

Supplementary information

mdclust - Exploratory Microarray Analysis by Multidimensional Clustering

M. Dugas, S. Merk, S. Breit, P. Dirschedl

Comparison of results from search.mdclust() with (Golub et al.1999)

6 out of 10 genes identified by mdclust are on the list of top genes published by Golub (Science 1999, p.534).

Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-537

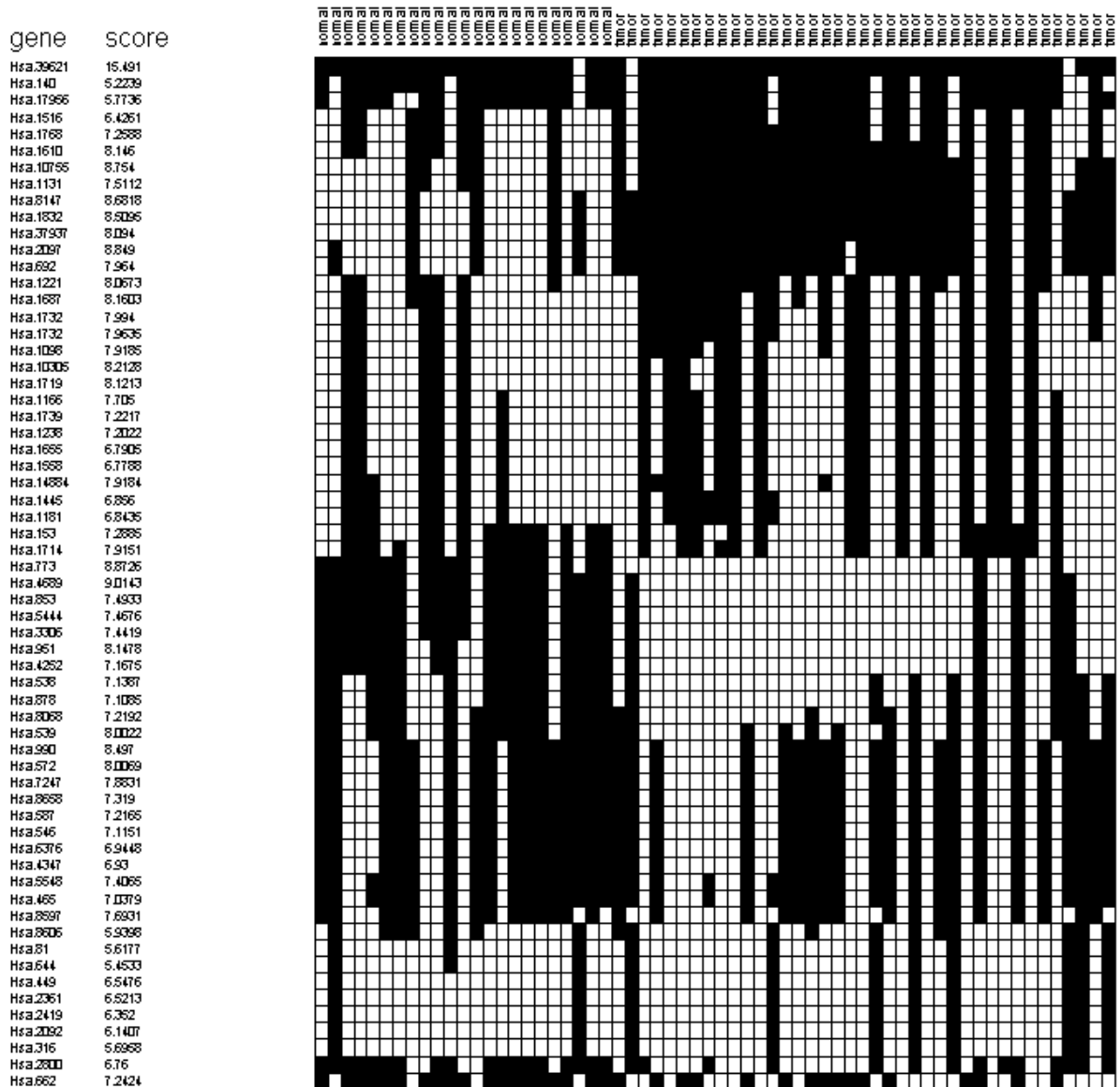
genes identified by mdclust	probe set	description	Golub's gene list
2124	X95735_at	Zyxin	yes
1778	X07743_at	PLECKSTRIN	no
766	M16038_at	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	yes
108	D14874_at	ADM Adrenomedullin	no
808	M23197_at	CD33 CD33 antigen (differentiation antigen)	yes
1037	M91432_at	ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	yes
1995	X74262_at	RETINOBLASTOMA BINDING PROTEIN P48	yes
2489	U22376_cds2_s_at	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds	yes
2851	U72936_s_at	X-LINKED HELICASE II	no
561	L07758_at	IEF SSP 9502 mRNA	no

mdclust (Alon et al. 1999)

The data set consists of 62 samples and 2000 genes. Colon-tumor versus normal samples are compared.

