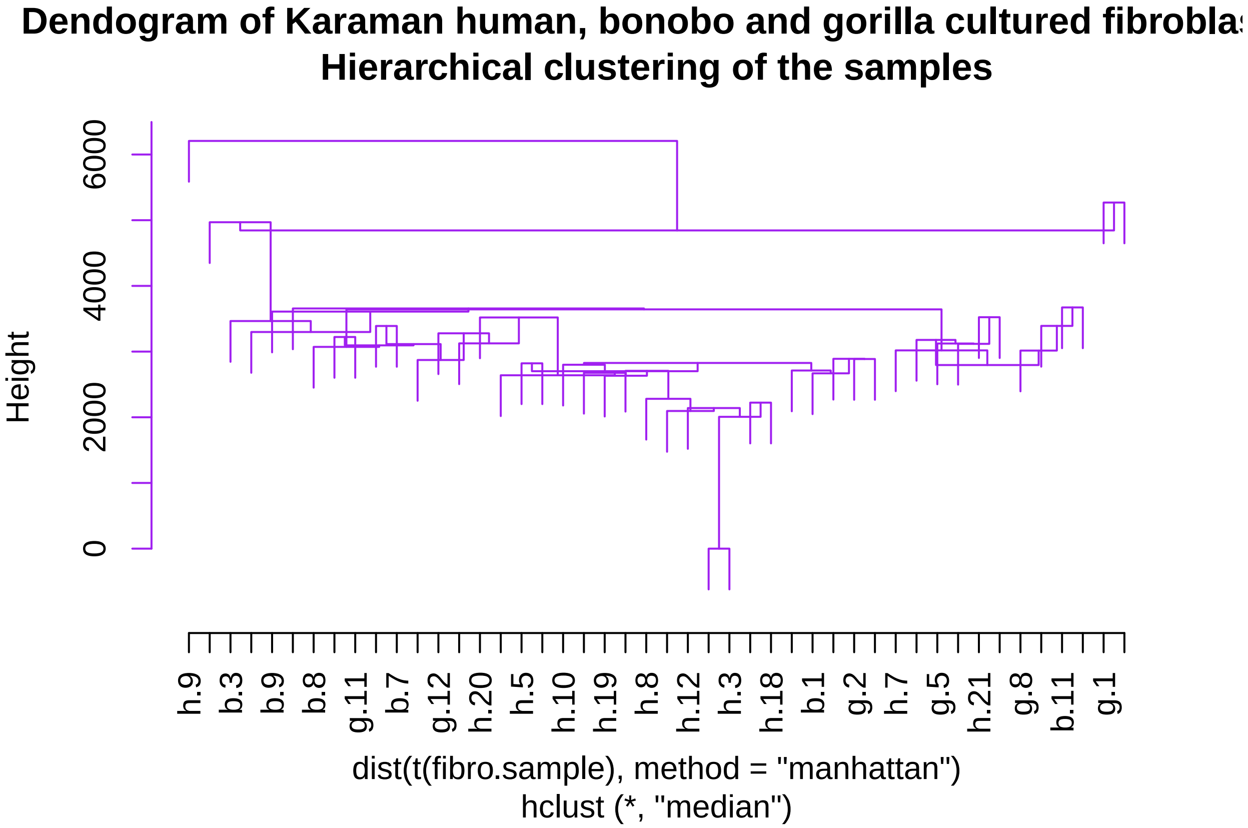
Run and plot hierarchical clustering of the samples using Manhattan distance metric and median linkage agglomeration (grouping) method

