Class discovery

Wessel van Wieringen
w.vanwieringen@vumc.nl

Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

VU medisch centrum

Outline

Topics discussed

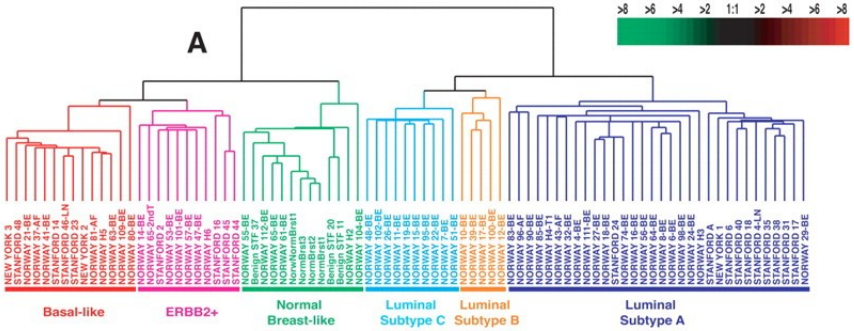
1.	Motivation
2.	Exploratory data analysis
3.	Hierarchical clustering (dendrogram & heatmap)
4.	Hierarchical clustering (similarity & linkage)
5.	Consensus clustering
6.	*K*-means
7.	Principal component analysis
8.	A published example

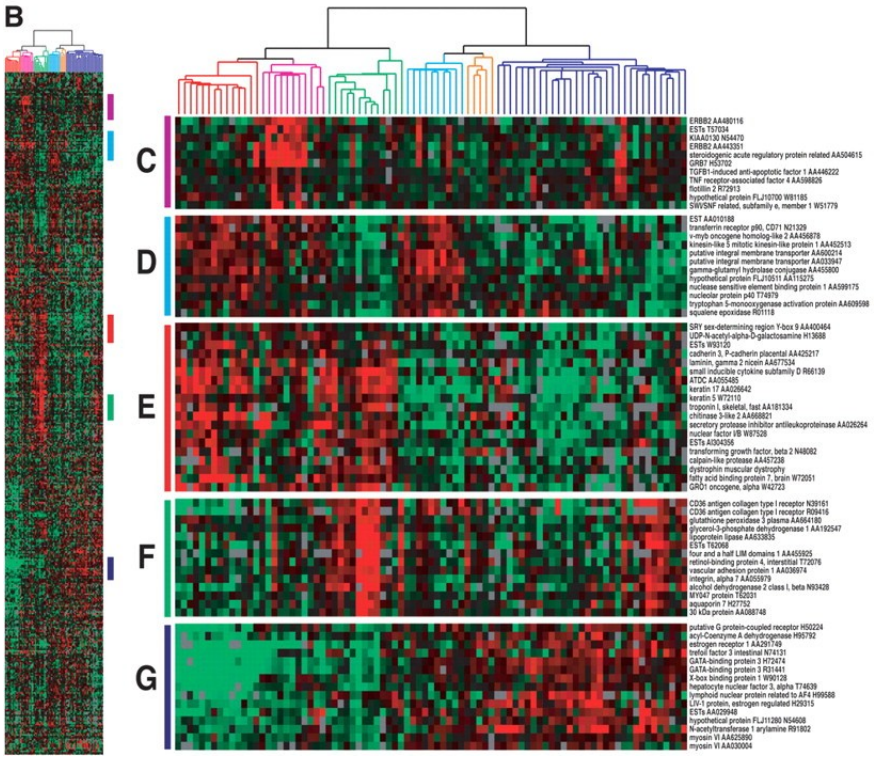
Motivation

Motivation

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]





Using 78 breast cancer profiles, five subtypes are distinguished:

- Basal
- ERBB2
- Luminal A
- Luminal B
- Normal

Motivation

Traditional medicine

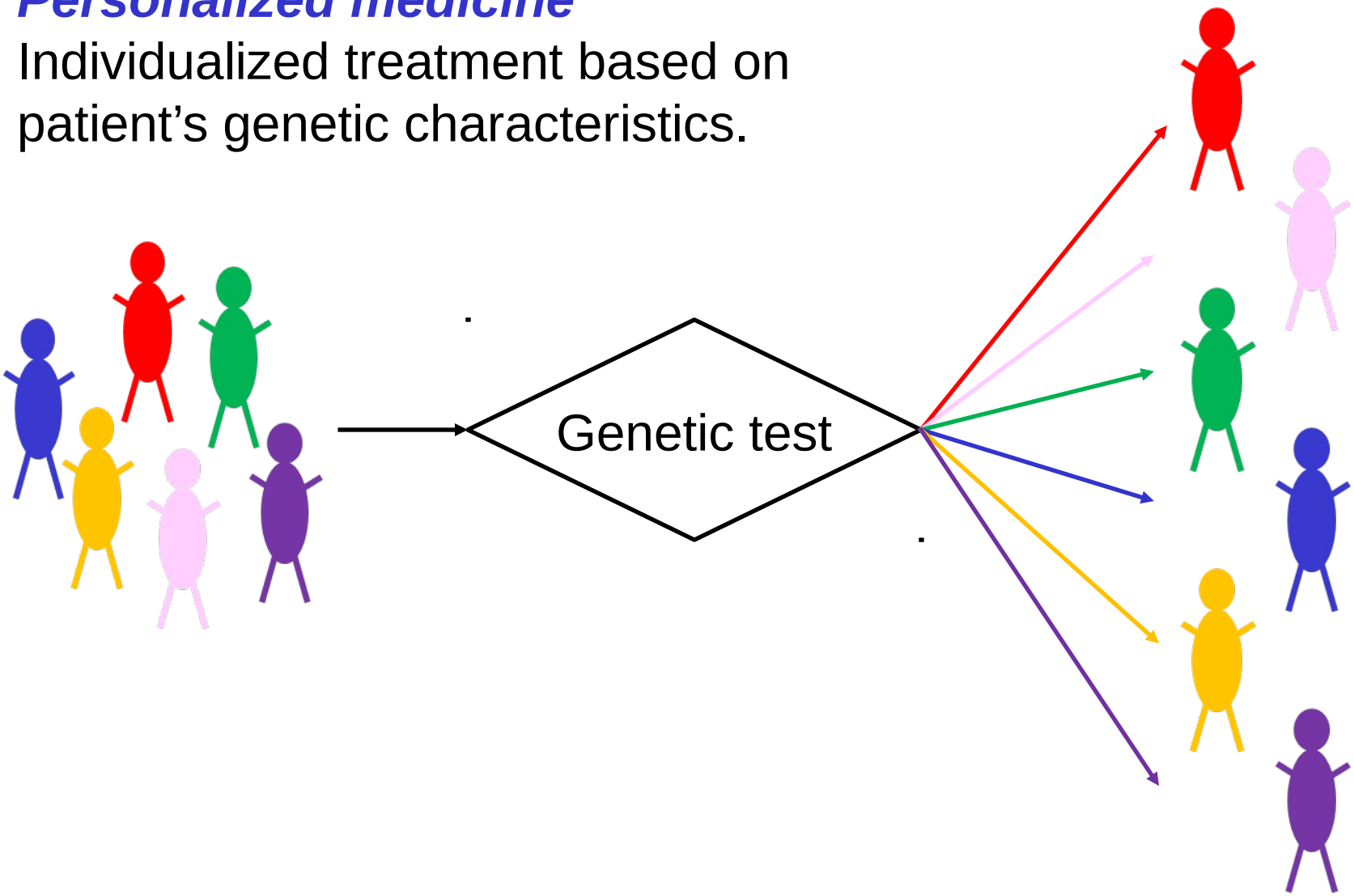
Standard treatment may not be beneficial to everyone.

Subgrouping of breast cancers suggest patients from different subgroups may need different therapy.

Motivation

Personalized medicine

Individualized treatment based on patient's genetic characteristics.



Exploratory data analysis

Exploratory data analysis

Exploratory data analysis (EDA) is a hypothesis identification approach. It consists of the analysis of observational data, often collected without well-defined hypotheses, with the purpose of finding clues that could inspire ideas and hypotheses.

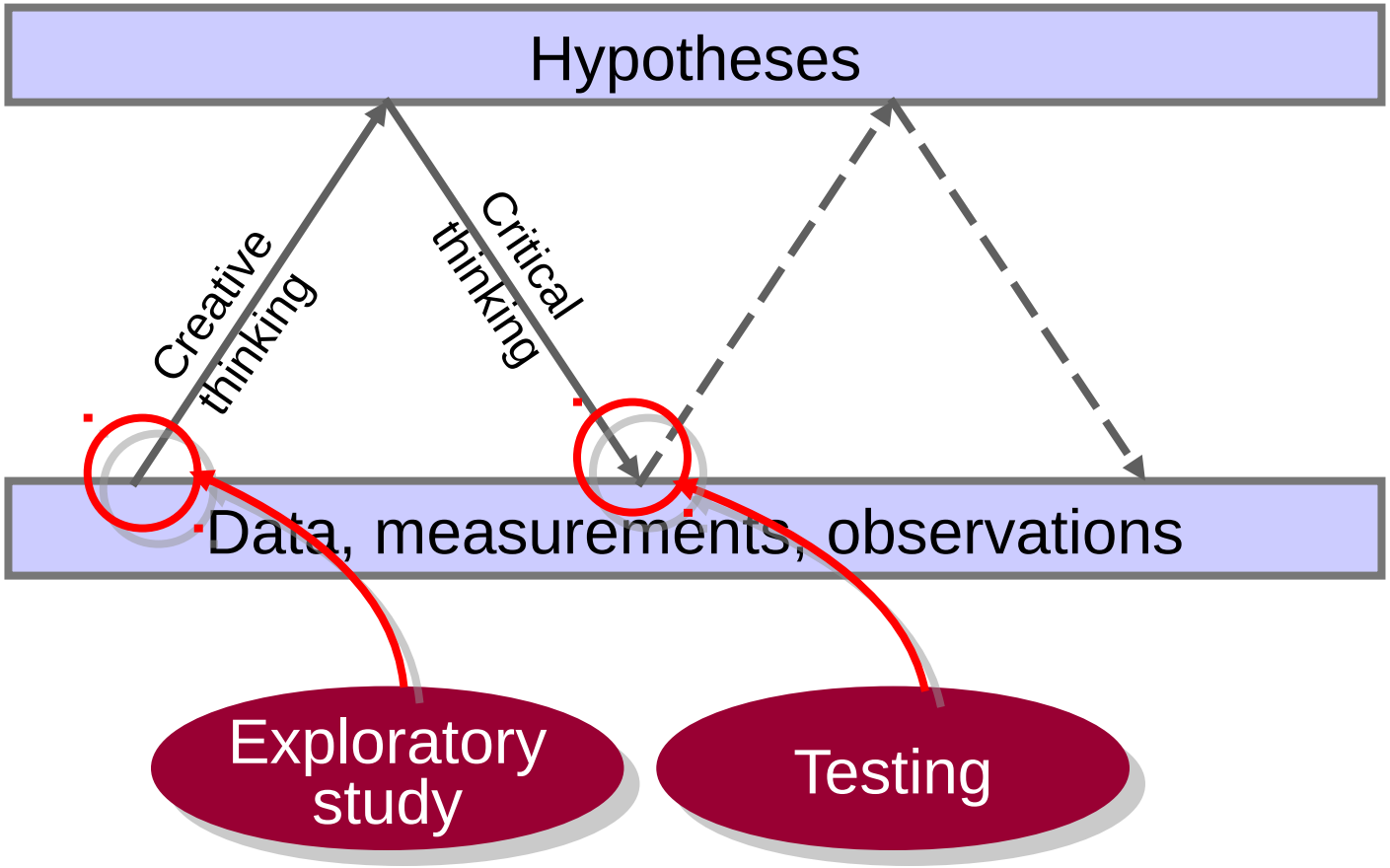
The EDA toolbox encompasses class discovery.

Main steps of EDA:		Clustering:
1) display the data,	⟶	dendrogram,
2) identify salient features,	⟶	samples clusters,
3) interpret salient features.	⟶	interpret clusters.

De Mast, Kemper (2009) in *Quality Engineering*

Exploratory data analysis

How does EDA fit in with the scientific method?



Box (2001) in *Journal of Applied Statistics*

Hierarchical clustering

(dendrogram, heatmap)

Hierarchical clustering

Objective of cluster analysis

Cluster analysis seeks

- meaningful data-determined groupings of samples, s.t.
- samples are more “similar” within than across groups,
- this similarity in gene expression profiles is assumed to imply some form of phenotypic similarity of the samples.

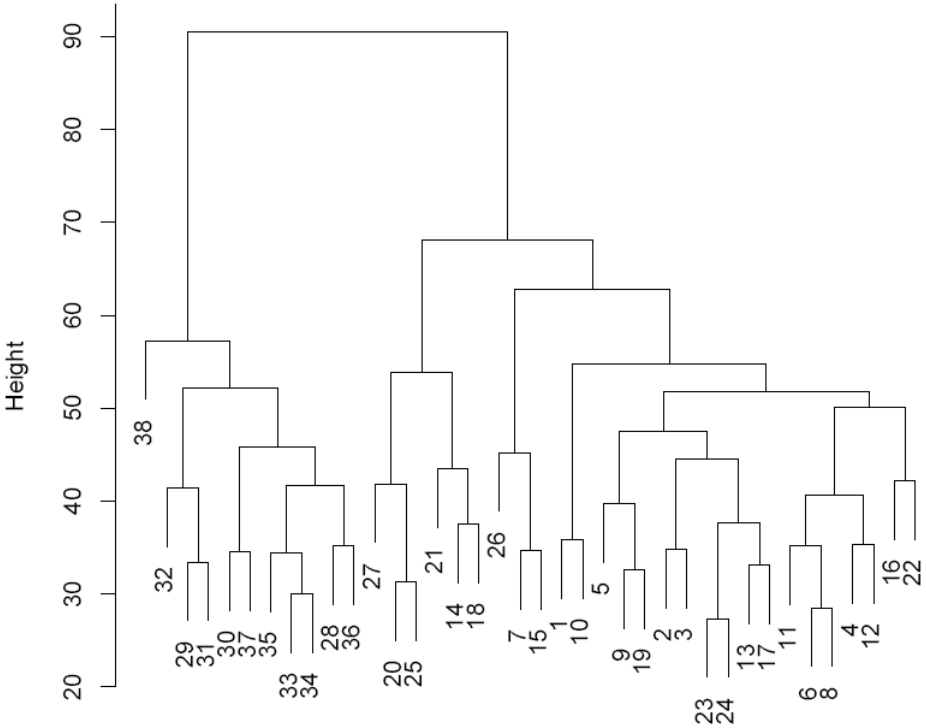
Cluster analysis is also known as unsupervised learning, class discovery, and data segmentation

Hierarchical clustering

Hierarchical clustering produces a nested sequence of clusters. It starts with all objects apart, and at each step two clusters are merged until only one is left.

The nested sequence can be represented by a dendrogram.

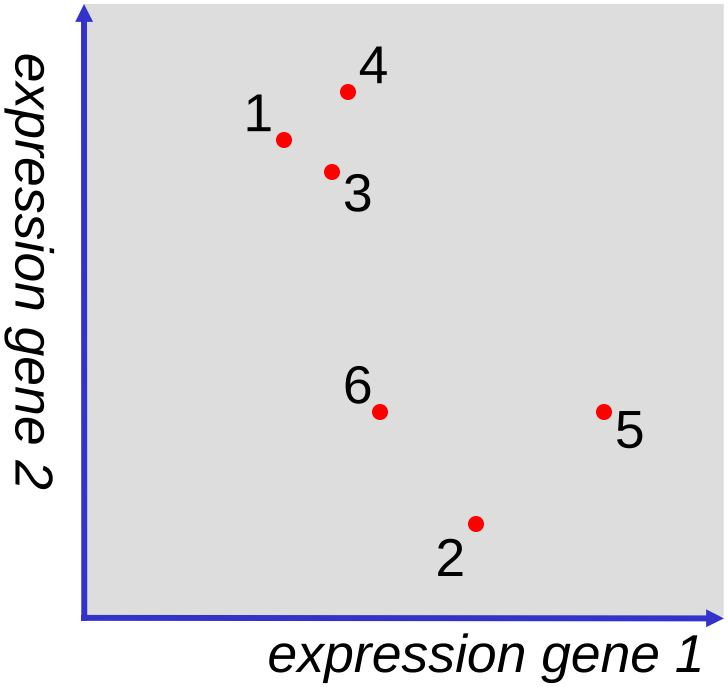
A ***dendrogram*** is a two-dimension diagram, a tree. Each fusion of clusters is plotted at a height equal to the dissimilarity of the two clusters which are joined.

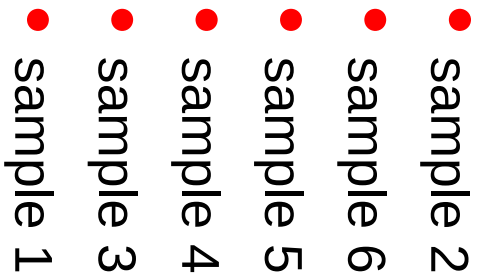


Hierarchical clustering

Building a dendrogram (loosely):

- Find samples that have most similar gene expression profiles.

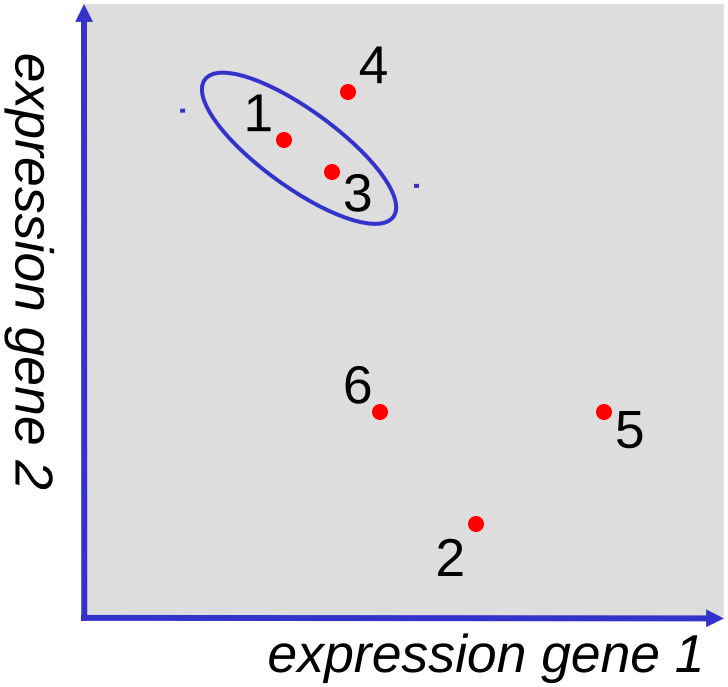


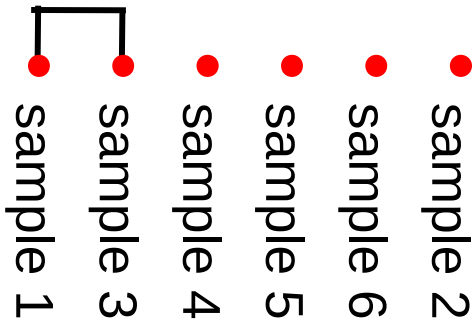


Hierarchical clustering

Building a dendrogram (loosely):

- Samples 1 and 3 have most similar gene expression profiles. Let these samples form a cluster.
- Repeat this exercise.

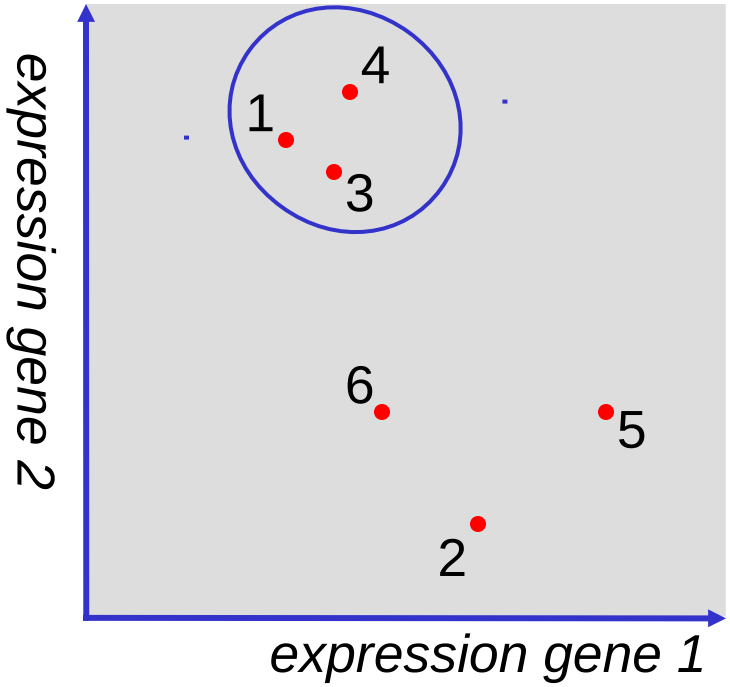


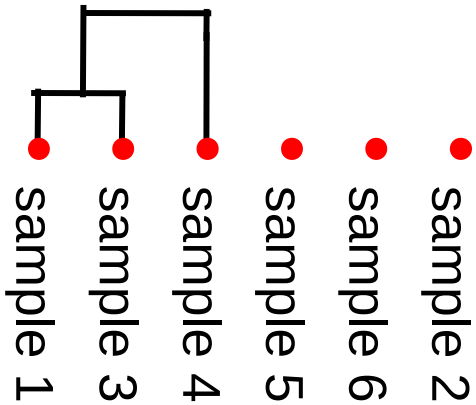


Hierarchical clustering

Building a dendrogram (loosely):

- Look for samples or clusters that have most similar gene expression profiles.