Two main groups can be identified by mdclust, but the result is less clear than with the Alizadeh data.

U Alon, N Barkai, D A Notterman, K Gish, S Ybarra , D Mack, A J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 1999, 96: 6745–6750

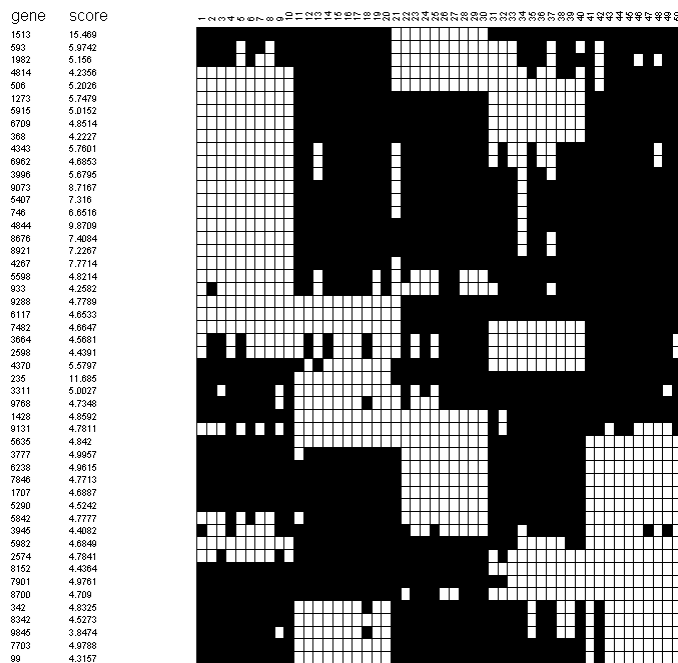


Simulated data with 5 groups

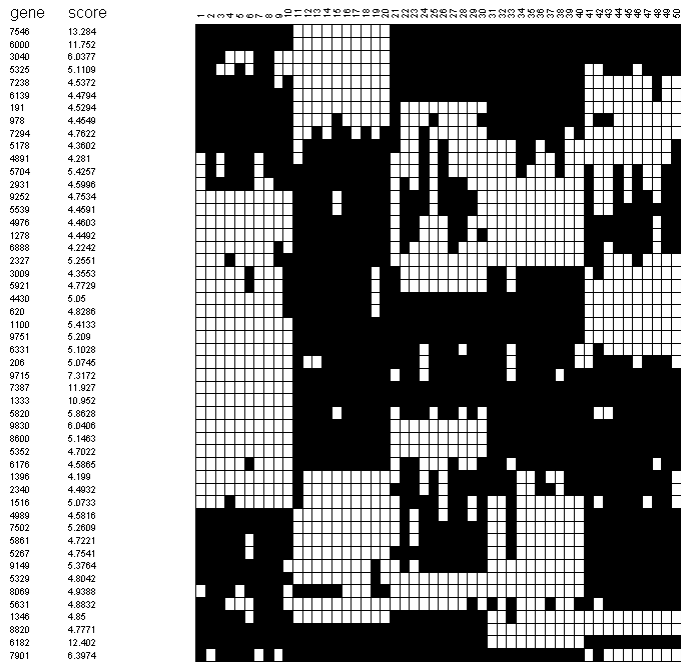
The simulated data set consists of 50 samples, 10,000 genes and 100 differential genes specific for 5 groups. The parameter delta defines the signal-to-noise ratio in standard deviation units. For $\delta \geq 2$ mdclust is capable to detect 5 groups.

R code to generate the data

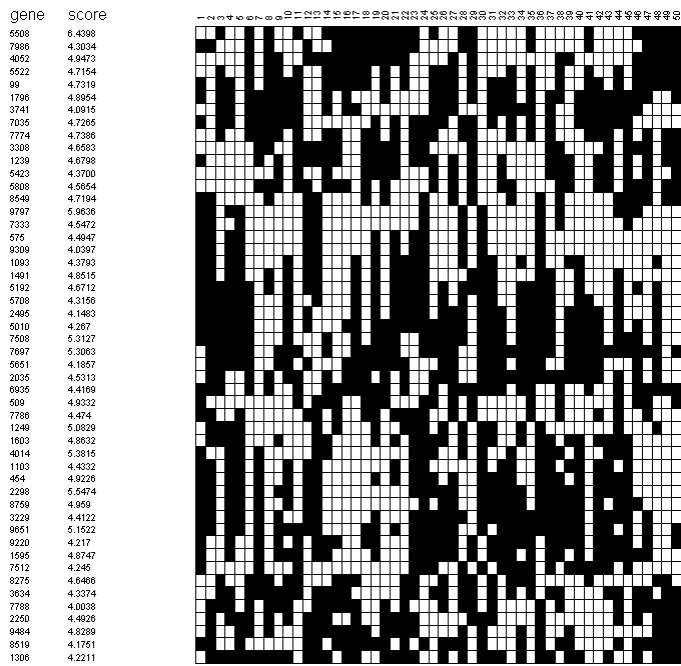
```
xdata <- matrix(rnorm(5E5),ncol=50)
diffgenes <- sample(1:10000,100)
# delta in SD
delta <- 3 # 1 and 2 respectively
for (i in 1:5) {
  start <- (i-1)*20 + 1;
  stop <- i*20;
  start2 <- (i-1)*10 + 1;
  stop2 <- i*10;
  for (g in diffgenes[start:stop]) {
    xdata[g,start2:stop2] <- xdata[g,start2:stop2] + delta
  }
}
```