See https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust for details

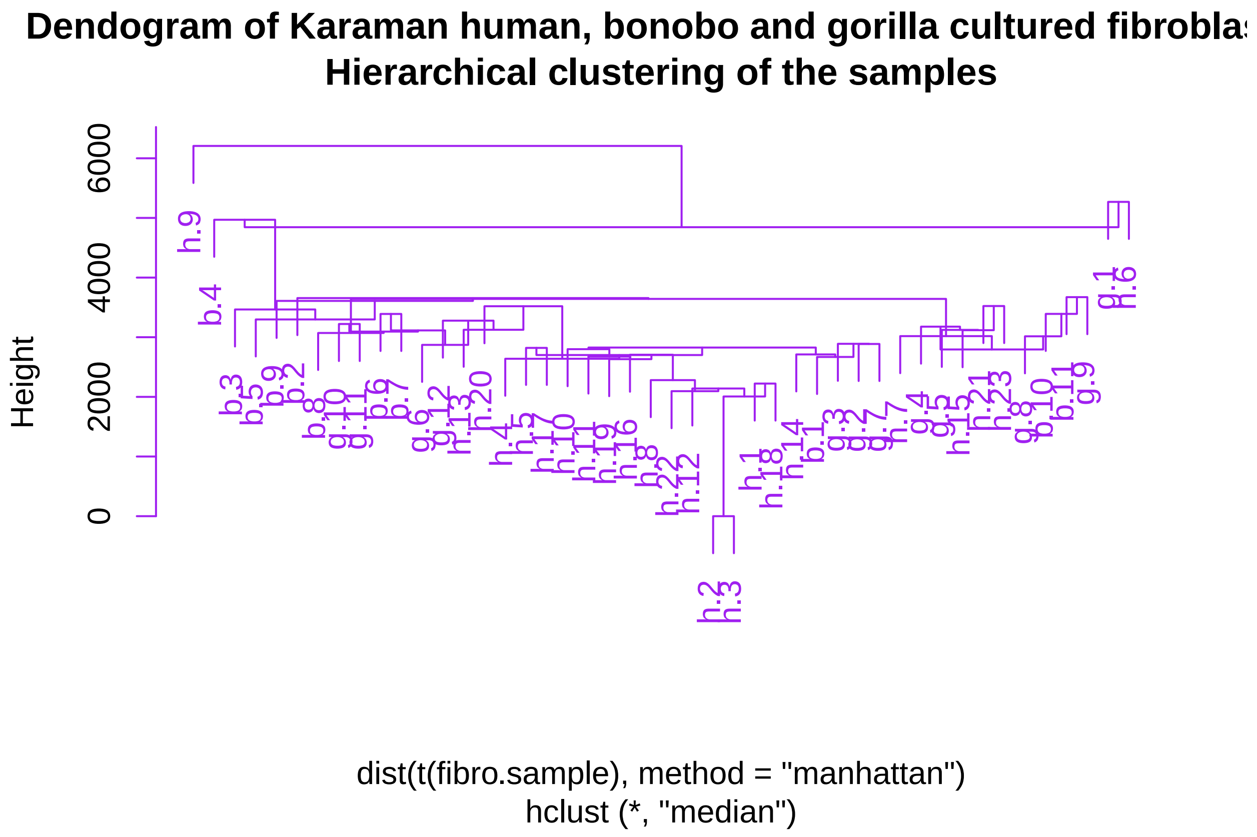
```
fibro.sample.dist <- dist(t(fibro.sample), method='manhattan')
fibro.hclust <- hclust(fibro.sample.dist, method='median')
# an object of class hclust which describes the tree produced by the iterative clustering process
unclass(fibro.hclust)
```

```
## $merge
##       [,1] [,2]
##  [1,]  -25  -26
##  [2,]  -24  -41
##  [3,]    1    2
##  [4,]  -35    3
##  [5,]  -45    4
##  [6,]  -31    5
##  [7,]  -39    6
##  [8,]  -42    7
##  [9,]  -34    8
## [10,]  -33    9
## [11,]  -28  -40
## [12,]  -13  -18
## [13,]  -14   12
## [14,]   -1   13
## [15,]  -37   14
## [16,]   10   15
## [17,]   11   16
## [18,]  -27   17
## [19,]  -21  -22
## [20,]   -6   -7
## [21,]  -43   18
## [22,]  -36   21
## [23,]  -23   22
## [24,]  -17   23
## [25,]   20   24
## [26,]   19   25
## [27,]   -8   26
## [28,]  -44  -46
## [29,]  -38   28
## [30,]  -16   29
## [31,]  -15   30
## [32,]  -11  -20
## [33,]  -10   32
## [34,]  -19   33
## [35,]   31   34
## [36,]  -30   35
## [37,]   27   36
## [38,]   -2   37
## [39,]   -9   38
## [40,]   -5   39
## [41,]   -3   40
## [42,]   -4   41