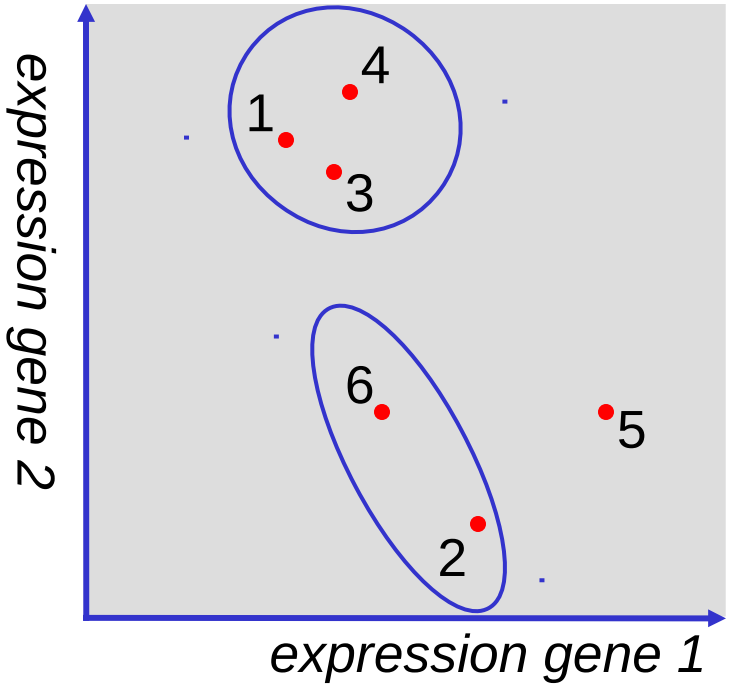


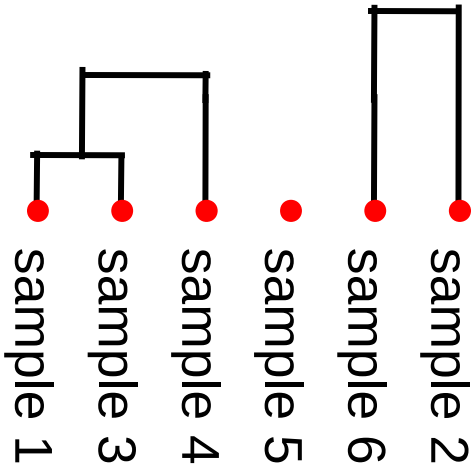


Hierarchical clustering

Building a dendrogram (loosely):

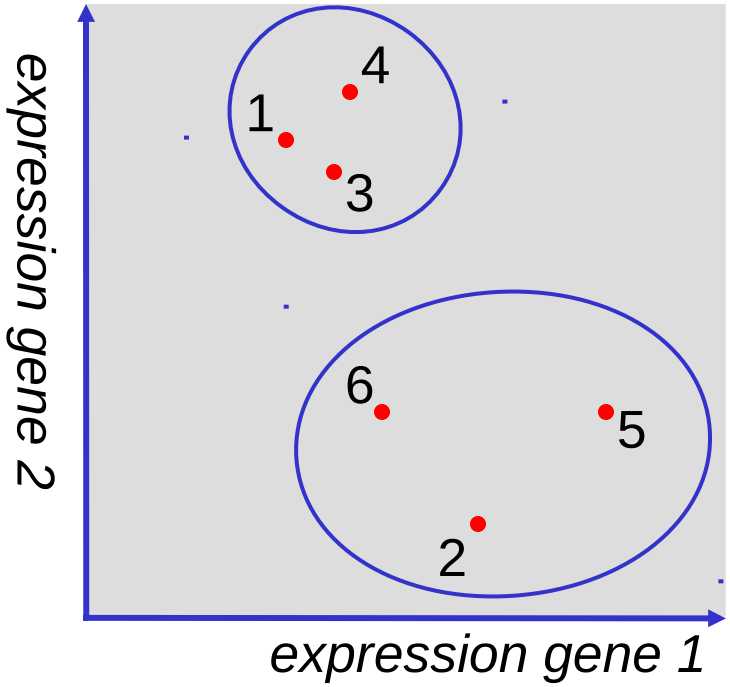
- New clusters may form: samples 2 and 6.

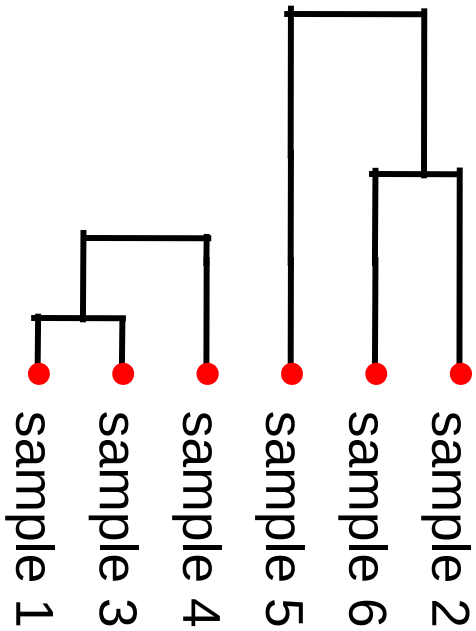




Hierarchical clustering

Building a dendrogram (loosely).

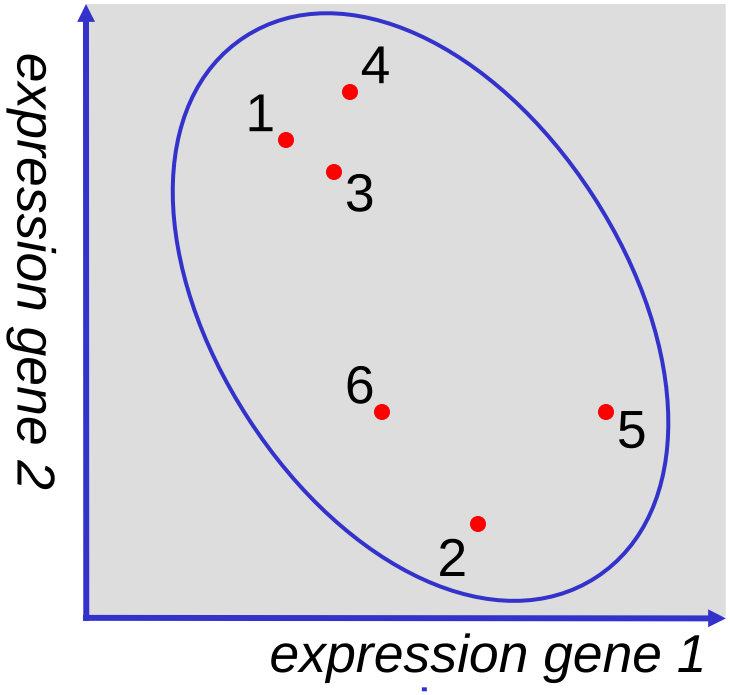


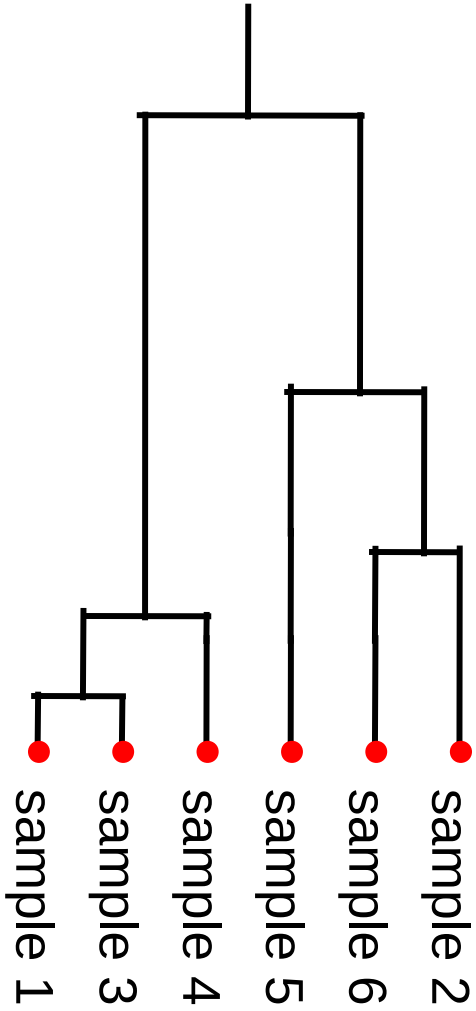


Hierarchical clustering

Building a dendrogram (loosely):

- Finally, all samples/clusters are merged into one big cluster.





Hierarchical clustering

Algorithm

Step 1: Form initial clusters, each containing one sample.

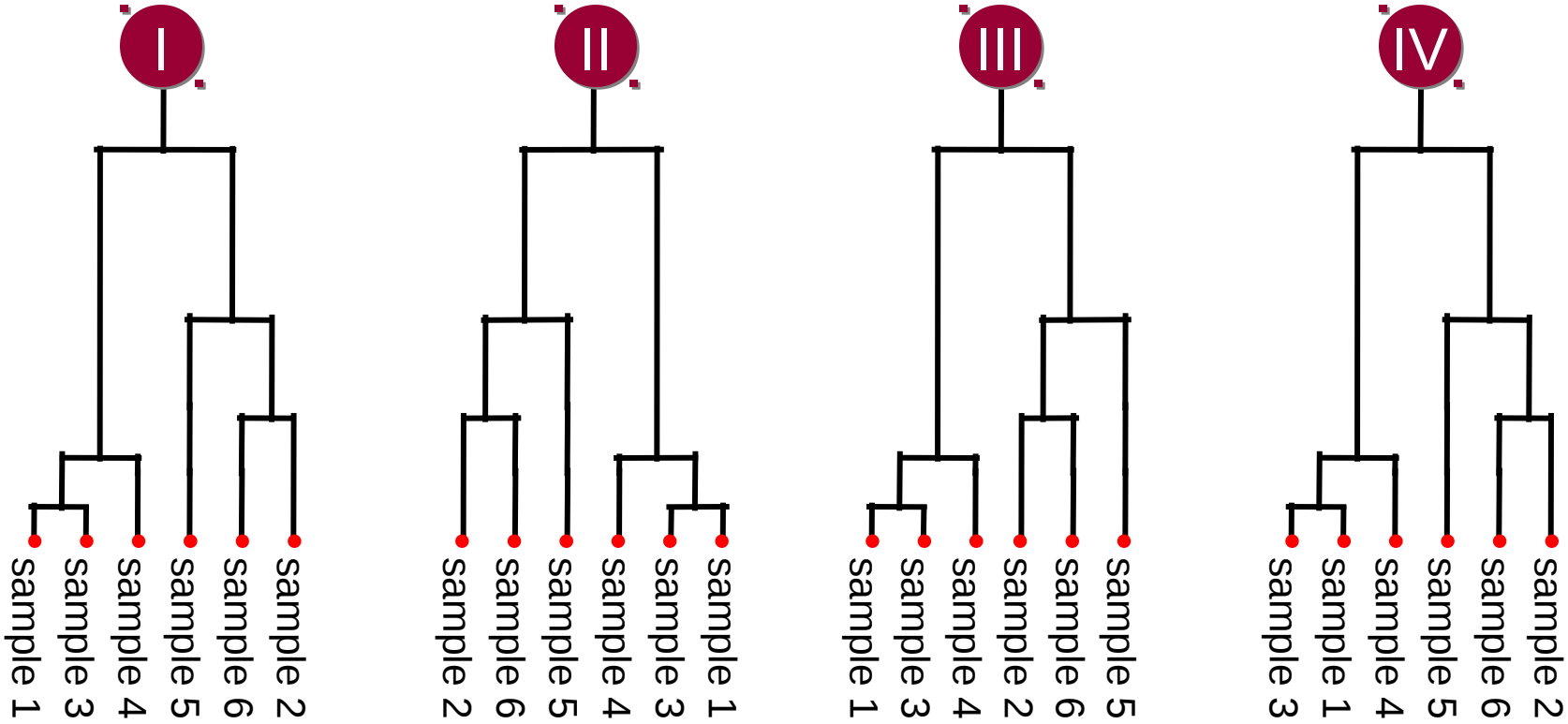
Step 2: Calculate the distance between all cluster pairs.

Step 3: Merge the two clusters with smallest distance.

Step 4: Iterate between step 3 and 4 until one final cluster remains.

Step 5: Draw dendrogram and heatmap.

Hierarchical clustering



Question

Consider the four hierarchical clusterings of six samples. Which of them do you prefer? Motivate your answer.

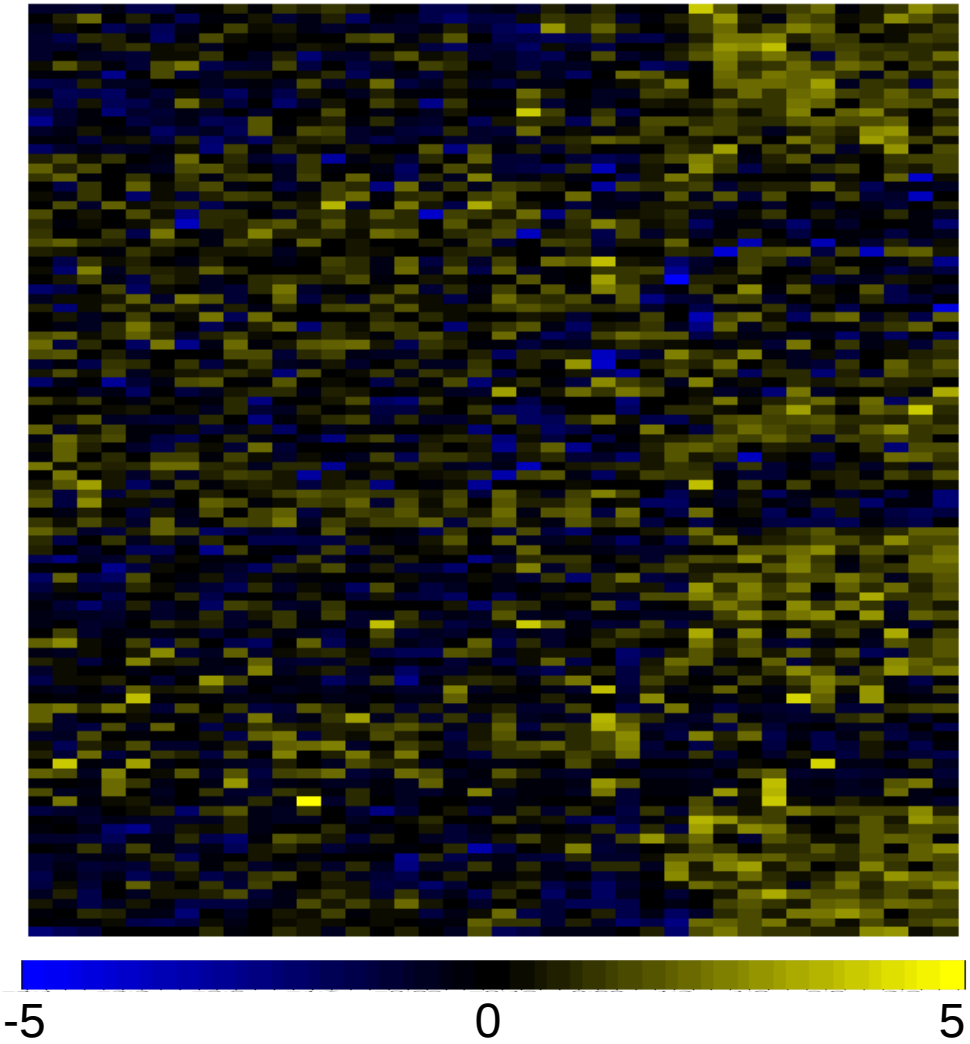
Hierarchical clustering

Heatmap

A dendrogram is often used in combination with heatmap.

A ***heatmap*** is a graphical representation of data where the values taken by a variable in a two-dimensional map are represented as colors.

Heatmap of a expression matrix

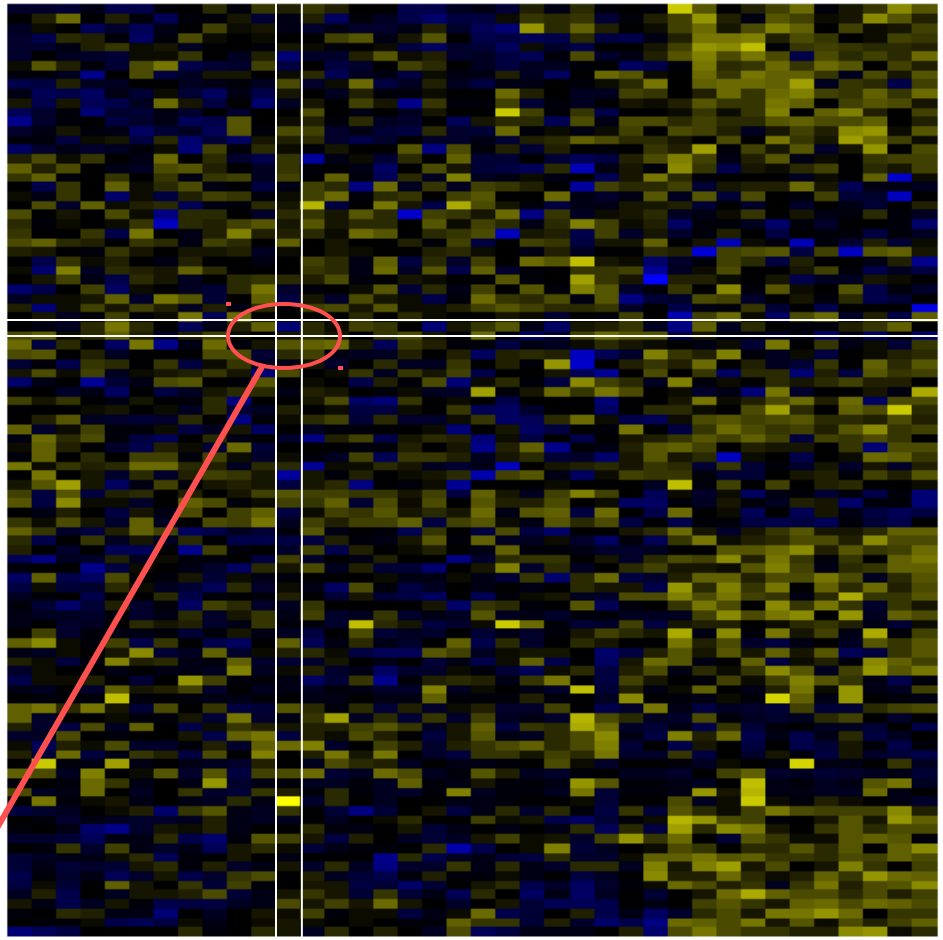


Hierarchical clustering

Expression matrix

	S_1		S_i		S_50
g_1	-1.3	…	0.1	…	-1.2
g_2	-0.1	…	2.4	…	0.3
…	…	…	…	…	…
…	…	…	…	…	…
g_j	0.4	…	1.5	…	-0.2
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
g_846	-0.9	…	-0.8	…	0.4

Heatmap

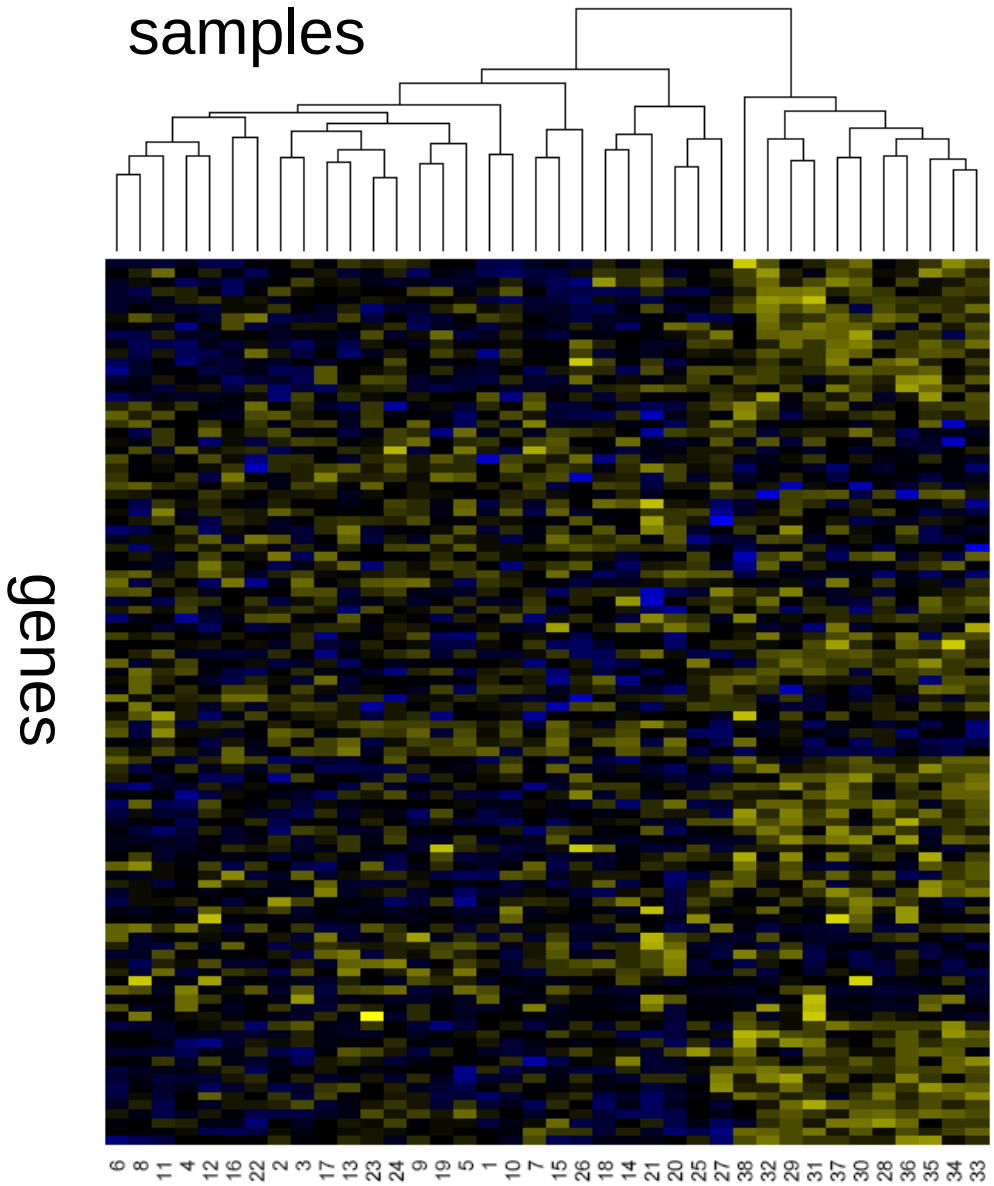


expression of gene *j* in sample *i*

Hierarchical clustering

Visualization of hierarchical clustering results:

dendrogram and heatmap combined

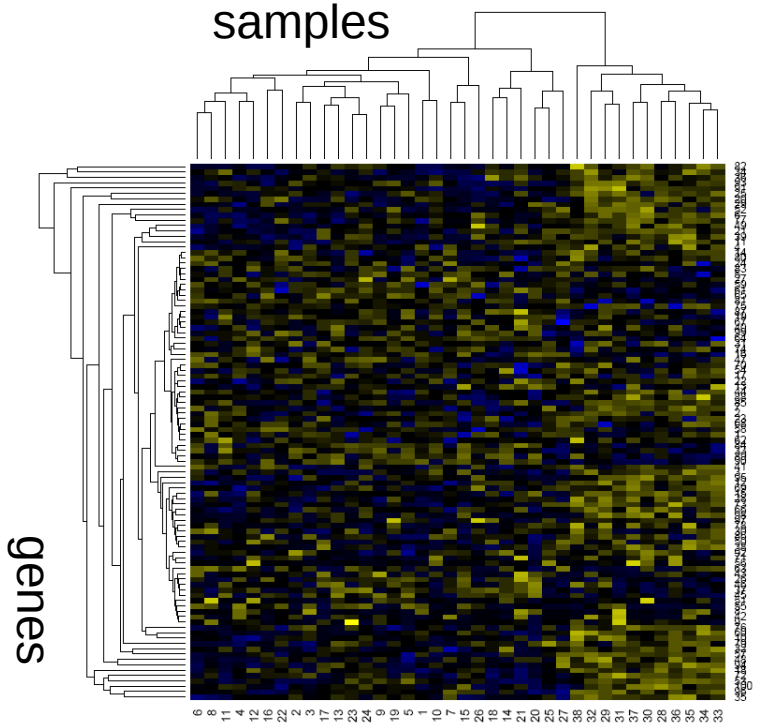


Hierarchical clustering

Hierarchical clustering of genes

- Genes that cluster together are believed to be functionally related (modules / pathway / GO node).
- This may help to characterize unknown genes.