delta= 3SD



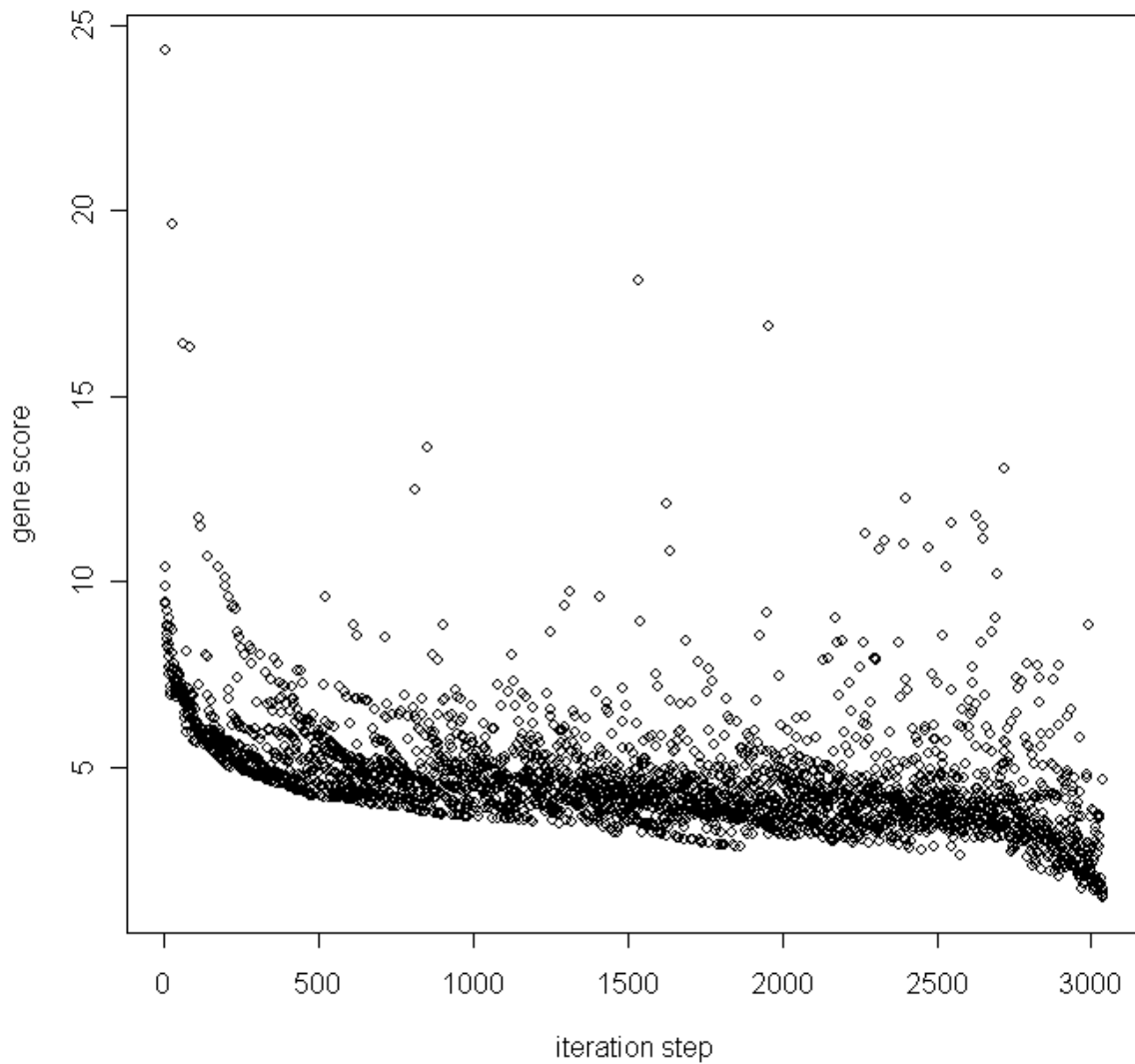
delta = 2 SD



delta = 1 SD

Gene scores by iteration step

The number of clusters (respectively genes) to be selected in the first step of the mdclust algorithm is subject to discussion. A suitable choice depends on the data. For n genes at most n clusters could be determined. The next figure shows gene scores by iteration step for Golub's data. There is a clear trend that genes with high scores are selected first, however, this is not guaranteed.



Reproducibility of mdclust

To assess the reproducibility of mdclust we performed 10 replications of the analysis on the same data set. The results are very similar, but not identical. Because the genes are sorted according to Hamming-distance, a gene with a very high score can change the order of sorting.