## [43,]  -12  -29
## [44,]   42   43
## [45,]  -32   44
##
## $height
##  [1]    0.000 2223.000 2006.750 2140.188 2095.547 2280.887 2706.847 2632.274
##  [9] 2675.959 2799.693 2821.000 2887.000 2889.750 2668.938 2711.734 2826.994
## [17] 2701.180 2639.340 3222.000 3390.000 3519.267 3125.100 3277.959 2871.927
## [25] 3114.009 3094.374 3071.371 3523.000 3117.750 3122.938 3177.484 3672.000
## [33] 3391.000 3015.250 2794.965 3018.351 3642.326 3656.175 3608.926 3298.186
## [41] 3465.158 4969.024 5269.000 4844.264 6206.742
##
## $order
##  [1] 32  4  3  5  9  2  8 21 22  6  7 17 23 36 43 27 28 40 33 34 42 39 31 45 35
## [26] 25 26 24 41 37  1 14 13 18 30 15 16 38 44 46 19 10 11 20 12 29
##
## $labels
##  [1] "b.1"  "b.2"  "b.3"  "b.4"  "b.5"  "b.6"  "b.7"  "b.8"  "b.9"  "b.10"
## [11] "b.11" "g.1"  "g.2"  "g.3"  "g.4"  "g.5"  "g.6"  "g.7"  "g.8"  "g.9"
## [21] "g.10" "g.11" "g.12" "h.1"  "h.2"  "h.3"  "h.4"  "h.5"  "h.6"  "h.7"
## [31] "h.8"  "h.9"  "h.10" "h.11" "h.12" "h.13" "h.14" "h.15" "h.16" "h.17"
## [41] "h.18" "h.19" "h.20" "h.21" "h.22" "h.23"
##
## $method
## [1] "median"
##
## $call
## hclust(d = fibro.sample.dist, method = "median")
##
## $dist.method
## [1] "manhattan"
```