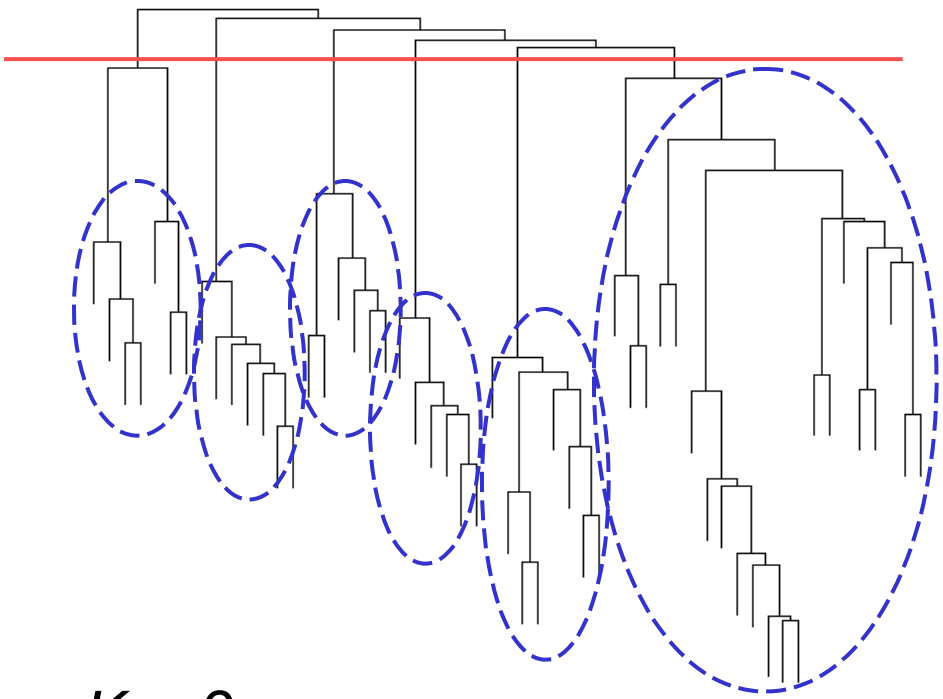
May also cluster samples and genes simultaneously.



Hierarchical clustering

Selection of *K* clusters from a hierarchical clustering corresponds to cutting the dendogram with a horizontal line at an appropriate height.

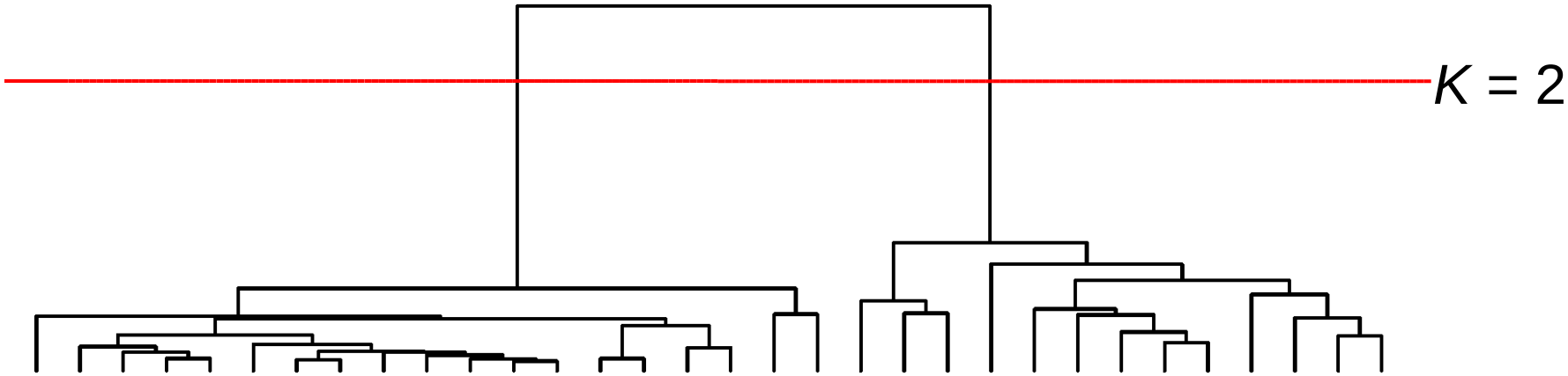
Each branch cut by the horizontal line corresponds to a cluster.



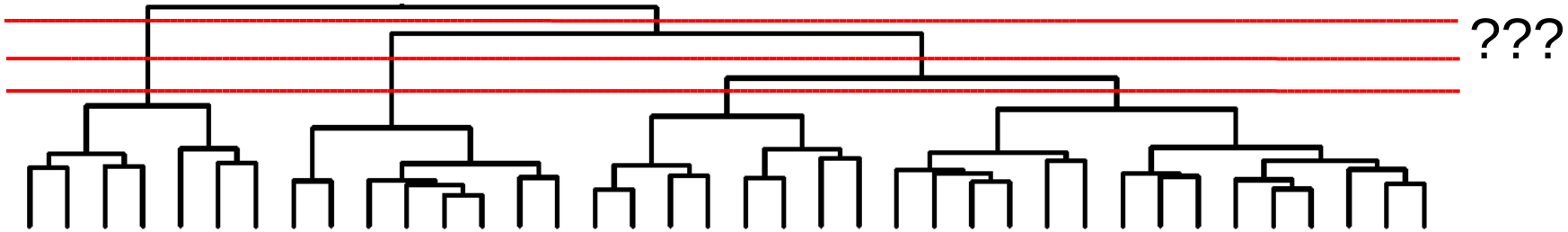
K = 6

Hierarchical clustering

K is often chosen by visual inspection, which may seem reasonable …



… but is often highly disputable.



Hierarchical clustering

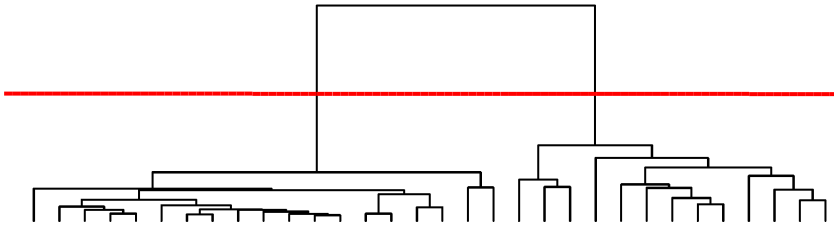
Q : How then to decide on the number of clusters *K* from a dendrogram?

A1 : This is often done through a robustness / stability analysis, e.g., ***consensus clustering***.

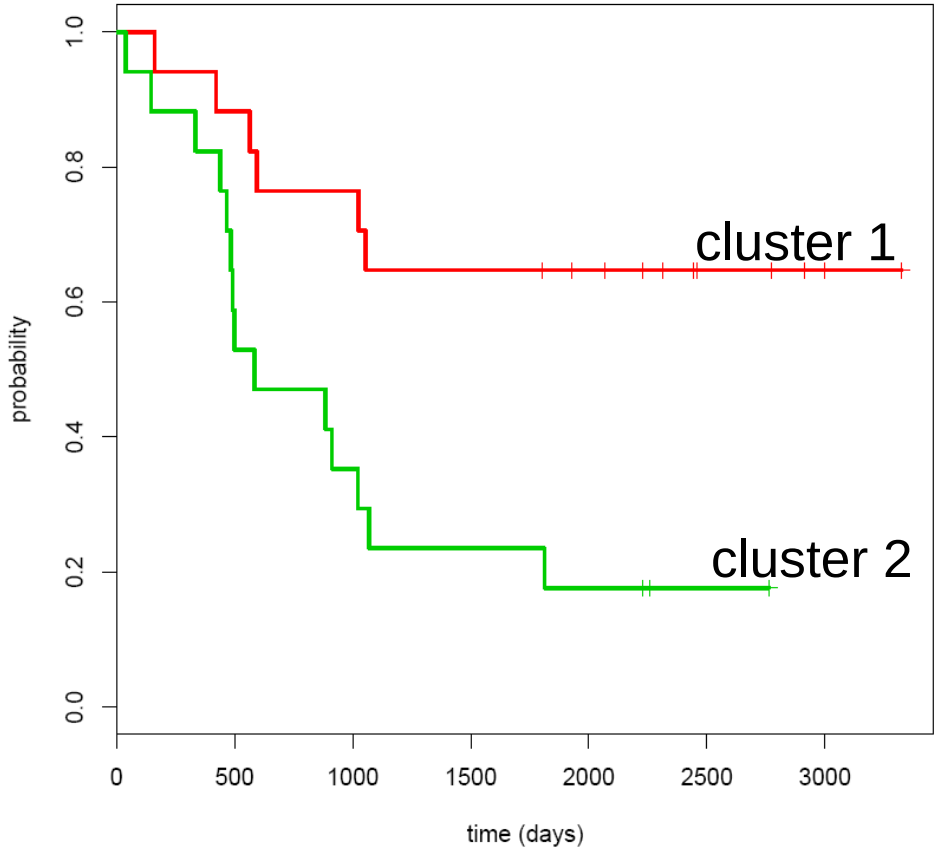
A2 : To obtain confidence in one's choice of *K*, also re-analyze the data with another exploratory technique, e.g. **PCA**, ***K*-means** and investigate whether it yields a grouping resembling the chosen one (with *K* clusters) produced by hierarchical clustering.

Hierarchical clustering

Once decided upon a K, interpret the K clusters, e.g., by means of clinical data.

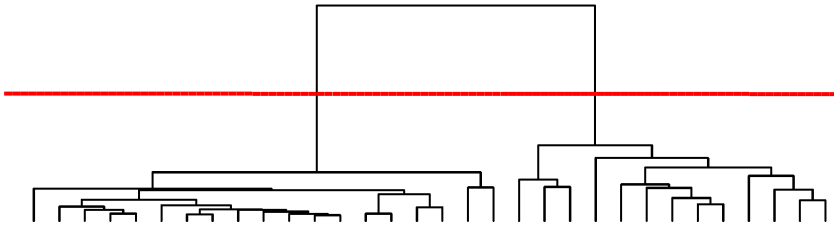


	ER-	ER+
Cluster 1	6	7
Cluster 2	10	8



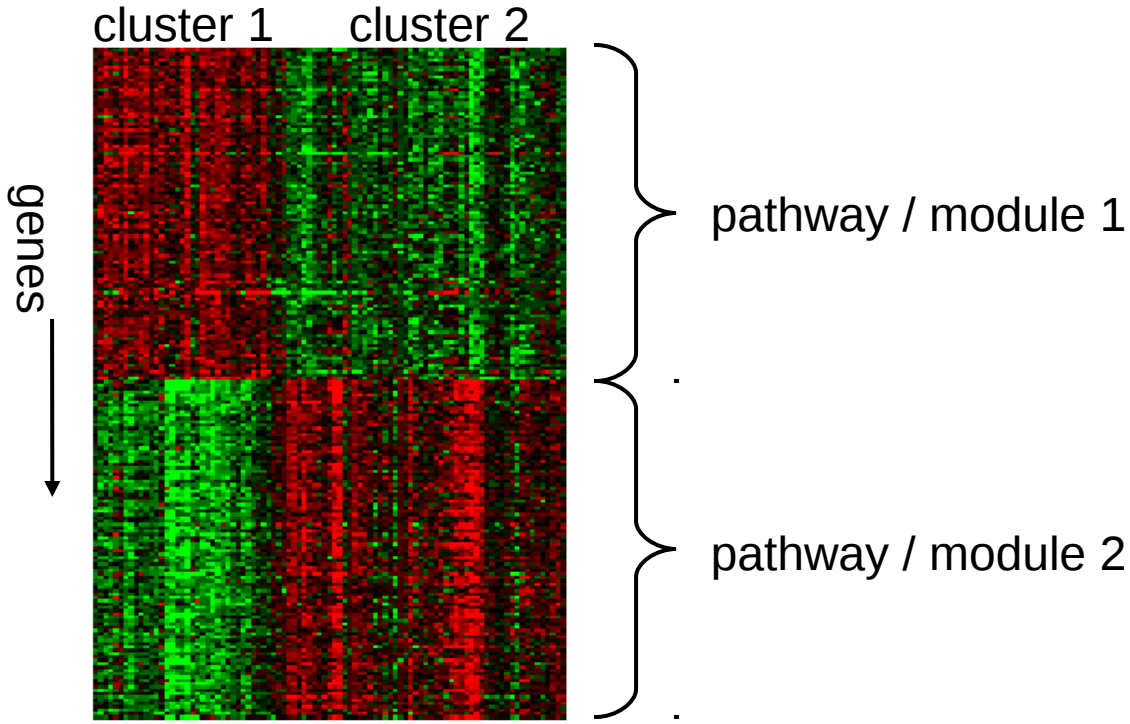
Hierarchical clustering

But also by the gene expression data itself…



genes have also been clustered

only part of full heatmap shown



Hierarchical clustering

(distance, linkage)

Hierarchical clustering

Distance

Central to cluster analysis is the notion of distance (or dissimilarity) between objects being clustered.

Distance measures take on values between 0 and ∞:

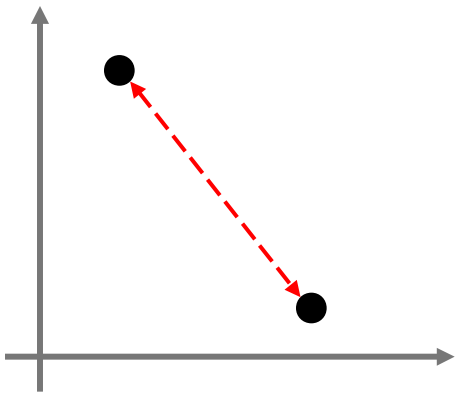
- 0 reflects maximum similarity between two samples,
- ∞ means that two samples are not similar at all, and
- values inbetween indicate various degrees of resemblance.

Hierarchical clustering

Some distance measures (for continuous data)

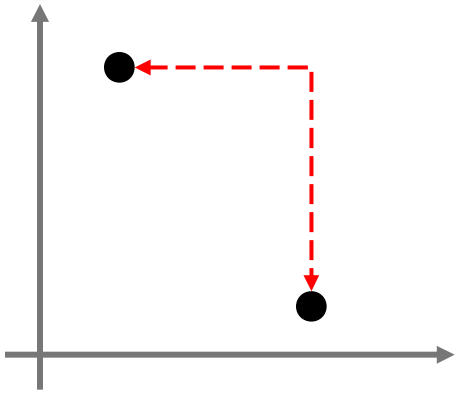
Euclidean distance:

$$D(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sqrt{\sum_{j=1}^{p} (x_{i_1,j} - x_{i_2,j})^2}$$



Manhattan distance:

$$D(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sum_{j=1}^{p} |x_{i_1,j} - x_{i_2,j}|$$



Hierarchical clustering

Expression data

	s_1	s_2
g_1	-1.3	0.1
g_2	-0.1	2.4
g_3	0.4	-0.2
g_4	-0.9	0.4

Manhattan distance:

$| -1.3 – 0.1| + |-0.1 – 2.4| + |0.4 – -0.2| + |-0.9 – 0.4| = 5$

Euclidean distance:

$[(-1.3 – 0.1)^2 + (-0.1 – 2.4)^2 + (0.4 – -0.2)^2 + (-0.9 – 0.4)^2]^{1/2} = 2.97$

Hierarchical clustering

A ***distance matrix*** is a matrix containing the distances between all pairs of samples.

Expression data

	S_1	S_2	S_3
g_1	-1.3	0.1	-0.5
g_2	-0.1	2.4	0.0
g_3	0.4	-0.2	1.8
g_4	-0.9	0.4	-1.1

Distance matrix (euclidean)

	S_1	S_2	S_3
S_1	0.00	2.97	2.57
S_2	2.97	0.00	3.52
S_3	2.57	3.52	0.00

The distance matrix is often denoted with **D** and its elements as $d_{i1,\ i2}$ (the distance between samples i_1 and i_2).

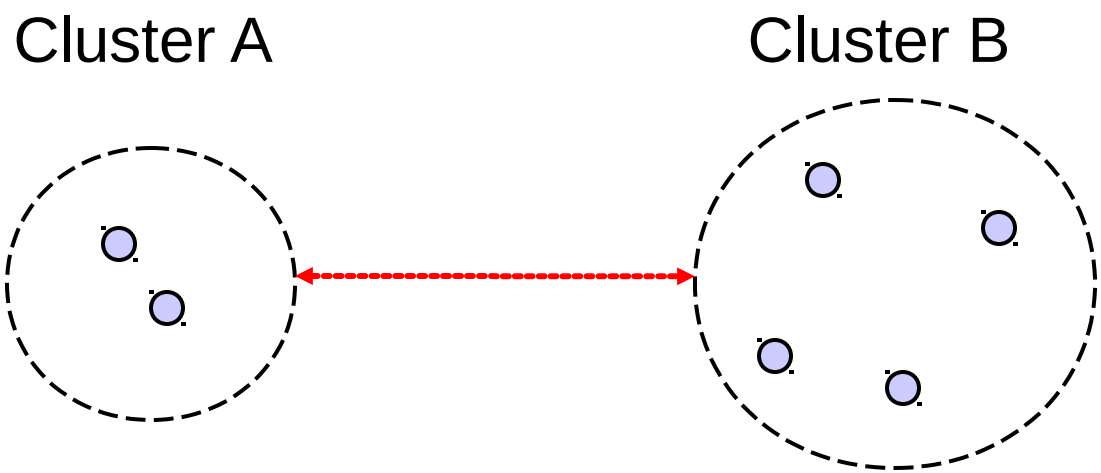
Hierarchical clustering

Distance between clusters

Distance measures are defined between two samples.

In hierarchical clustering, also the distance between groups of samples (clusters) needs to be assessed.

Linkage tells us how to do that.