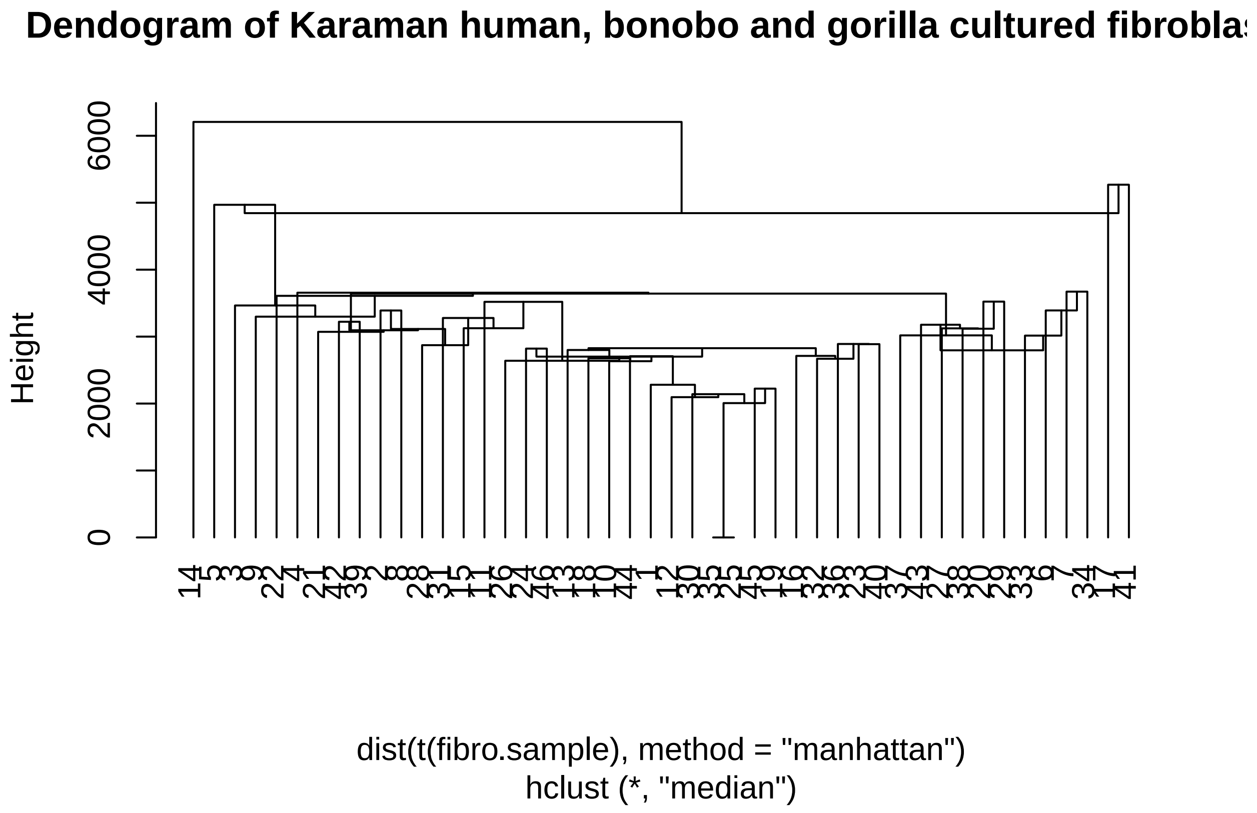
```
fibro.hclust <- hclust(dist(t(fibro.sample), method='manhattan'), method='median')

plot(fibro.hclust,
     main = 'Dendogram of Karaman human, bonobo and gorilla cultured fibroblasts\nHierarchical clusteri
axis(1, at = 1:length(fibro.hclust$labels), labels= fibro.hclust$labels[fibro.hclust$order], las=2)
```



```
plot(fibro.hclust,
     main = 'Dendogram of Karaman human, bonobo and gorilla cultured fibroblasts\nHierarchical clusteri
     col='purple')
```