Hierarchical clustering

Distance between clusters

Single linkage: the minimum distance between samples in different clusters.

$$D(C_1, C_2) \quad = \quad \min_{i_1 \in C_1, i_2 \in C_2} D(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$$

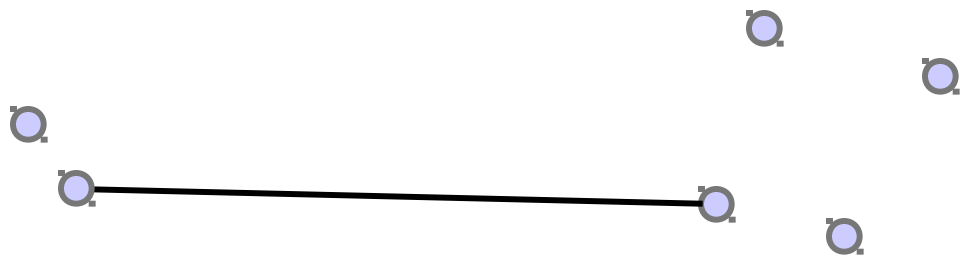
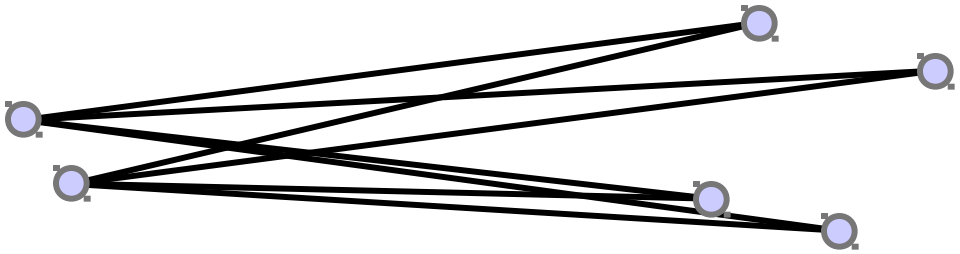
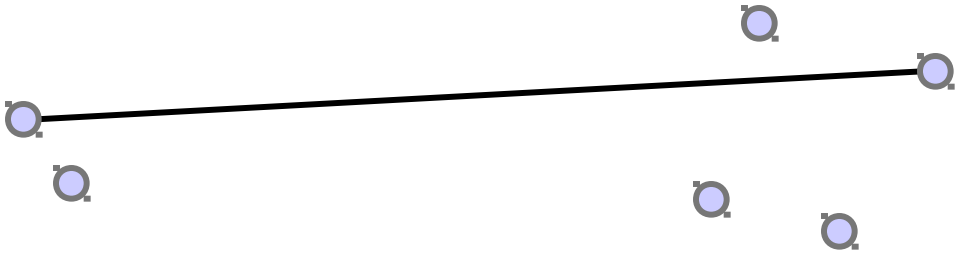
Complete linkage: the maximum distance between samples in different clusters.

$$D(C_1, C_2) \quad = \quad \max_{i_1 \in C_1, i_2 \in C_2} D(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$$

Average linkage: the average distance between all samples in different clusters.

$$D(C_1, C_2) \quad = \quad \frac{1}{n_1 \, n_2} \sum_{i_1 \in C_1} \sum_{i_2 \in C_2} D(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$$

Hierarchical clustering

		Cluster A	Cluster B
Single linkage	Minimum distance		
Average linkage	Average distance		
Complete linkage	Maximum distance		

Hierarchical clustering

Distance matrix (euclidean)

	S_1	S_2	S_3
S_1	0.00	2.97	2.57
S_2	2.97	0.00	3.52
S_3	2.57	3.52	0.00

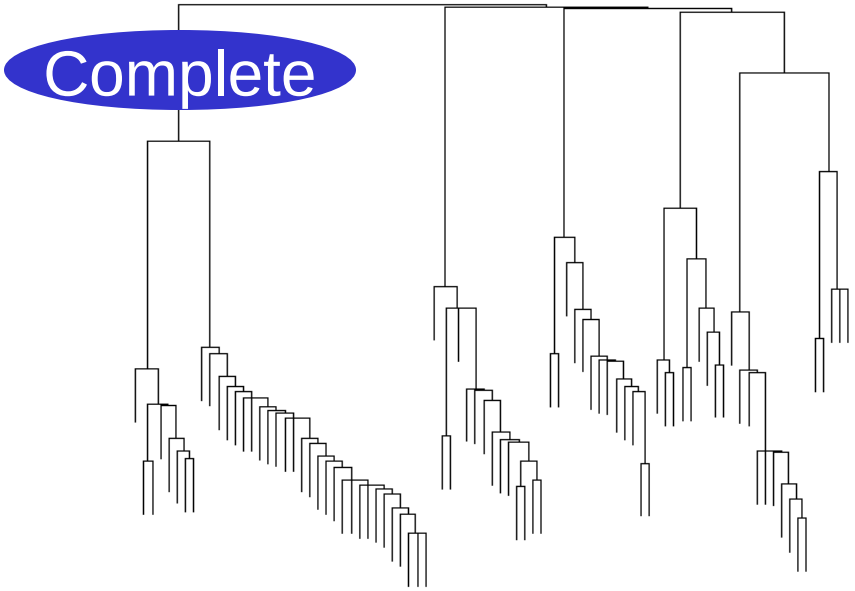
Assume sample 1 and 3 already form a cluster. Then,

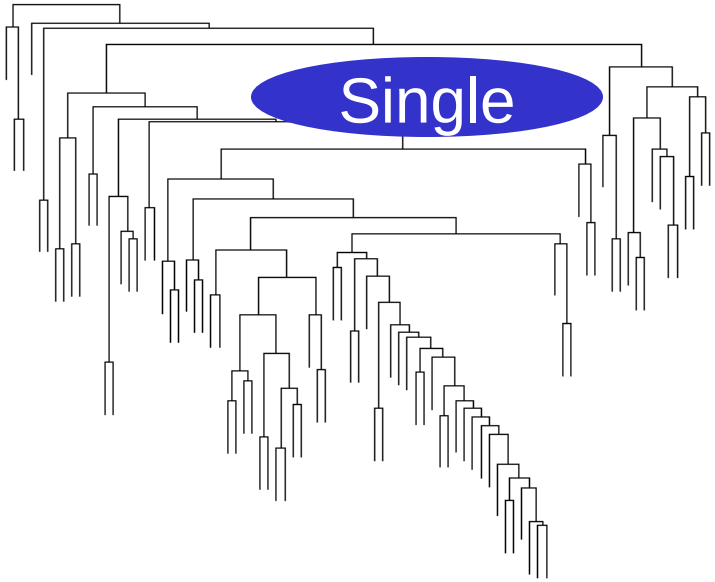
- single linkage : d({1,3}, {2}) = 2.97
- average linkage : d({1,3}, {2}) = 3.25
- complete linkage : d({1,3}, {2}) = 3.52

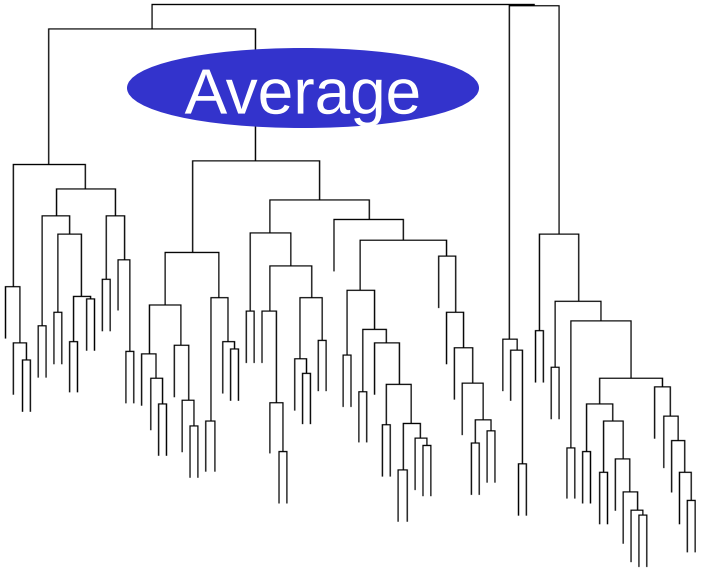
Hierarchical clustering

Effects of linkage

Complete yields a more compact clustering.

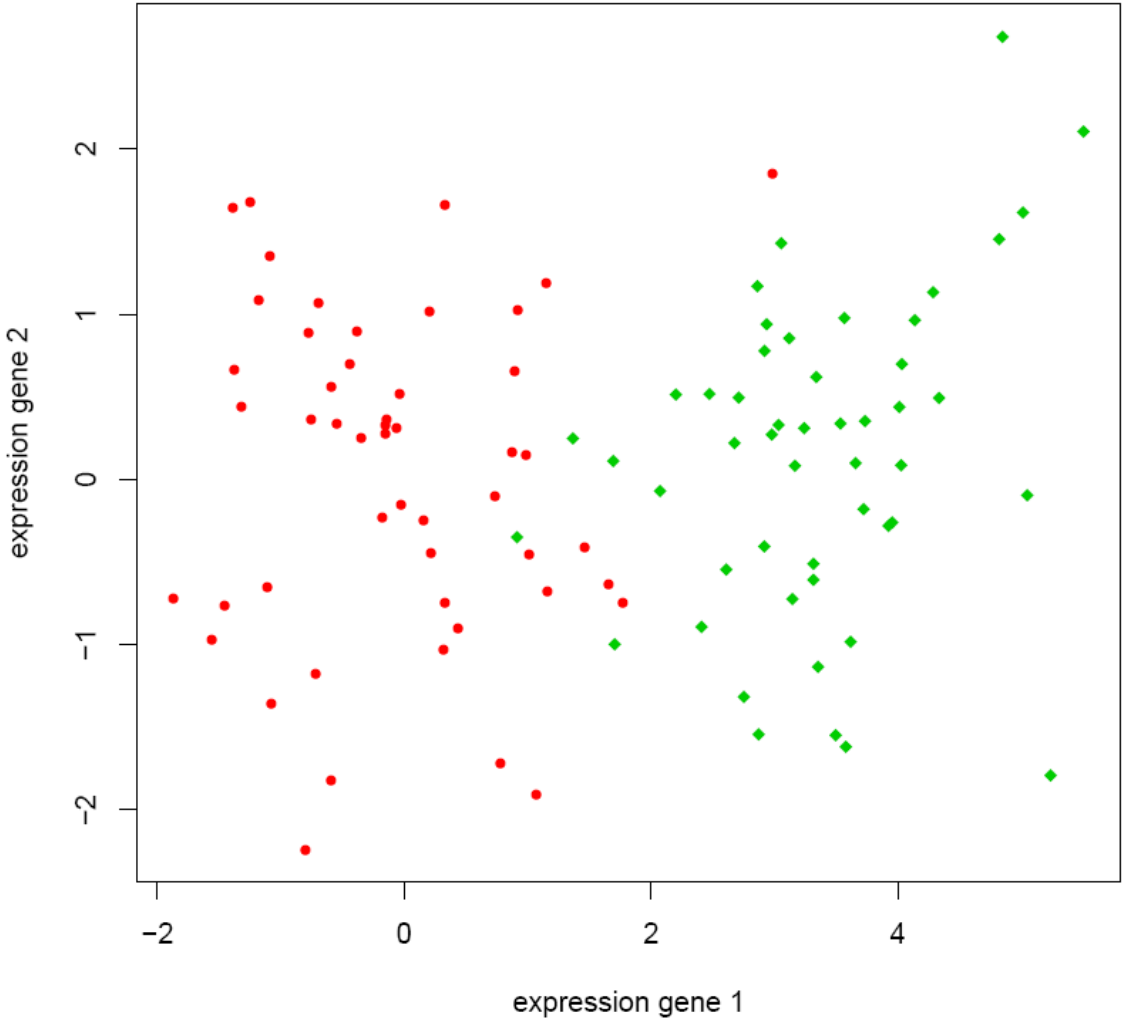






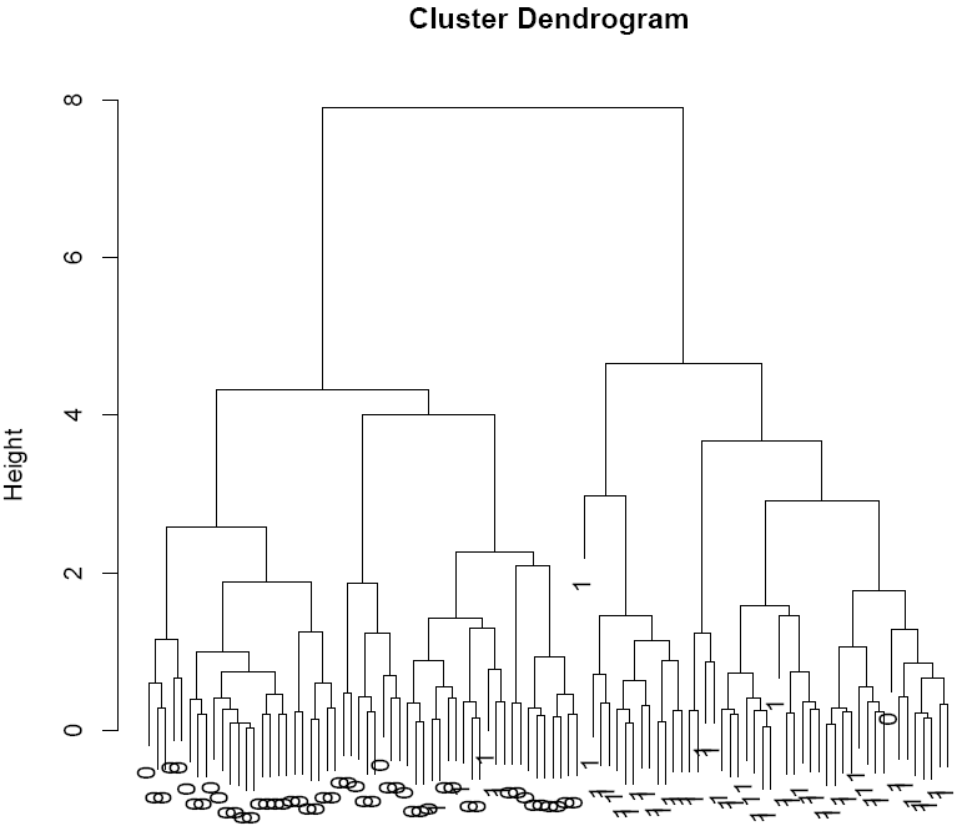
Hierarchical clustering

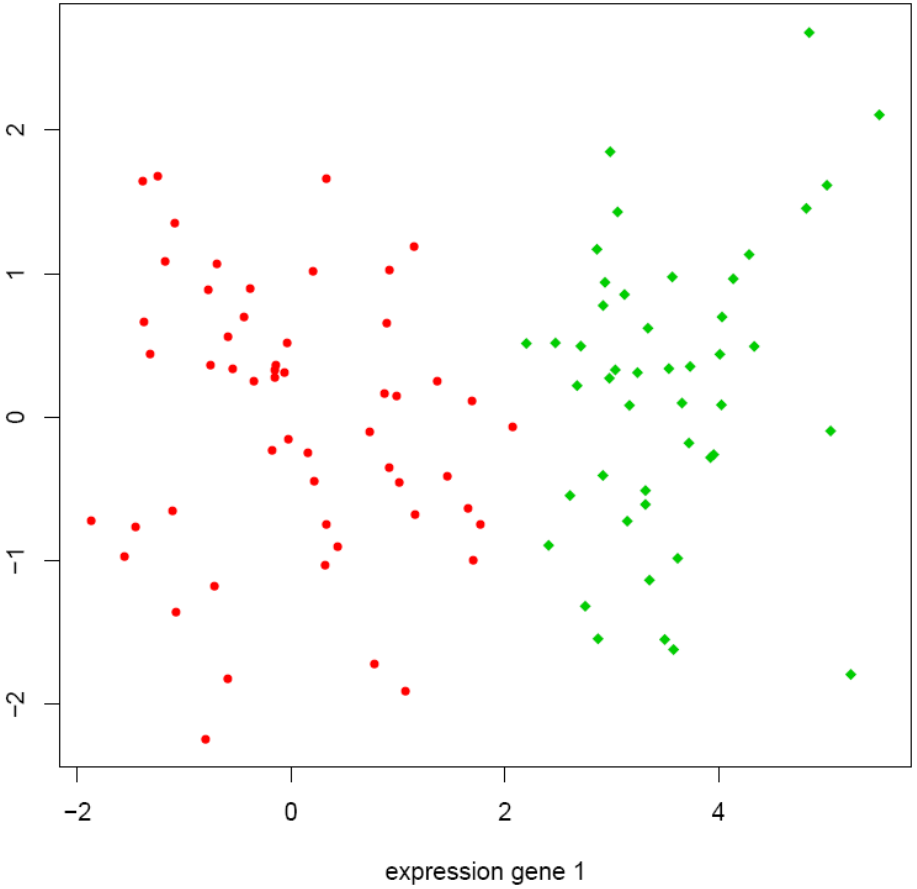
The effect of the spatial distribution of the data (I): groups separate by gene 1



Hierarchical clustering

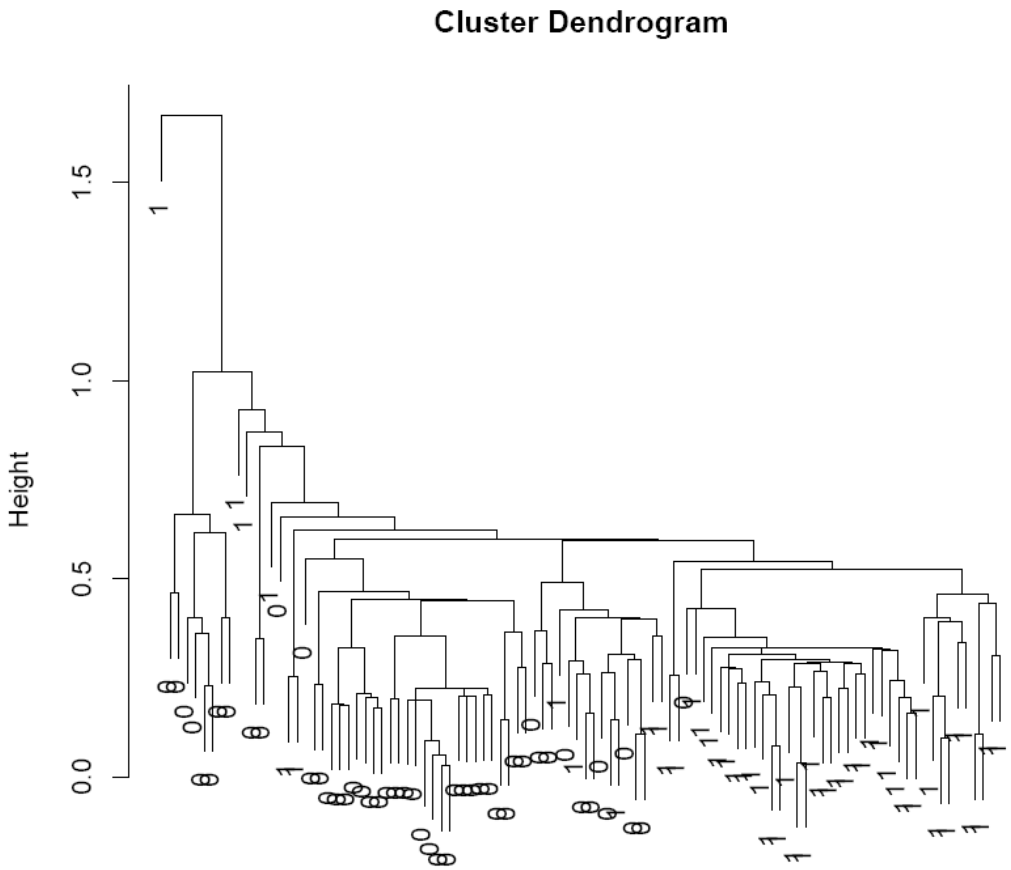
Hierarchical clustering results with complete linkage

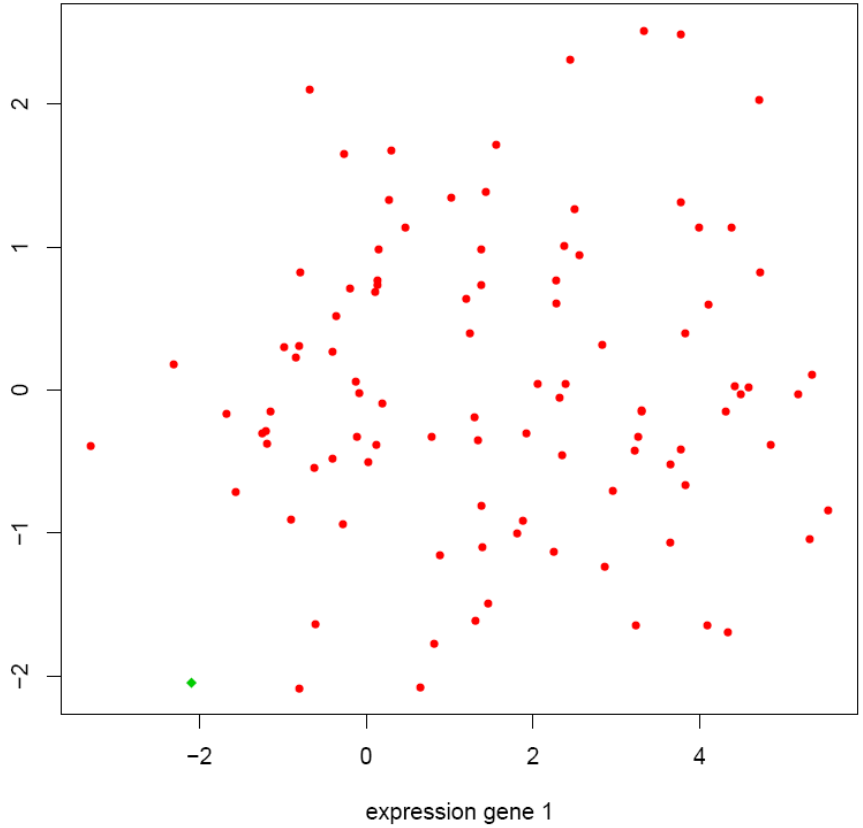




Hierarchical clustering

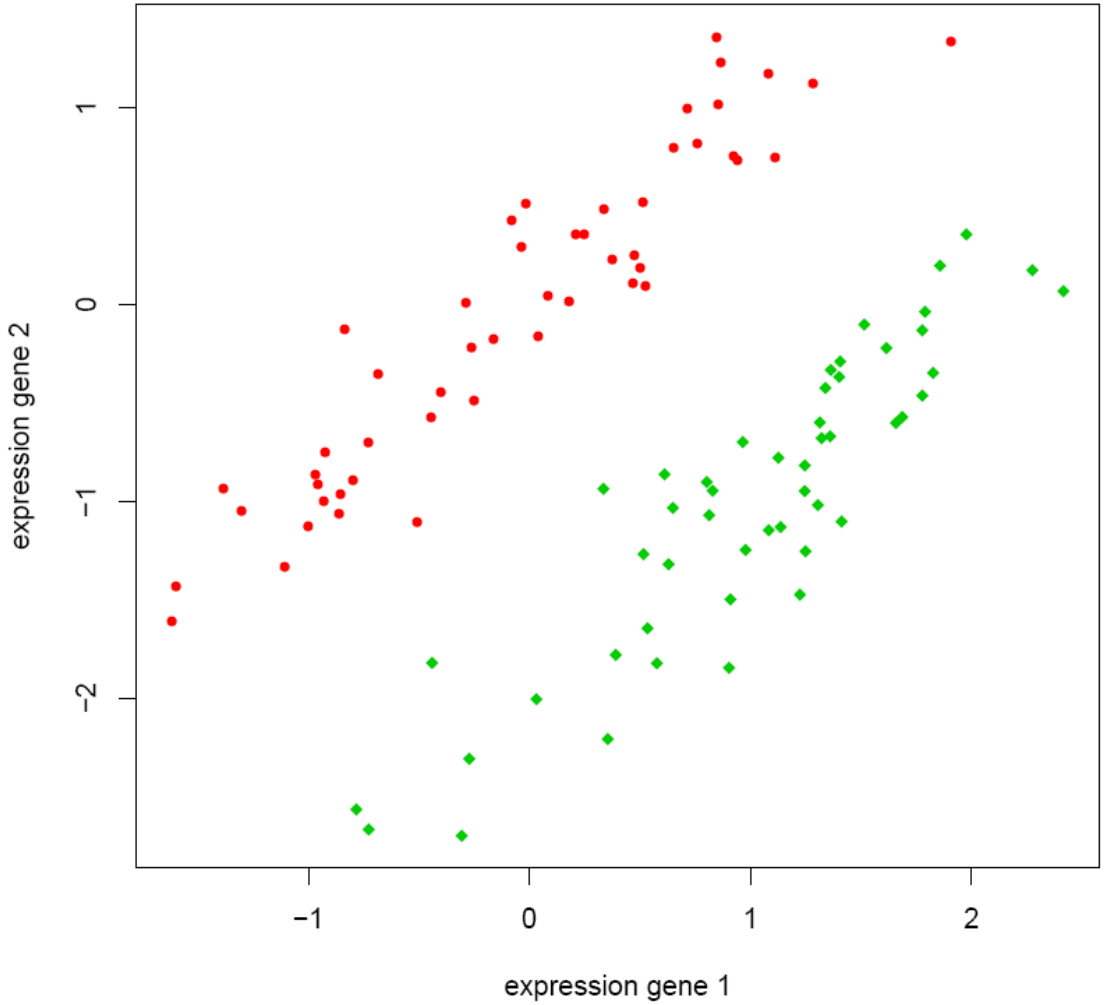
Hierarchical clustering results with single linkage





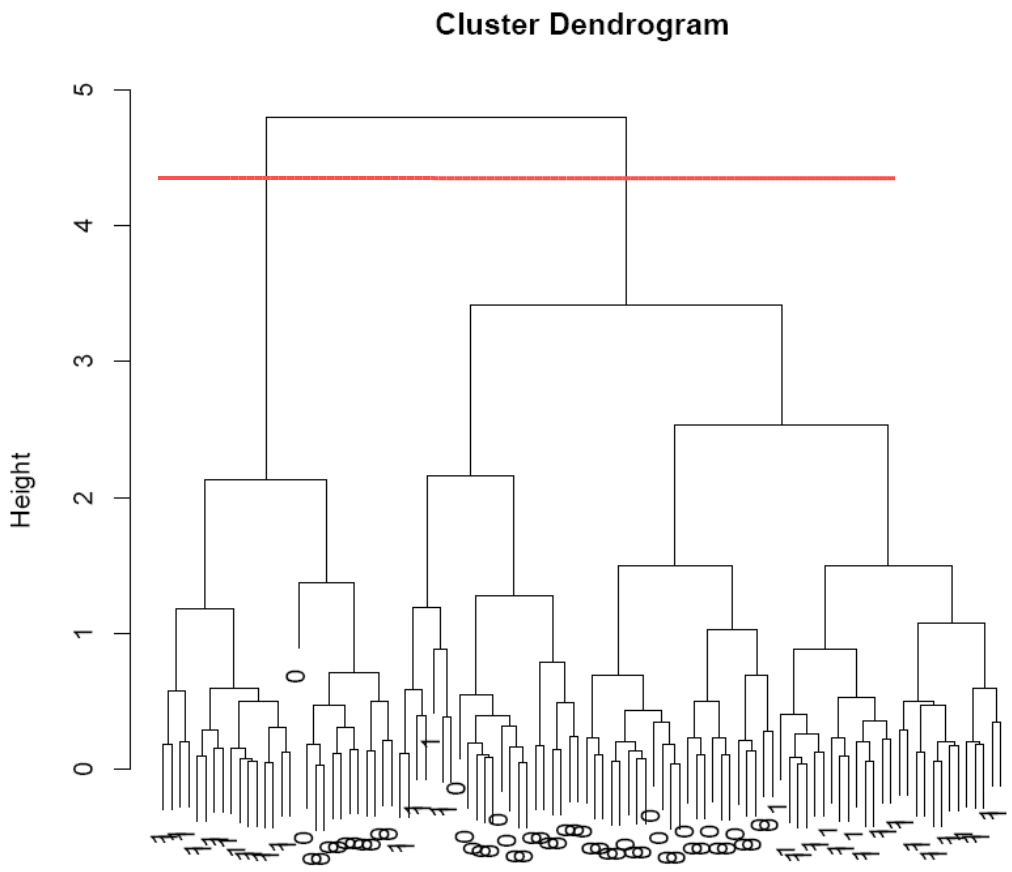
Hierarchical clustering

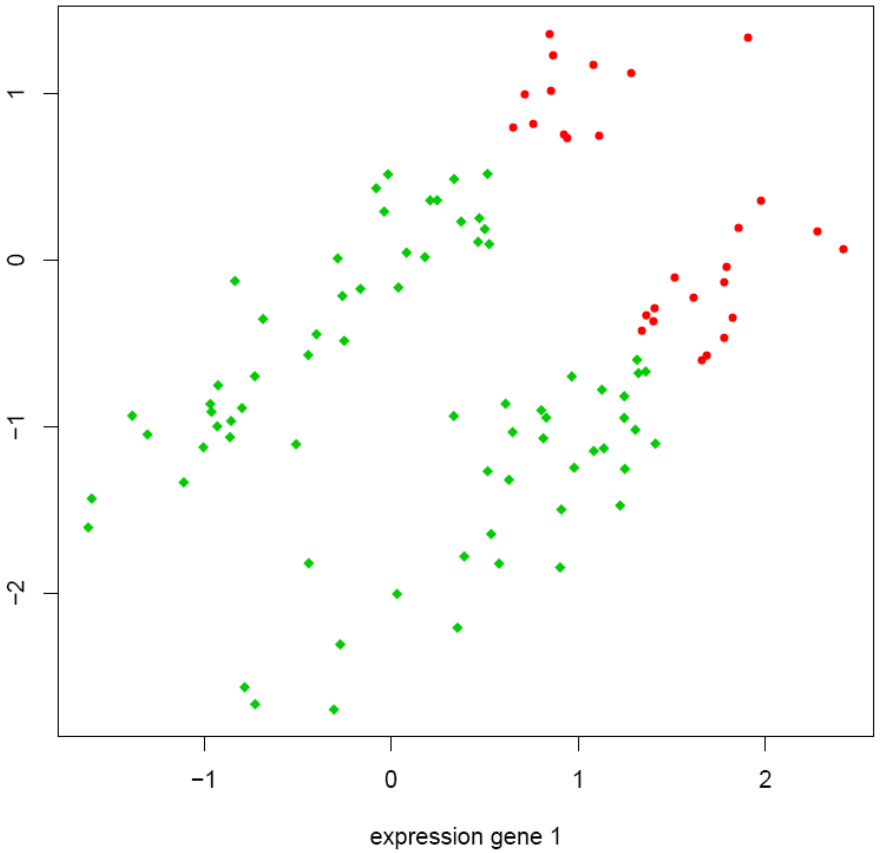
The effect of the spatial distribution of the data (II): groups separate by a combination of gene 1 and 2.



Hierarchical clustering

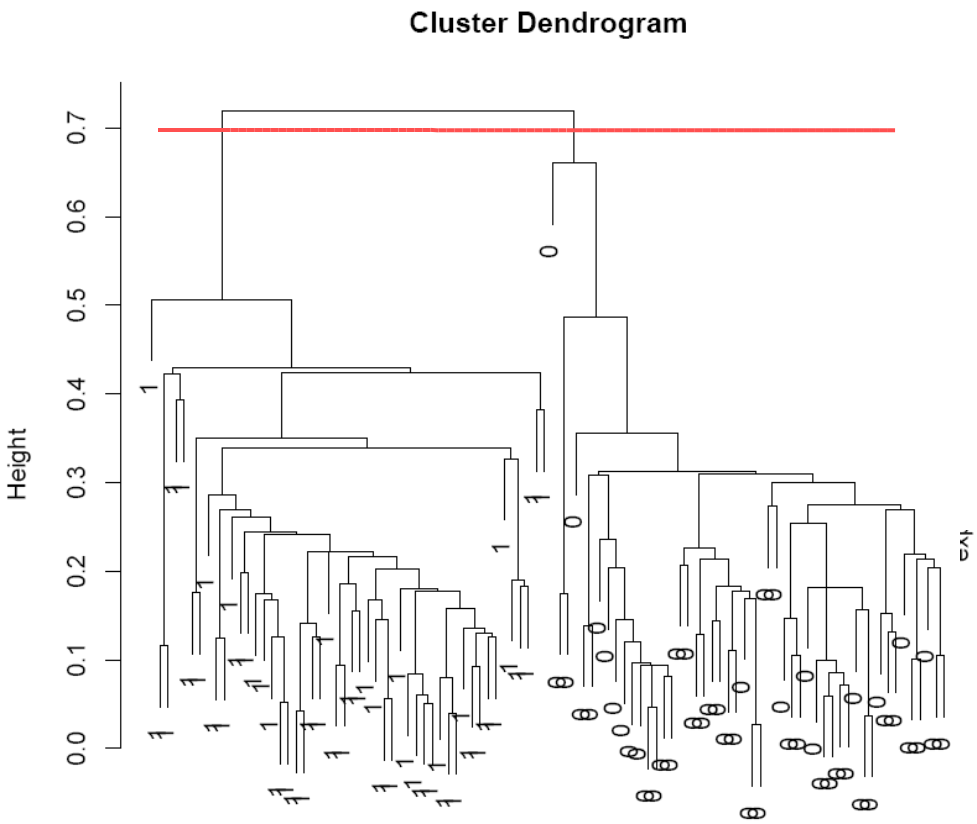
Hierarchical clustering results with complete linkage

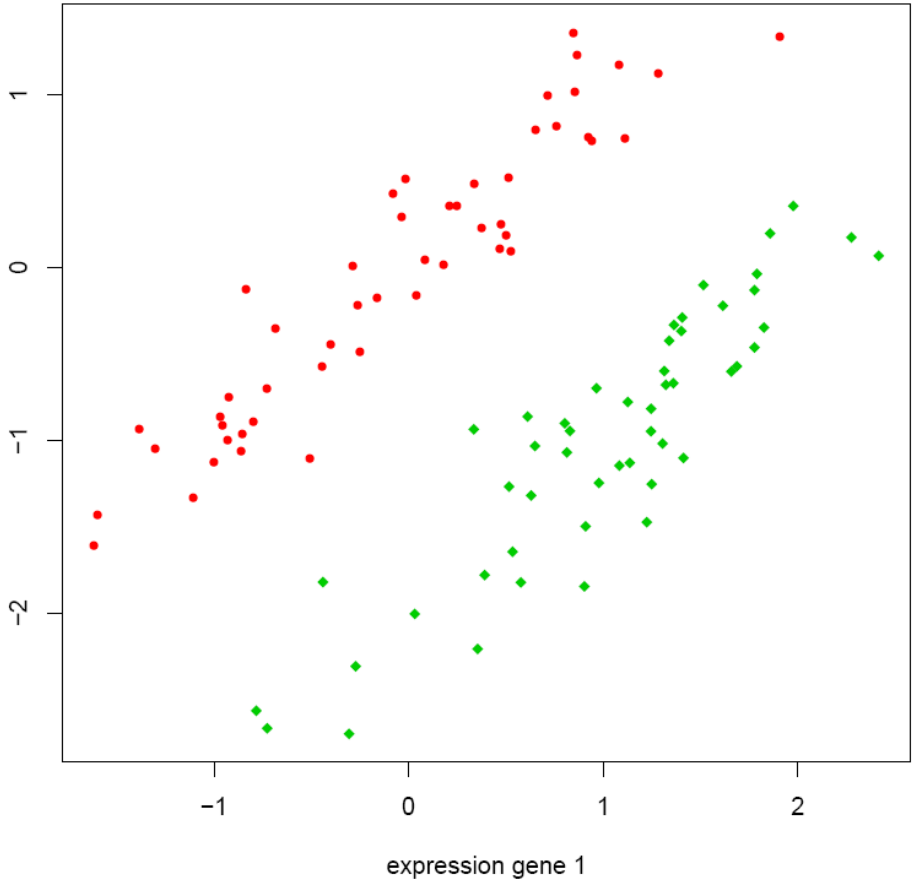




Hierarchical clustering

Hierarchical clustering results with single linkage





Consensus clustering

Consensus clustering

Given a clustering, including the number of clusters *K*.

How does one gain confidence in:

- the number of clusters, and
- the assignment of samples to clusters?

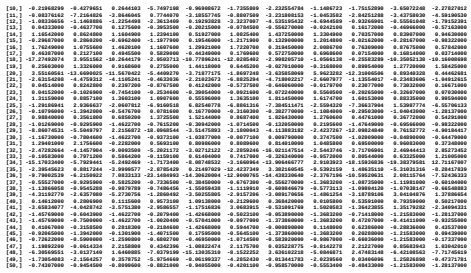
Consensus clustering is a stability analysis of a clustering, when the 'true' grouping is unknown.

Consensus clustering is also used to decide upon *K*.

Consensus clustering

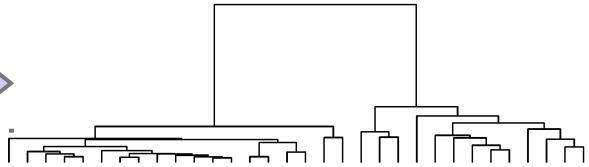
Consensus clustering in a nutshell

data

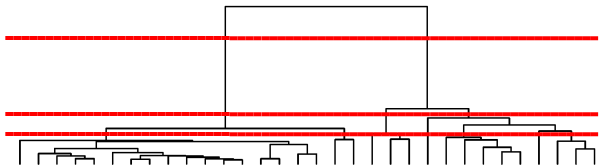


dendrogram

cluster

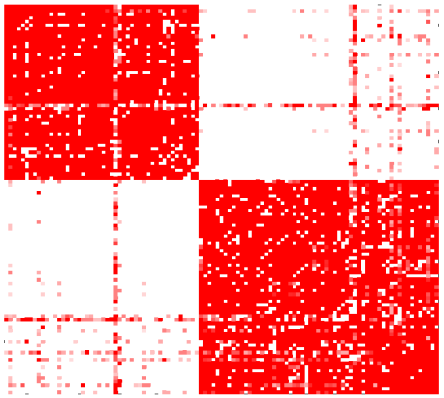


cut dendrogram
at multiple *K*



evaluate
stability of
clustering

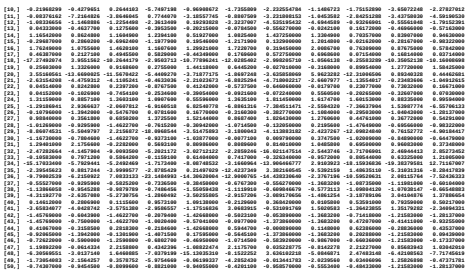
choose *K* with most
stable clustering



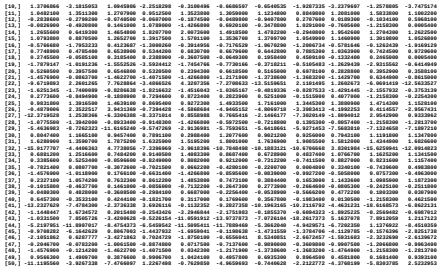
Consensus clustering

Evaluation of clustering stability

original data

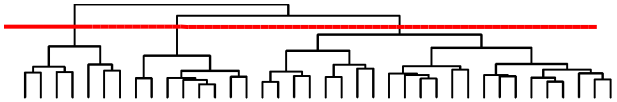


create perturbed versions of the data

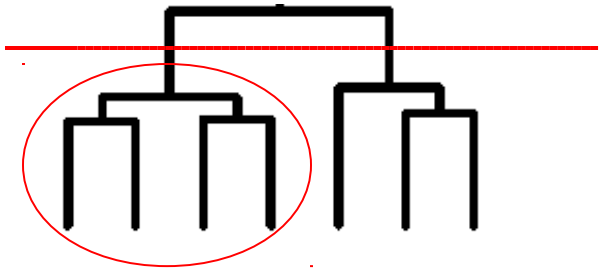


re-cluster

“modified clustering”



asses which samples cluster together



Consensus clustering

Motivation behind consensus clustering

Perturbed versions of the original data allow the assessment of the clustering stability with respect to sampling variability.

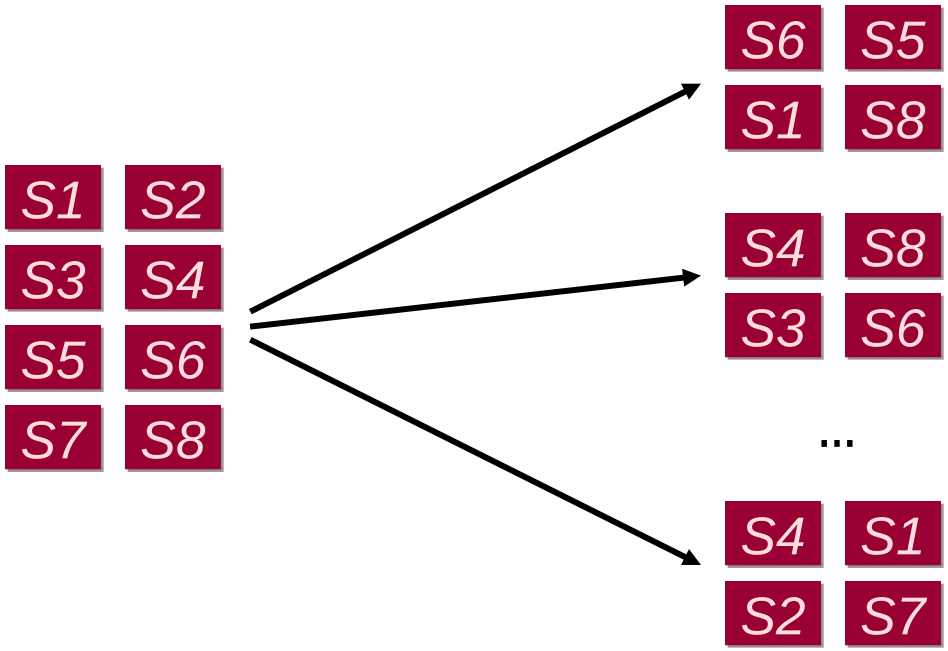
More stable clusterings are believed to be more reliable.

Ways to perturb the data:

- add noise,
- resample features, or
- resample samples.

Consensus clustering

Consensus clustering uses *subsampling without replacement* to perturb the data.

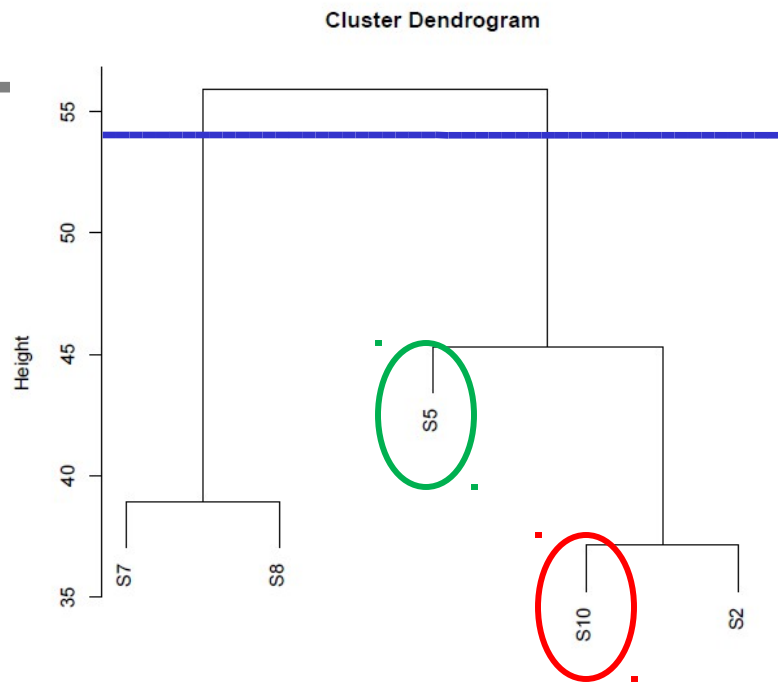


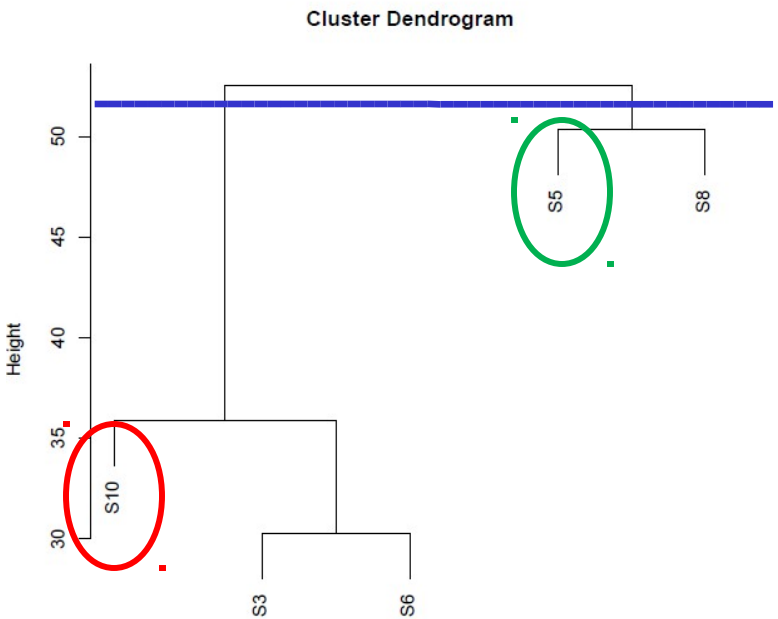
Now cluster with expression data from the selected samples only.