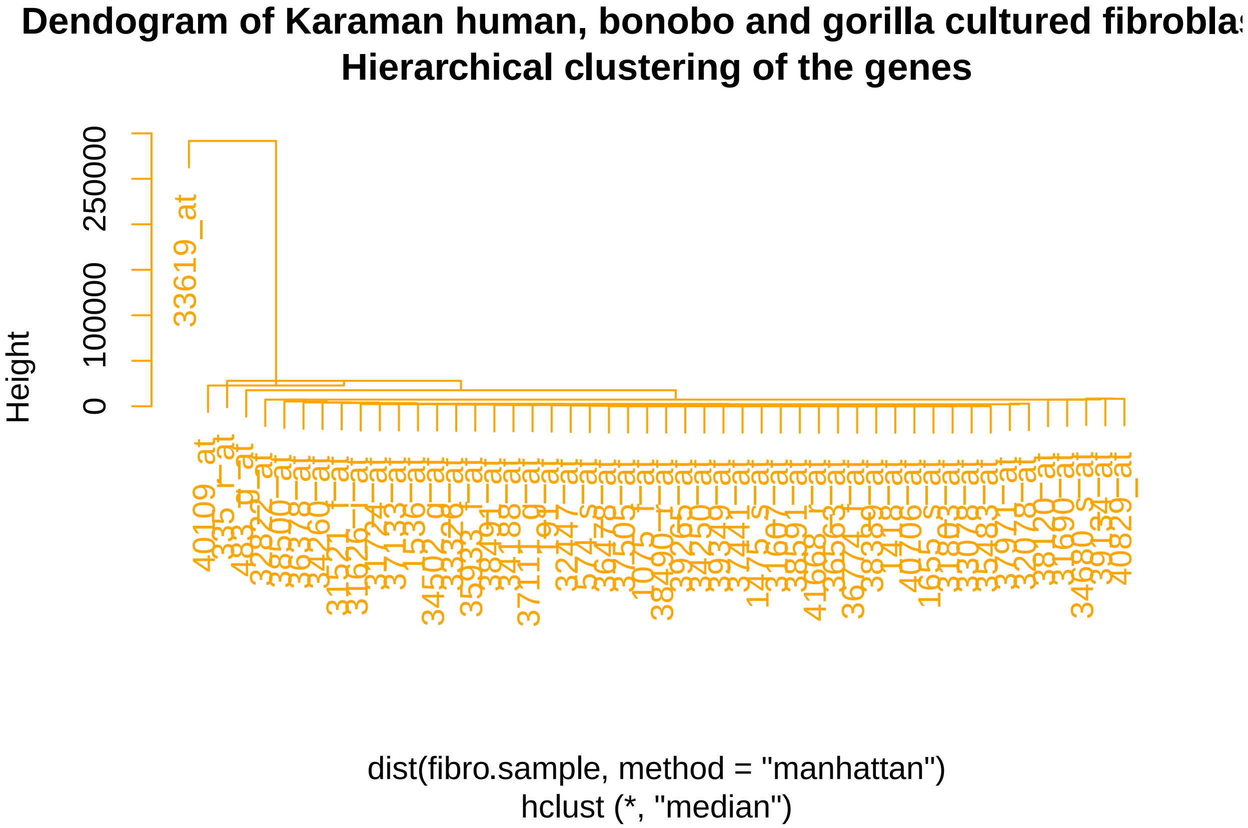


```r
plot(fibro.hclust,
     main = 'Dendogram of Karaman human, bonobo and gorilla cultured fibroblasts',
     labels= fibro.hclust$order,
     hang =-1)
```