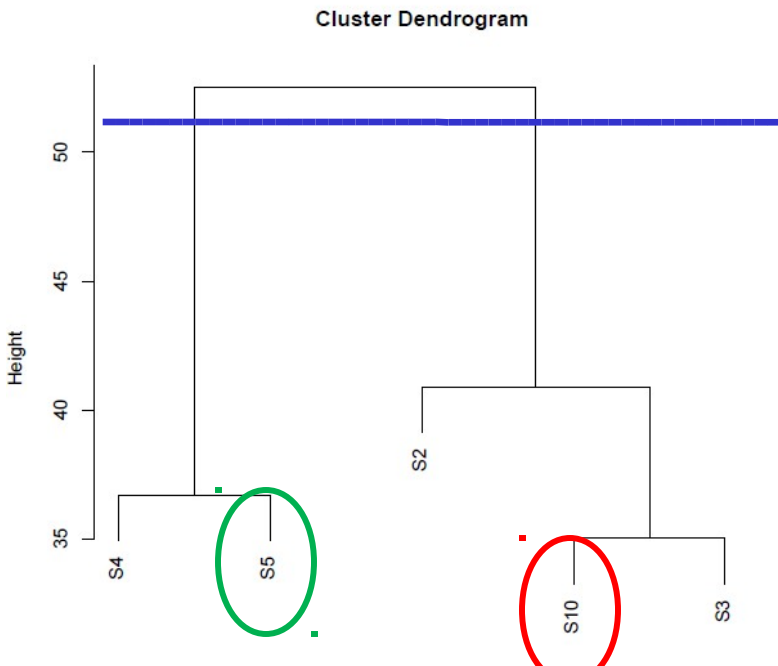
Consensus clustering

Ten samples,
three subsamples,
three clusterings

$K = 2$



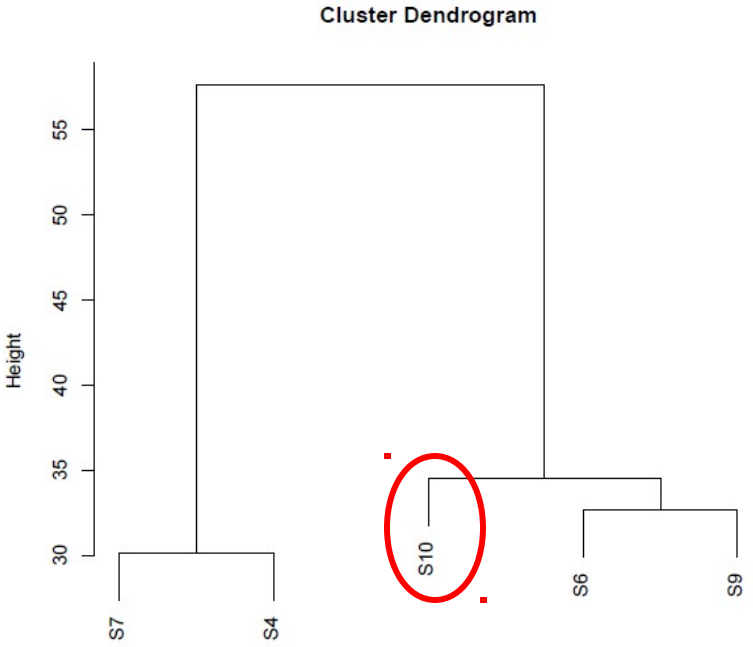


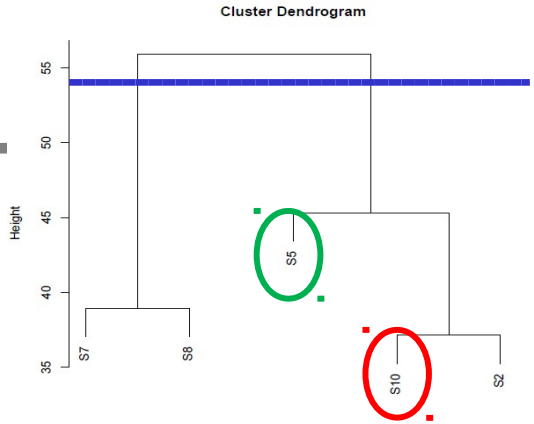


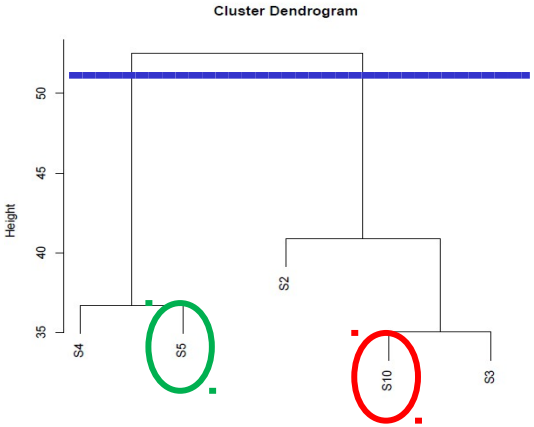
Consensus clustering

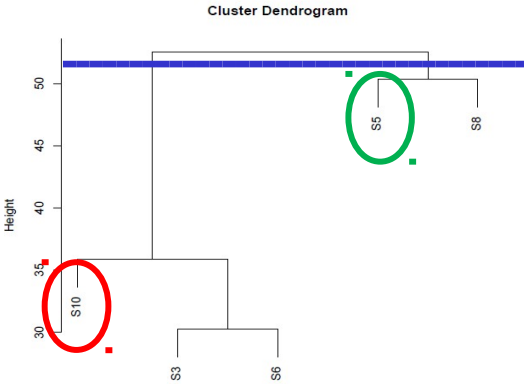
Samples 5 and 10 cluster together in 2 out 3 subsamples.

However, they are not sampled as a pair.









Consensus clustering

The ***consensus matrix M*** quantifies the agreement among clusterings from the perturbed data sets.

An element of **M** is proportion of clusterings in which the two samples are clustered together. Say, **M**[1,2]:

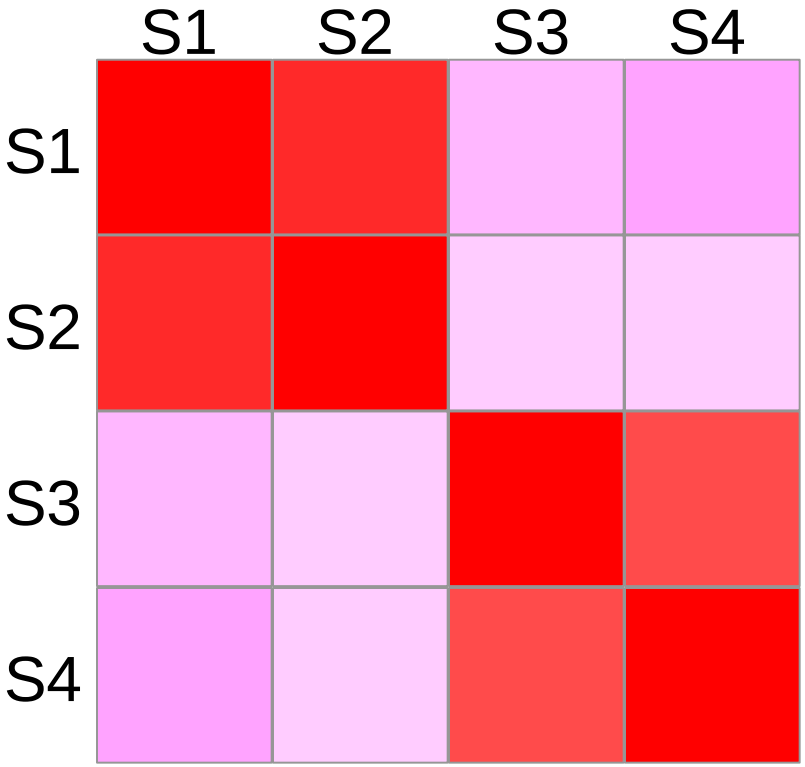
$$\frac{\#\ (\text{S1 \& S2 belong to same cluster})}{\#\ (\text{S1 \& S2 subsampled together})}$$

The consensus matrix is plotted as a heatmap, for which its rows and columns are ordered in accordance with the original clustering.

Consensus clustering

The consensus clustering heatmap

	S1	S2	S3	S4
S1	100%	97%	27%	32%
S2	97%	100%	18%	14%
S3	27%	18%	100%	83%
S4	32%	14%	83%	100%

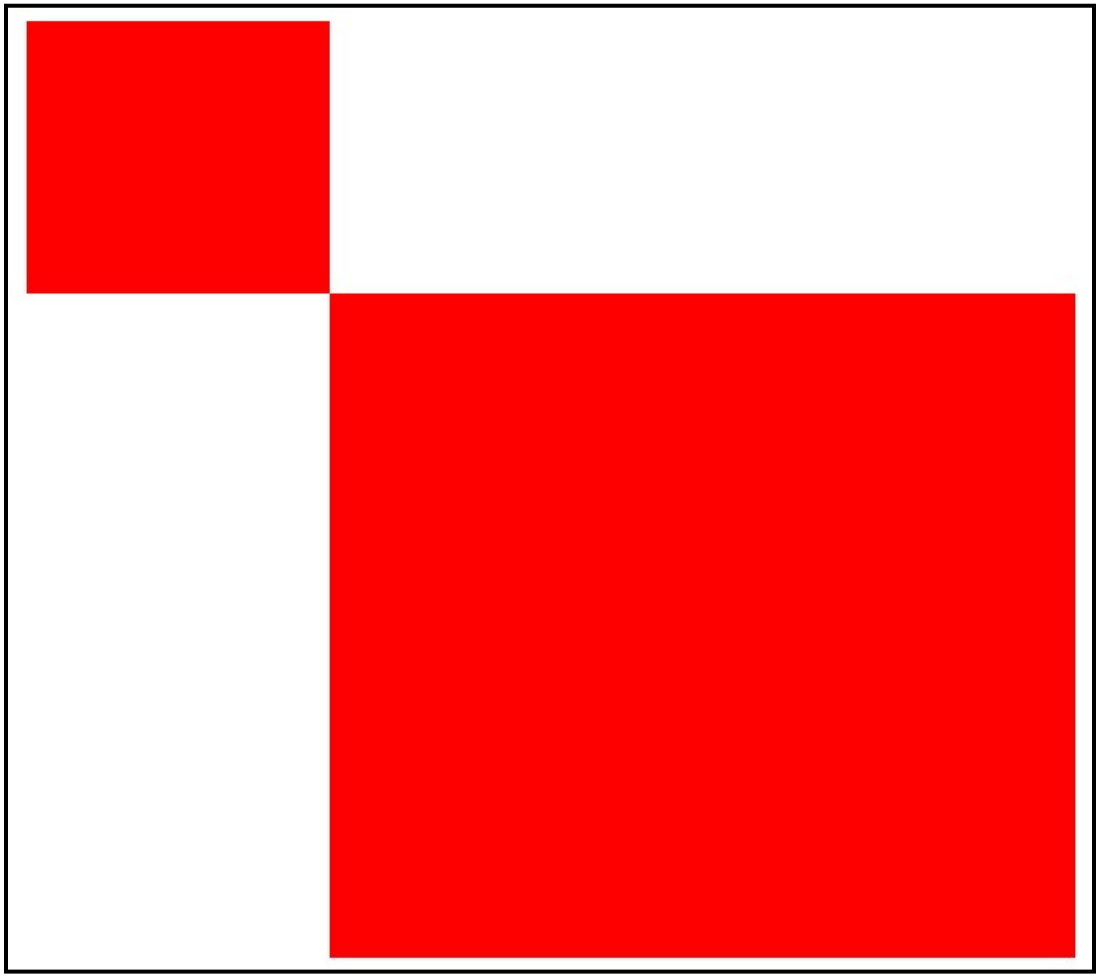


color legend

0%	20%	40%	60%	80%	100%

Consensus clustering

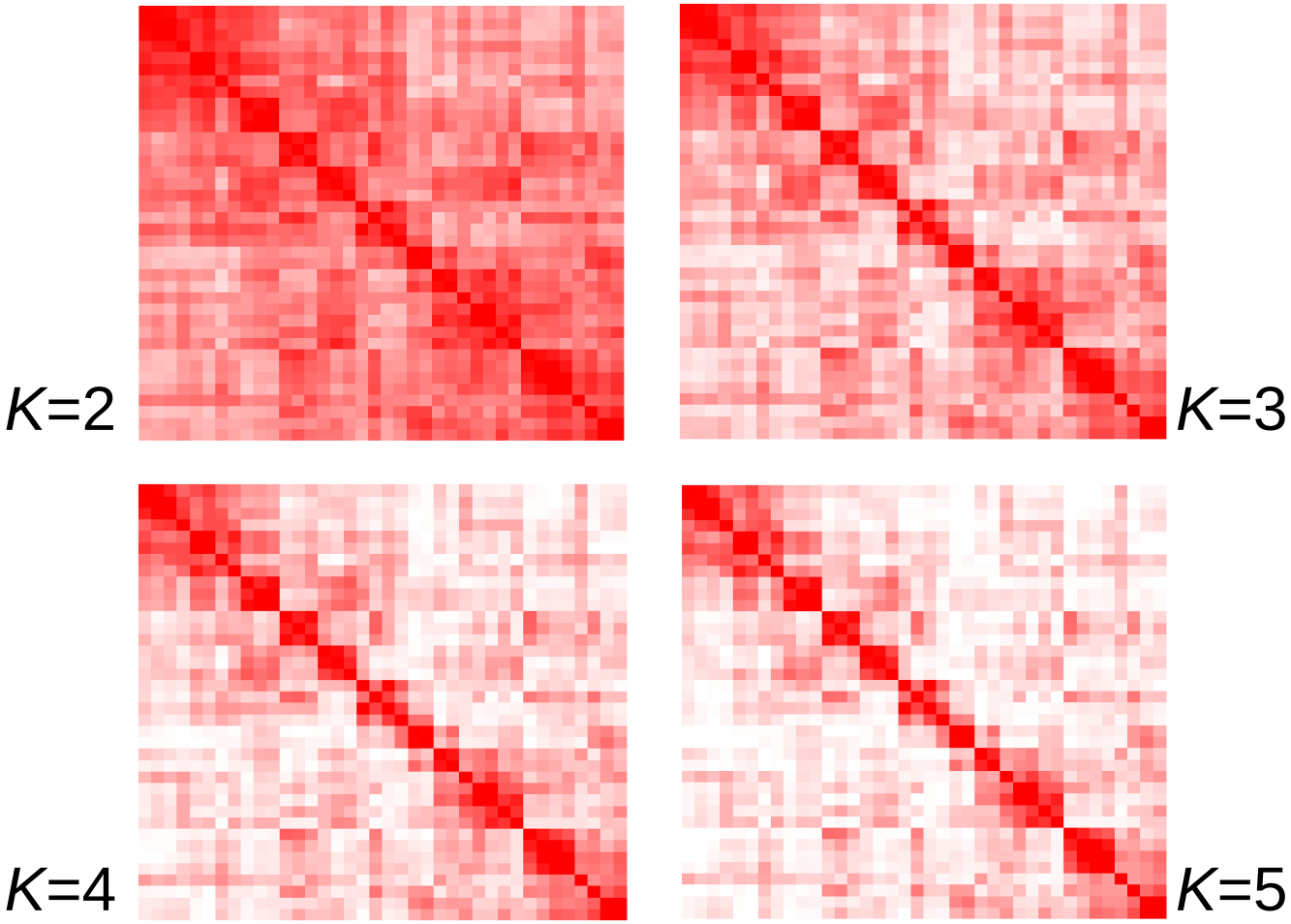
A heatmap characterized by red blocks along the diagonal on a white background reflects perfect consensus.



K=2

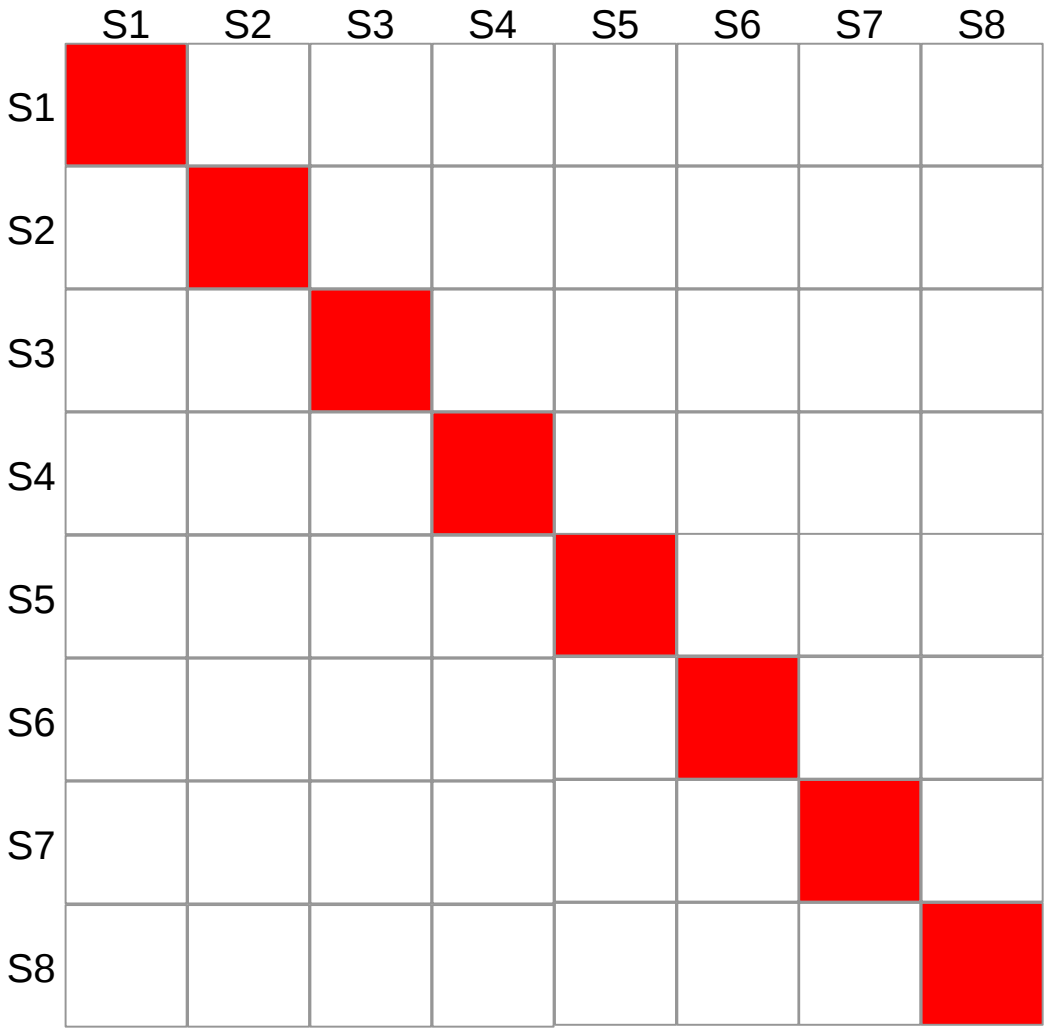
Consensus clustering

For random data (n=38, p=100, $X_{ij} \sim N(0,1)$):



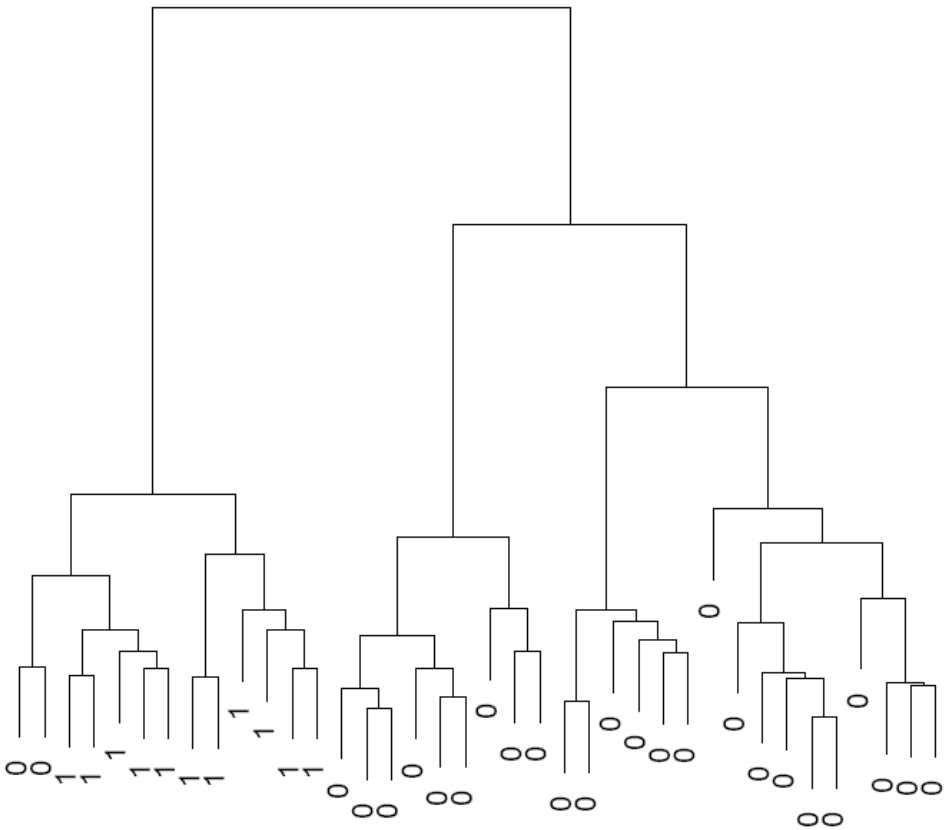
Consensus clustering

The consensus clustering heatmap when K = n



Consensus clustering

For the leukemia data of Golub *et al*. (1999):

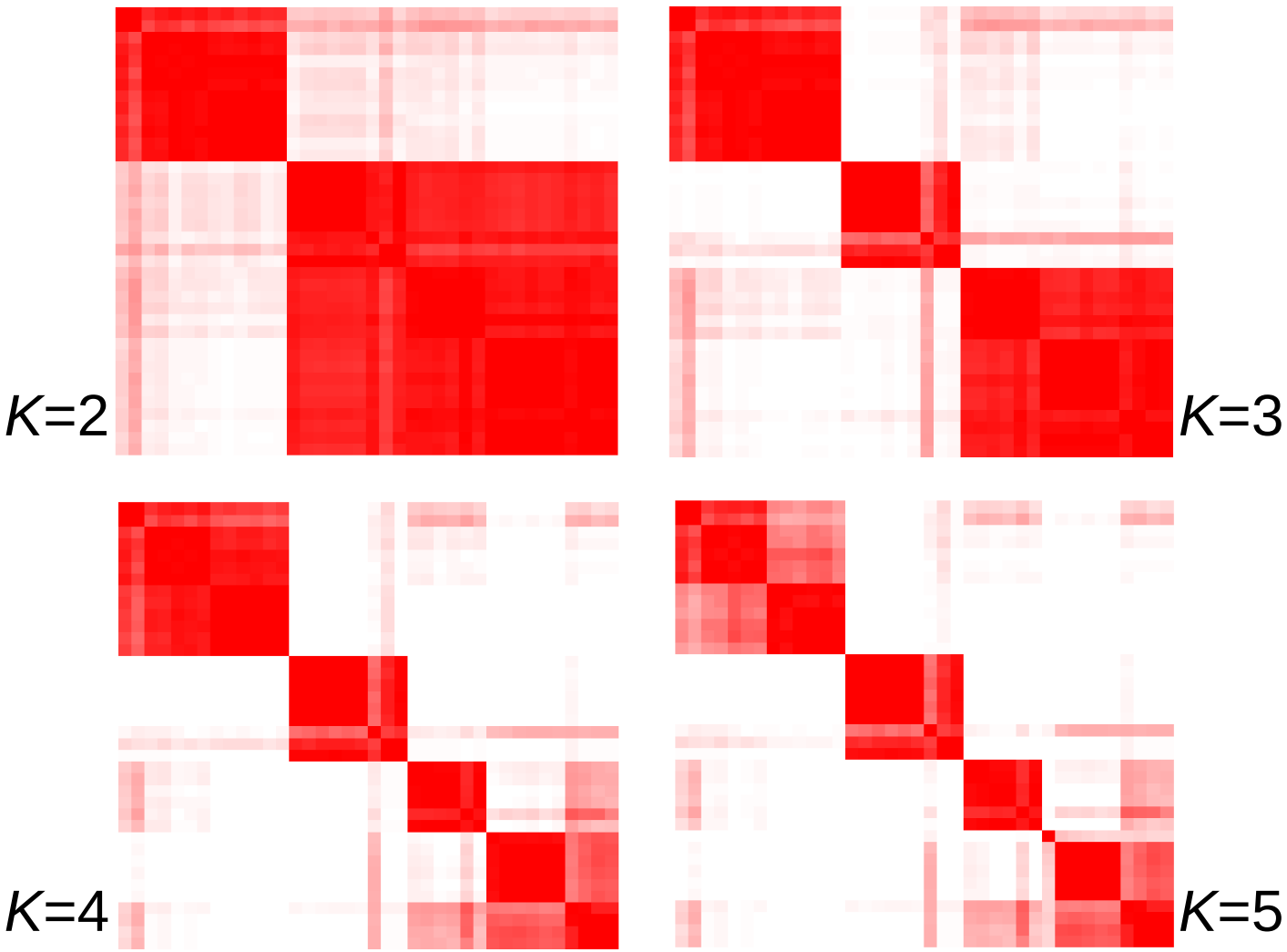


11 AML
8 ALL (T-lineage)
19 ALL (B-lineage)

ALL = 0
AML = 1

Consensus clustering

For the leukemia data of Golub *et al*. (1999):



Consensus clustering

A cluster's consensus index:

$$m(k) = \frac{2}{n_k(n_k-1)} \sum_{i,j \in C_k, i<j} (\mathbf{M})_{ij}$$

the average consensus between all samples pairs belonging to the same cluster.

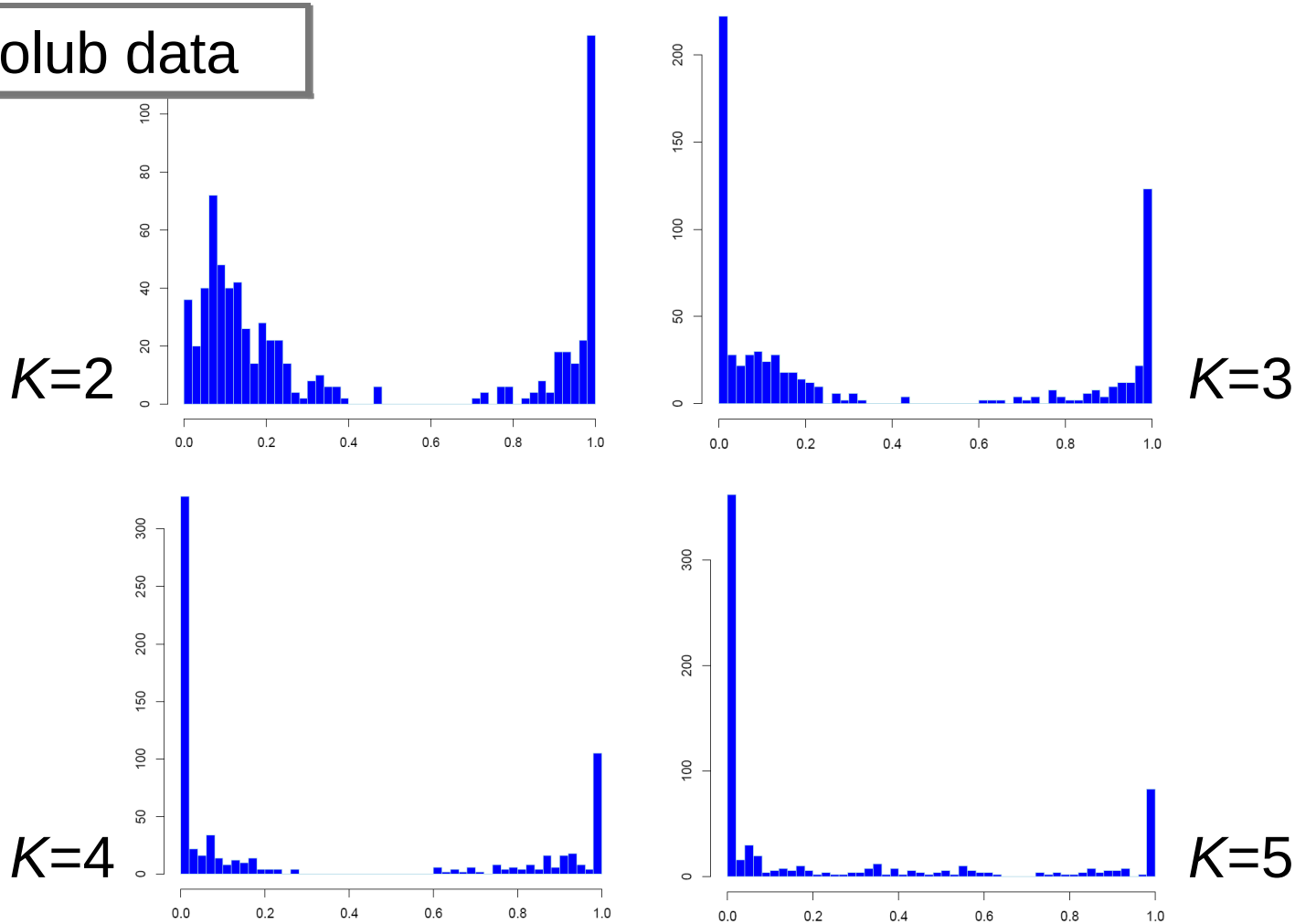
$$\mathbf{M} = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,n_k} & \cdots & \cdots \\ m_{2,1} & m_{2,2} & & m_{2,n_k} & \cdots & \cdots \\ \vdots & & \ddots & \vdots & & \\ m_{n_k,1} & m_{n_k,2} & \cdots & m_{n_k,n_k} & \cdots & \cdots \\ \vdots & \vdots & & \vdots & \ddots & \\ \vdots & \vdots & & \vdots & & \ddots \end{pmatrix}$$

Index is used to evaluate individual clusters.

Consensus clustering

Ideally, the histogram of the consensus indices is concentrated at 0 and 1. This provides another diagnostic.

Golub data



K-means clustering

K-means clustering

_K_-means

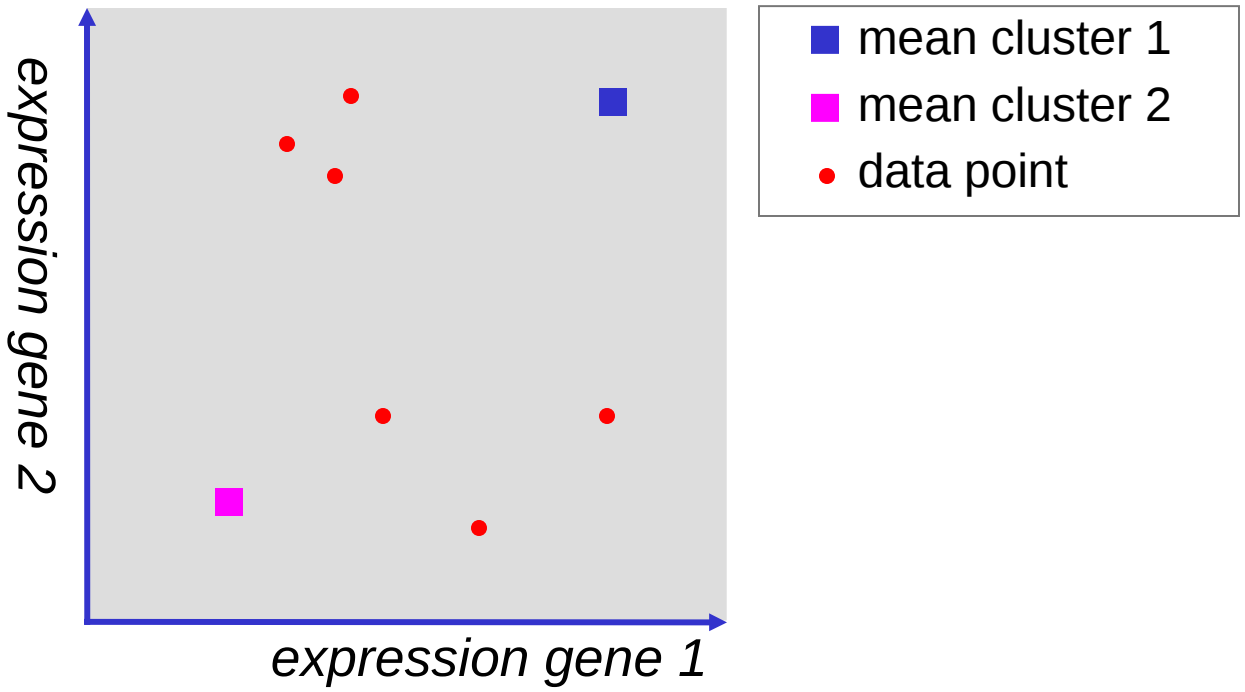
The *K*-means method is a clustering algorithm that assigns samples to *K* clusters.

Each cluster is represented by its mean, a ‘representative’ of the samples in the cluster.

- *K* is chosen before hand. Use consensus clustering to find the optimal *K*.
- In the case of clustering of samples on the basis of their gene expression, the resulting cluster mean is the typical expression profile for the samples in the cluster.

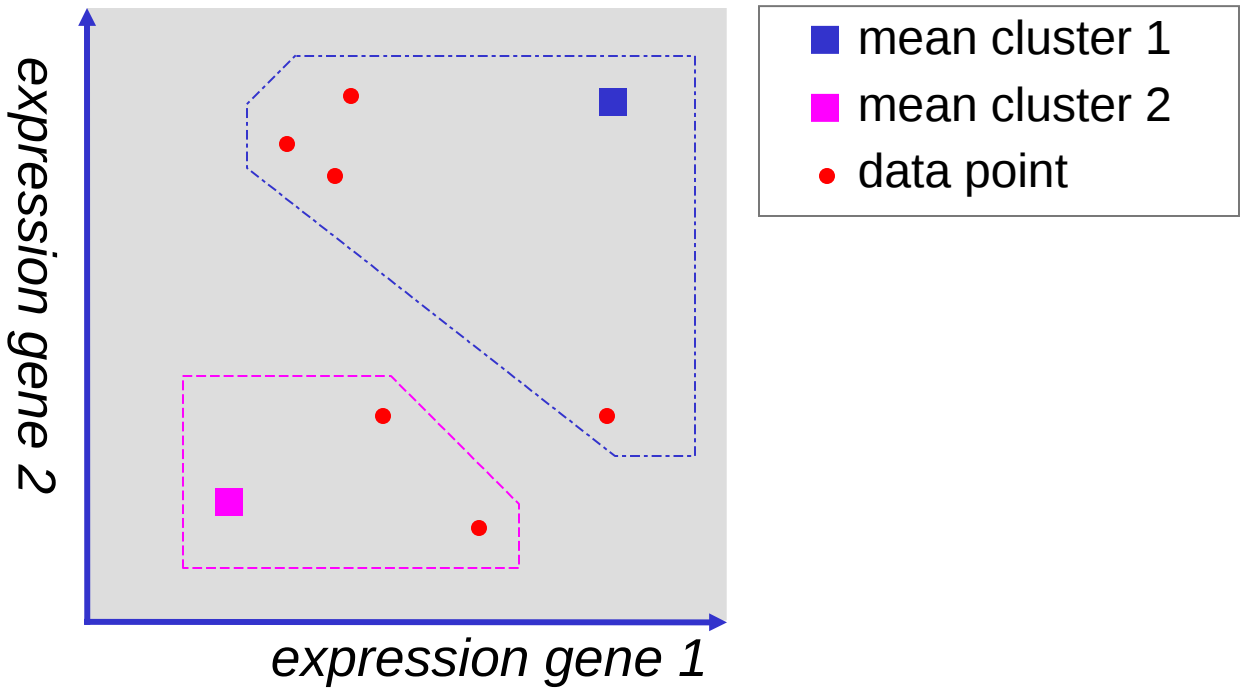
K-means clustering

Suppose *K*=2.
Randomly assign means to each cluster.



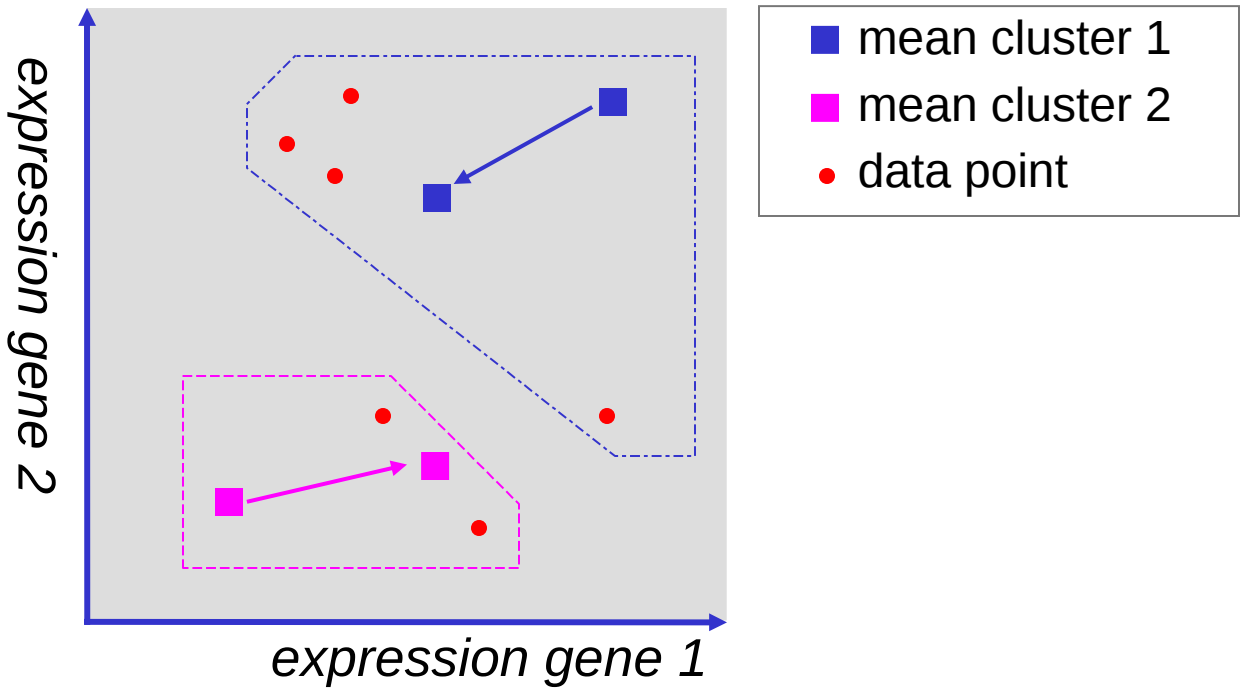
K-means clustering

For each sample determine the closest mean.



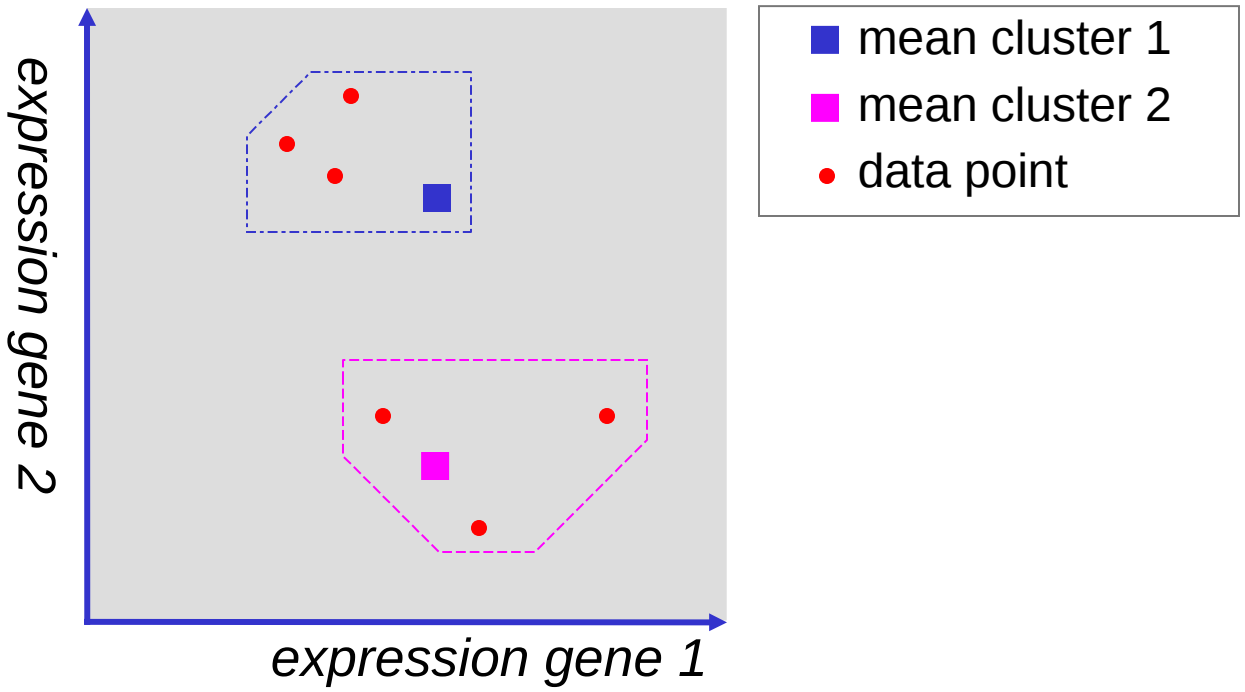
K-means clustering

Calculate means for each cluster.
Shift means to locations.



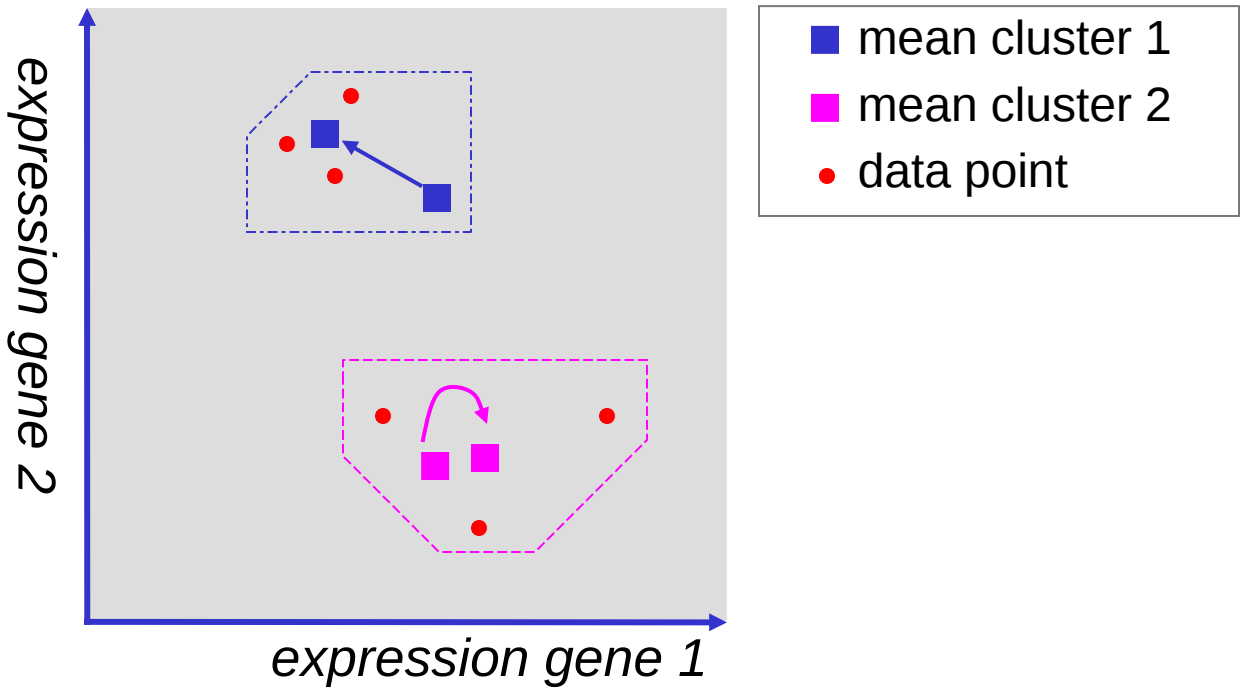
K-means clustering

For each sample determine the closest mean.