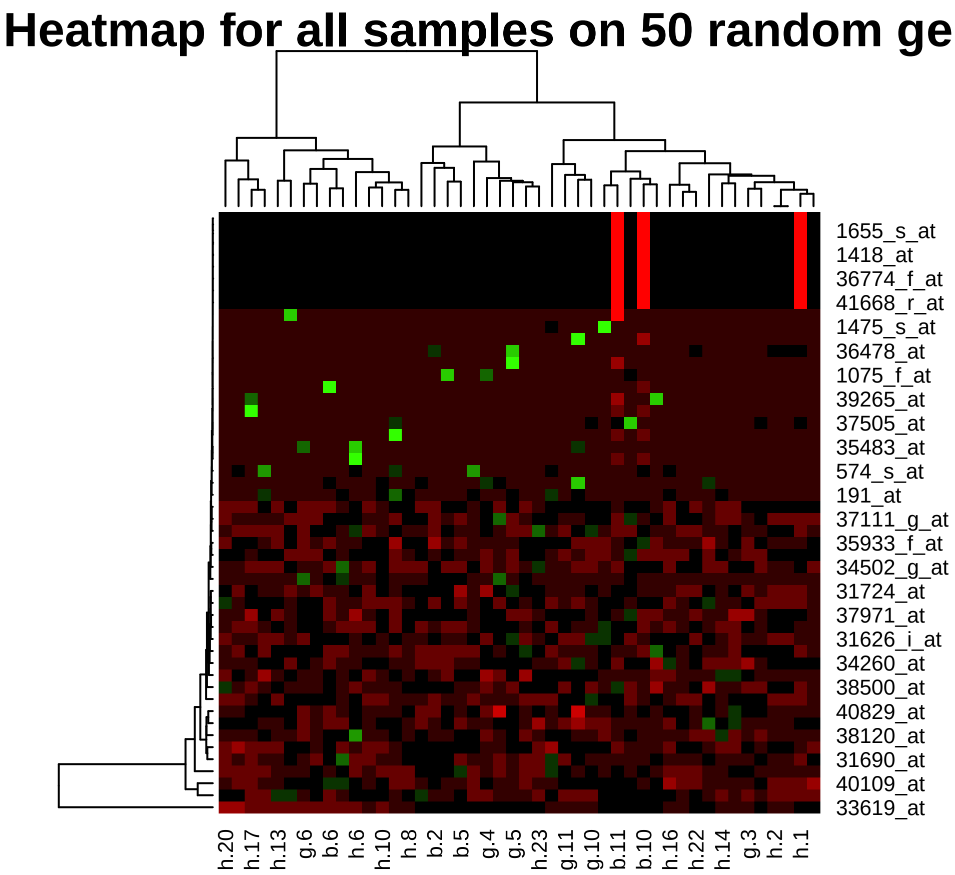
Hierachical clustering of the genes

```
fibro.hclust.g <- hclust(dist(fibro.sample, method='manhattan'),method='median')
plot(fibro.hclust.g,
     main = 'Dendogram of Karaman human, bonobo and gorilla cultured fibroblasts\nHierarchical clusteri
     col='orange')
```



Run hierarchical clustering and plot the results in two dimensions (on genes and samples): Plot a heatmap with genes on the y-axis and samples on the x-axis https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/heatmap

```
hm.rg <- c("#FF0000","#CC0000","#990000","#660000","#330000","#000000",
           "#000000","#0A3300","#146600","#1F9900","#29CC00","#33FF00")
heatmap(as.matrix(fibro.sample),
        main='Heatmap for all samples on 50 random genes',
        col=hm.rg, margins=c(2,2))
```



k-means clustering

Iterative algorithm that attempts to partition the dataset into k predefined distinct non-overlapping clusters where each data point belongs to only one cluster

Intra-cluster data points are as similar as possible while the clusters are kept as distinct (far) as possible.

Algorithm converges by minimization of distortion (solution is found by expectation-maximization method) Terminates the iterative assignment of data points to a cluster (E-step) when the sum of the squared distance between the data points and the cluster's centroid (calculated in M-step) is at the minimum The centroid is the arithmetic mean of all data points belonging to that cluster

The less variation within clusters, the more similar the data points are within the same cluster

Calculate PCA on the samples

Calculate k-means clustering on the first two principal components with k=3

```
fibro.sample.pca <- prcomp(t(fibro.sample))
fibro.loadings <- fibro.sample.pca$x[,1:2]
fibro.loadings
```

```
##              PC1          PC2
## b.1    132.85017  -232.184460
## b.2  -1673.70603   352.020949
## b.3  -2171.93436  1359.400301
## b.4    526.61363   781.939830
## b.5  -3054.05114   888.637131
## b.6   1426.70419  -301.003400
## b.7   1784.19973    17.032245
## b.8    656.28727  -340.179110
## b.9  -2715.38679   127.435928
## b.10 -1245.29663   679.499477
## b.11  -458.46306  1460.252731
## g.1   1432.58389  1458.267948
## g.2   -465.72565  -137.711595
## g.3   -301.58381  -459.009657
## g.4  -1716.36732  -649.626222
## g.5  -2127.52491  -203.905788
## g.6   2122.57987   -65.029685
## g.7   -539.05305  -374.371026
## g.8  -1326.40544   748.678227
## g.9   -529.36386   951.189311
## g.10  1021.33499  -215.307529
## g.11   -72.86215   -40.903551
## g.12  2277.31176   660.984424
## h.1   -609.43428  -849.536623
## h.2   -144.87934  -662.851925
## h.3   -144.87934  -662.851925
## h.4   2660.33069  -322.689707
## h.5   2269.86723  -907.631411
## h.6   1261.86535    38.261016
## h.7   -492.83171   715.517373
## h.8    692.56573   -61.522502
## h.9  -3249.65129  -808.602077
## h.10  1294.31552    -2.080419
## h.11   988.91323  -516.819315
## h.12  -180.06879  -243.517879
## h.13  2298.45014  1219.276262
## h.14   365.47541  -676.839499
## h.15 -2405.84299  -603.979399
## h.16    83.26905   135.690517
## h.17  2245.79991 -1041.940605
## h.18  -599.17857  -795.136386
## h.19  1053.51672   261.339919
## h.20  3764.60288   252.983752
## h.21 -1833.94866  -311.672232
## h.22  -493.81261   -13.394400
## h.23 -1807.18553  -608.109016
```

```
dim(fibro.loadings)
```

```
## [1] 46  2
```

```
cl <- kmeans(fibro.loadings, centers=3, iter.max=20)
cl
```

```
## K-means clustering with 3 clusters of sizes 15, 12, 19
##
## Cluster means:
```

```
##          PC1       PC2
## 1  1860.1584  35.70957
## 2 -2110.6084  80.81477
## 3  -135.5303 -79.23267
## 
## Clustering vector:
##  b.1  b.2  b.3  b.4  b.5  b.6  b.7  b.8  b.9 b.10 b.11  g.1  g.2  g.3  g.4  g.5
##    3    2    2    3    2    1    1    3    2    2    3    1    3    3    2    2
##  g.6  g.7  g.8  g.9 g.10 g.11 g.12  h.1  h.2  h.3  h.4  h.5  h.6  h.7  h.8  h.9
##    1    3    2    3    1    3    1    3    3    3    1    1    1    3    3    2
## h.10 h.11 h.12 h.13 h.14 h.15 h.16 h.17 h.18 h.19 h.20 h.21 h.22 h.23
##    1    1    3    1    3    2    3    1    3    1    1    2    3    2
## 
## Within cluster sum of squares by cluster:
## [1] 14773581 10193951 10737372
##  (between_SS / total_SS =  74.8 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
cluster1 <- cl$cluster[cl$cluster==1]
print("Cluster1 membership")
```

```
## [1] "Cluster1 membership"
```

```r
cluster1
```

```
##  b.6  b.7  g.1  g.6 g.10 g.12  h.4  h.5  h.6 h.10 h.11 h.13 h.17 h.19 h.20
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
```

```r
cluster2 <- cl$cluster[cl$cluster==2]
print("Cluster2 membership")
```

```
## [1] "Cluster2 membership"
```

```r
cluster2
```

```
##  b.2  b.3  b.5  b.9 b.10  g.4  g.5  g.8  h.9 h.15 h.21 h.23
##    2    2    2    2    2    2    2    2    2    2    2    2
```

```r
cluster3 <- cl$cluster[cl$cluster==3]
print("Cluster3 membership")
```

```
## [1] "Cluster3 membership"
```

```r
cluster3
```

```
##  b.1  b.4  b.8 b.11  g.2  g.3  g.7  g.9 g.11  h.1  h.2  h.3  h.7  h.8 h.12 h.14
##    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3    3
## h.16 h.18 h.22
##    3    3    3
```

```
length(cluster1);length(cluster2);length(cluster3)
```

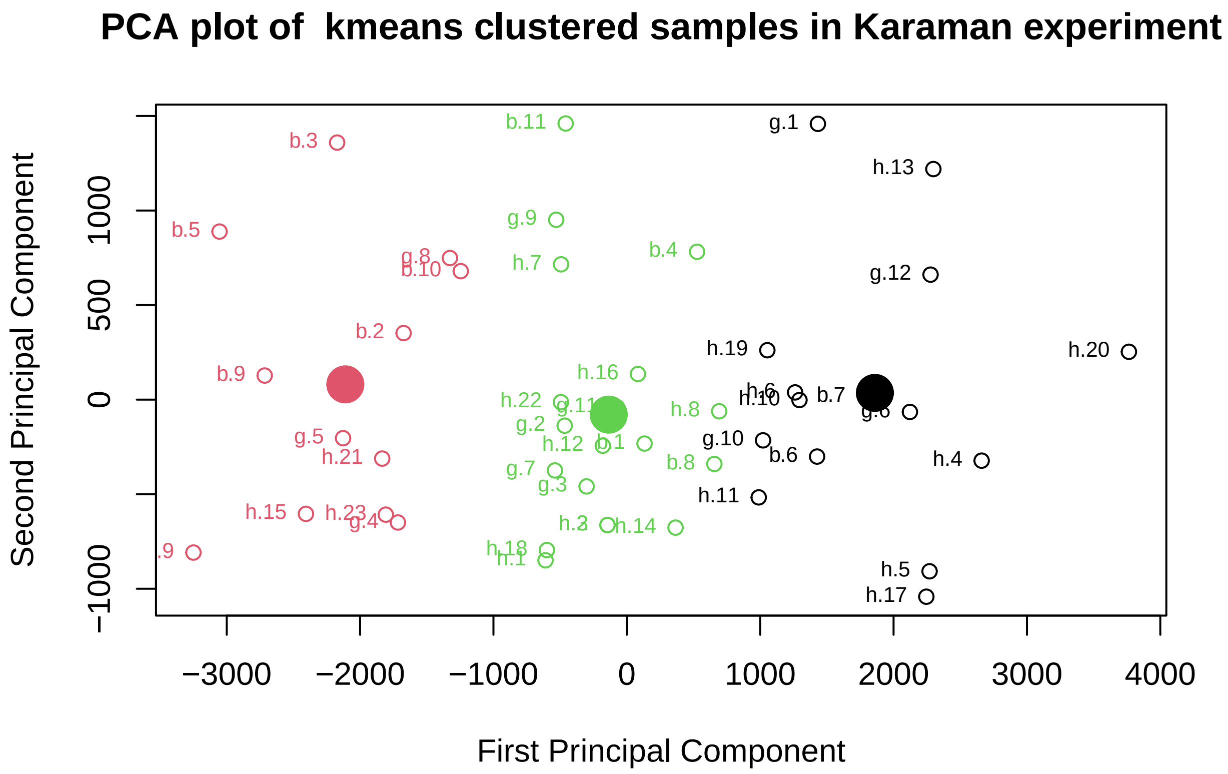
```
## [1] 15
```

```
## [1] 12
```

```
## [1] 19
```

Plot a two-dimensional scatter plot of the sample classification labels, embedded with the first two PCA eigenfunctions

```
plot(fibro.loadings, col = cl$cluster,cex=1,
     main='PCA plot of  kmeans clustered samples in Karaman experiment',
     xlab='First Principal Component', ylab='Second Principal Component')
text(fibro.loadings[,1], fibro.loadings[,2],
     col= cl$cluster,cex=0.7,
     labels= row.names(fibro.loadings), pos=2)
points(cl$centers, col = 1:3, pch = 19, cex=2.5)
```


PCA plot of kmeans clustered samples in Karaman experiment
Second Principal Component
First Principal Component
b.11
g.1
b.3
h.13
b.5
g.9
b.4
g.8
b.10
h.7
g.12
b.2
h.19
h.20
b.9
h.16
h.6
h.10
b.7
h.22
g.11
g.6
h.8
g.2
g.5
h.12
b.1
g.10
h.21
b.6
h.4
g.7
b.8
g.3
h.11
h.15
h.23
g.4
h.3
h.14
h.18
h.1
h.5
h.17