K-means clustering

Re-calculate means for each cluster.
Shift means to new locations.
Repeat this until means do not change.



K-means clustering

Algorithm

Step 1: Generate randomly *K* cluster means: $\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(K)}$

Step 2: Assign sample *i* to the cluster with the nearest mean:

$$\hat{k}^{(i)} = \arg \min_{k=1,\ldots,K} \{d(\mathbf{m}^{(k)}, \mathbf{X}_i)\}$$

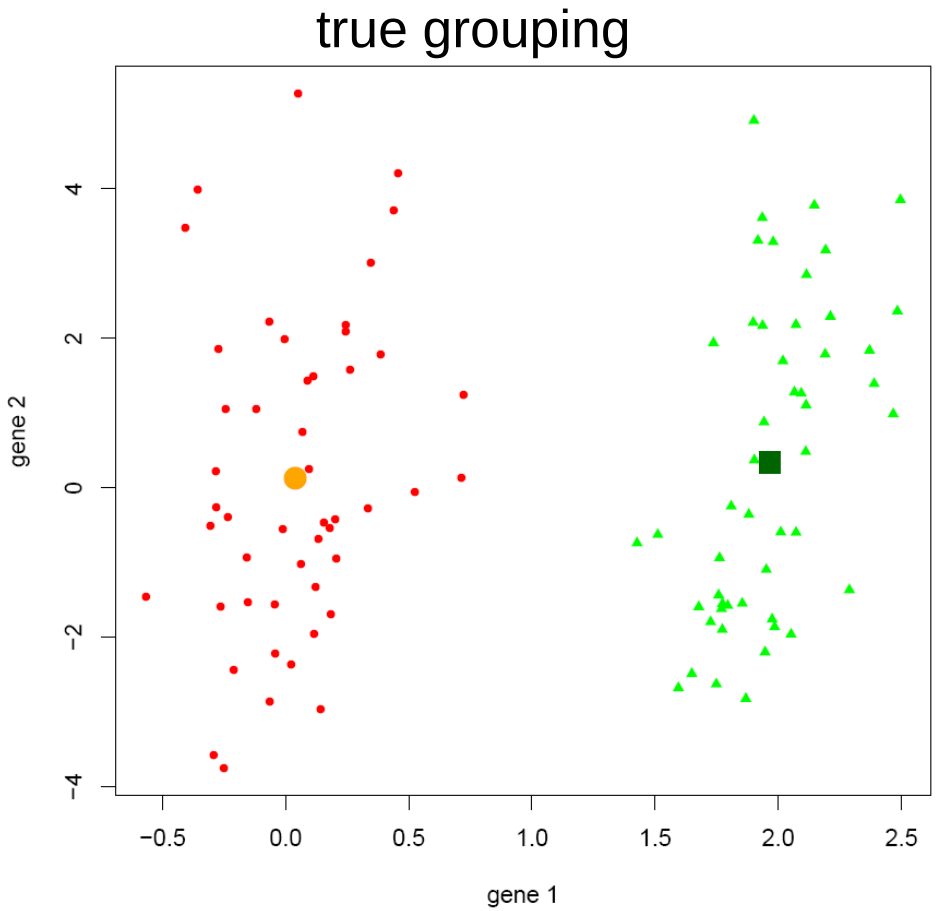
Step 3: Update the cluster means to latest assignment:

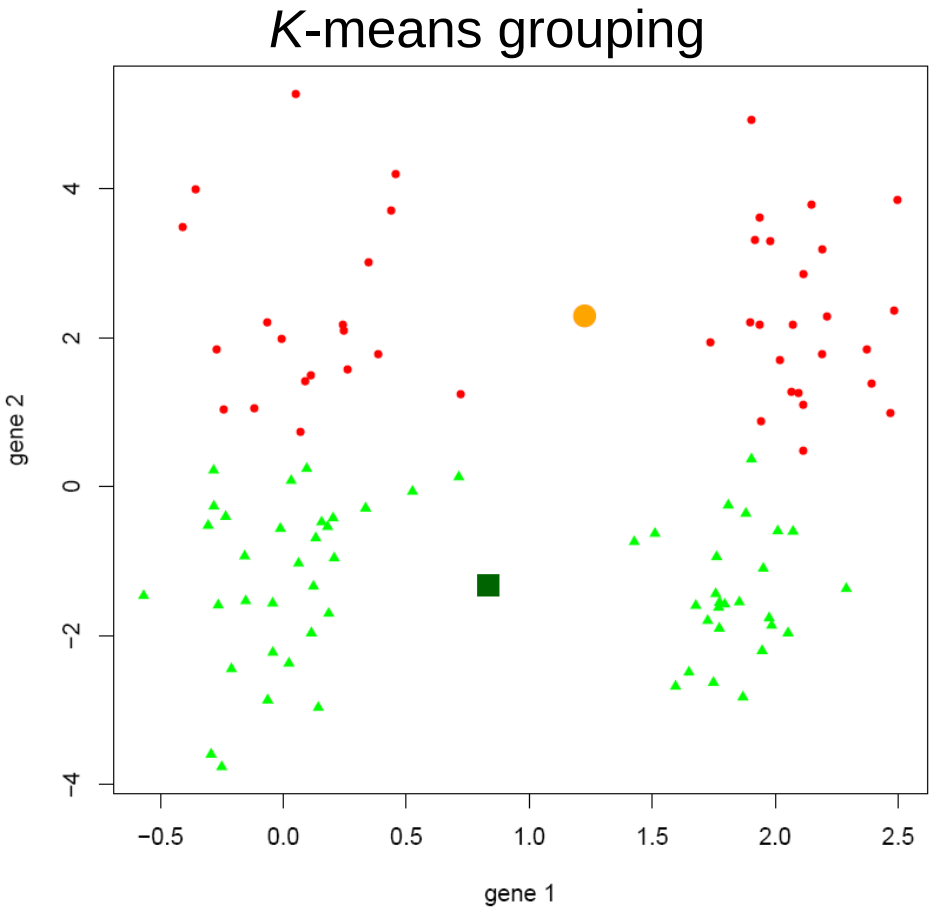
$$\mathbf{m}^{(k)} = \frac{1}{\{\#i \mid \hat{k}^{(i)} = k\}} \sum_{\{i \mid \hat{k}^{(i)} = k\}} \mathbf{X}_i$$

Step 4: Iterate between steps 2 and 3 ’til means don’t change.

K-means clustering

The effect of the spatial distribution of the data (I)

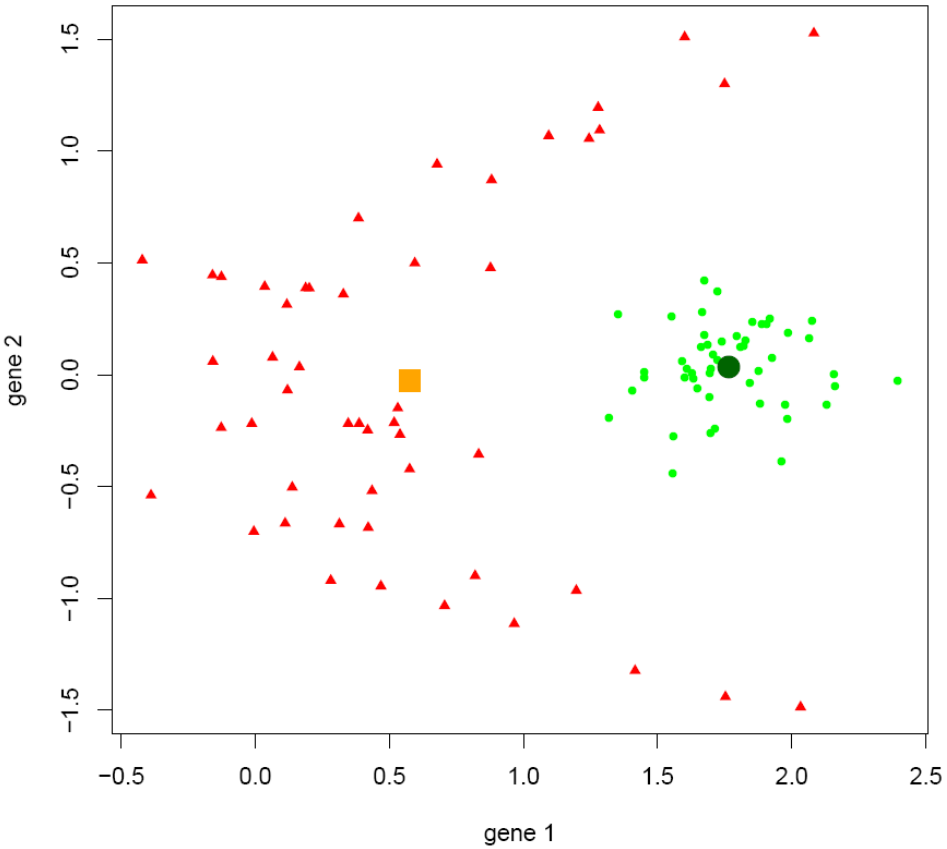




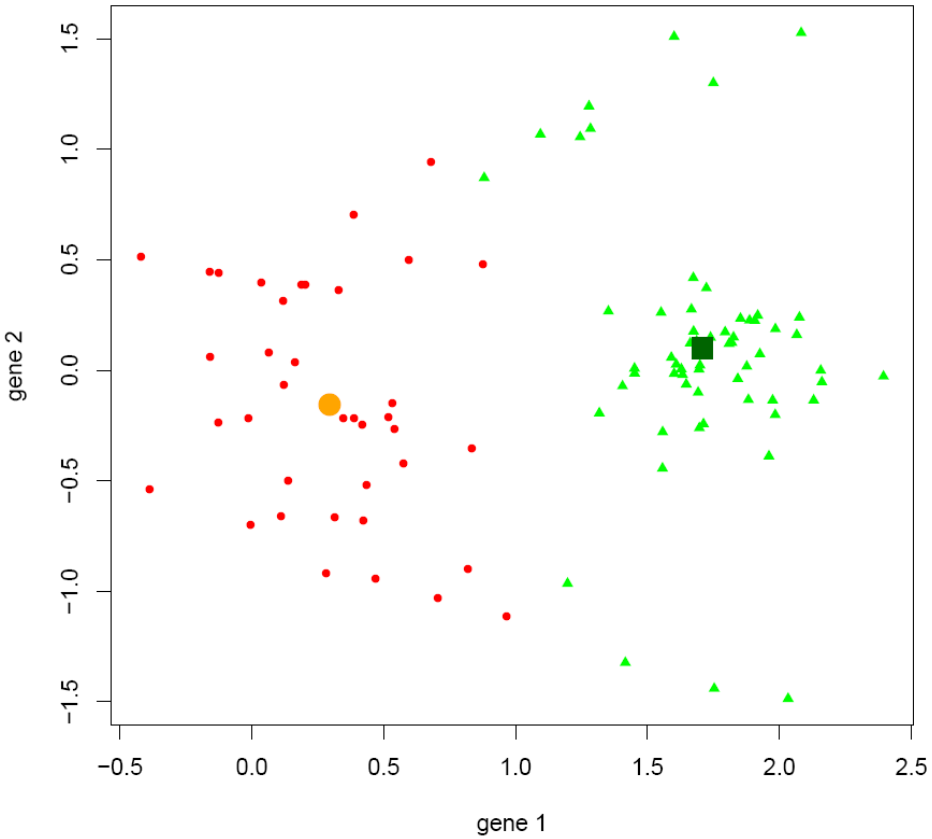
K-means clustering

The effect of the spatial distribution of the data (II)

true grouping

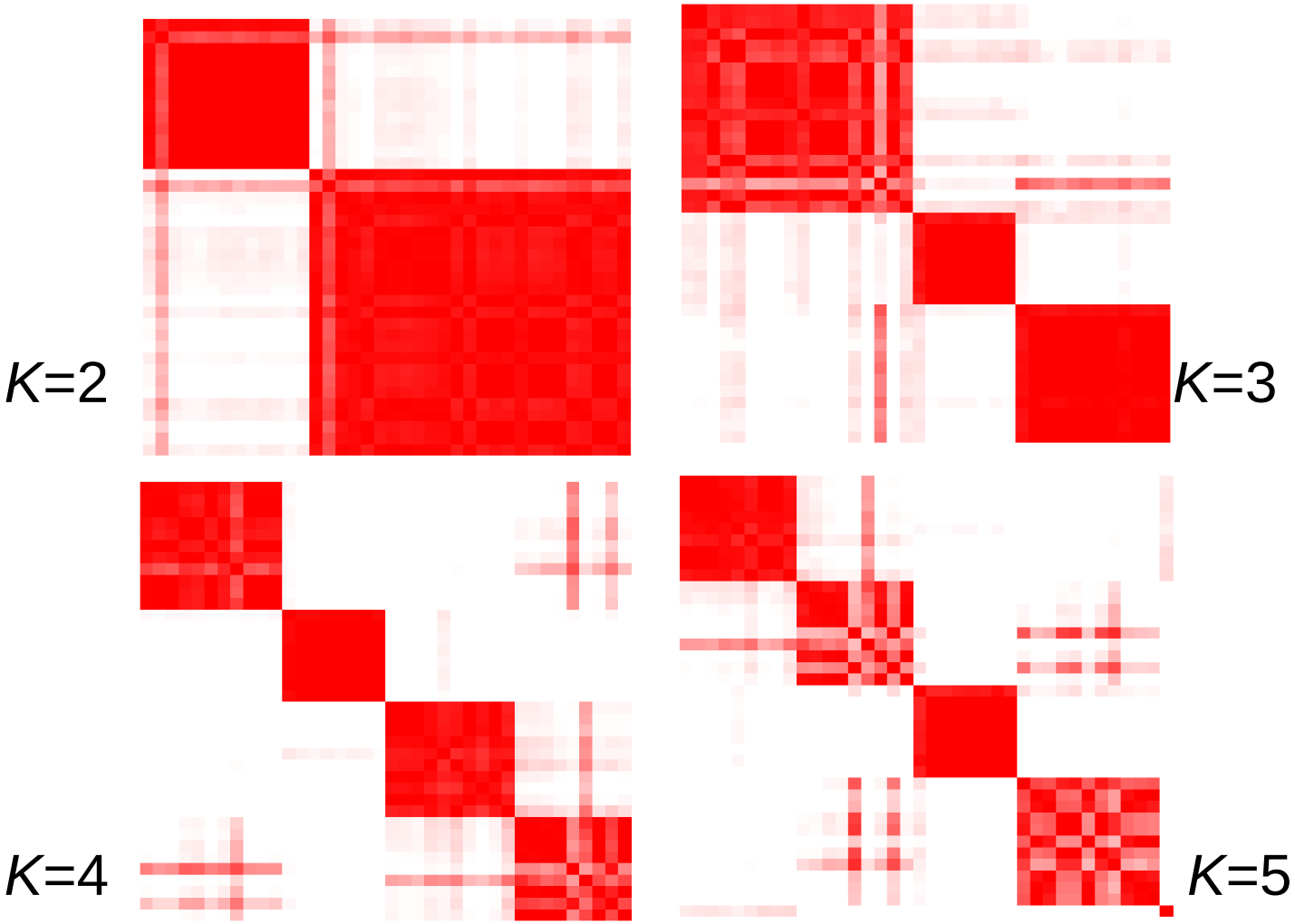


K-means grouping



K-means clustering

Choose *K* on the basis of prior knowledge or consensus clustering. The Golub data set revisited for the latter.



K-means clustering

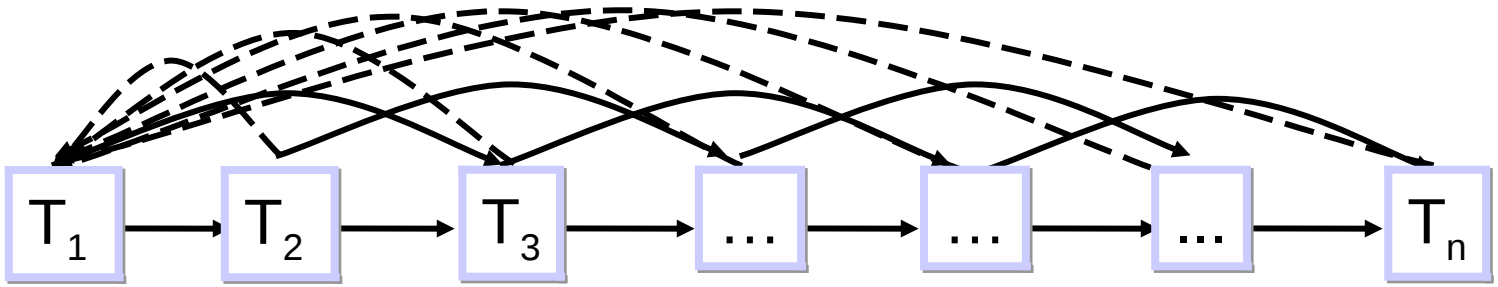
(of time courses)

K-means clustering

Time-course experiments

A time-course experiment is a single factor experiment with time being the factor. This factor has a natural ordening (as opposed to, e.g., the factor being placebo and treatment A, B, C).

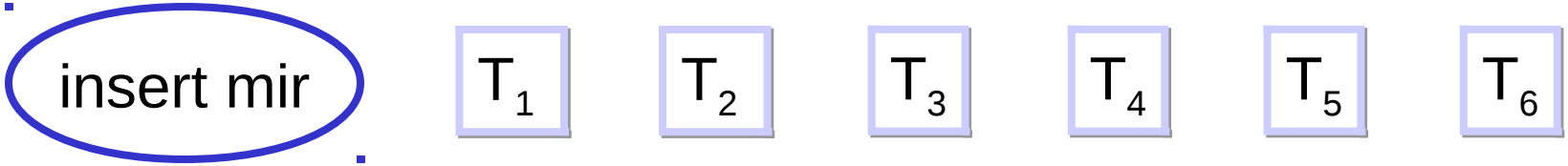
In general, a design for a time-course experiment is a selection of the hybridizations illustrated below:



K-means clustering

Example

Consider an RNAseq experiment in which a cell line has been followed of time after insertion of a microRNA:



Question

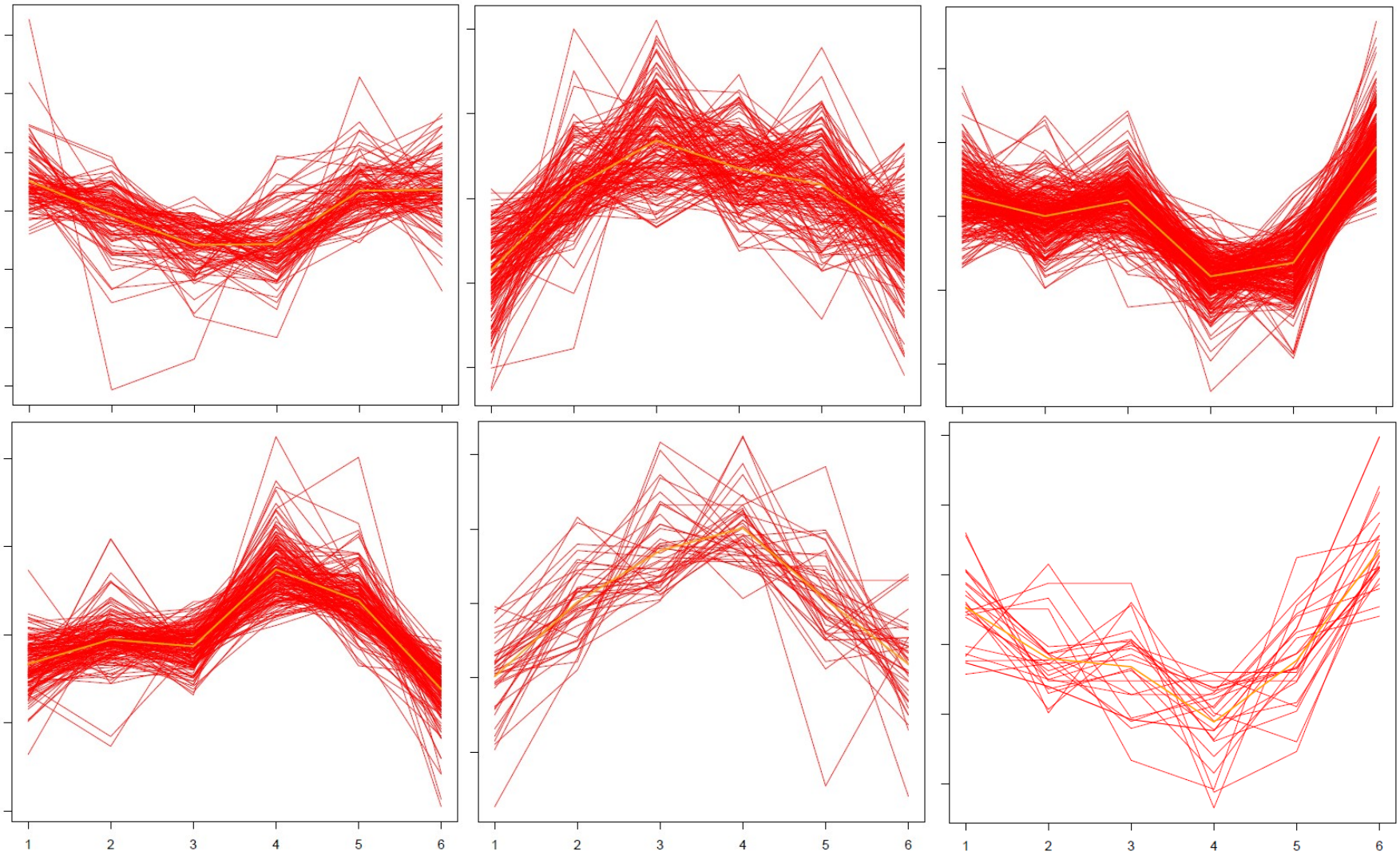
Which genes behave similarly over time?

Solution

Cluster genes by means of *K*-means.

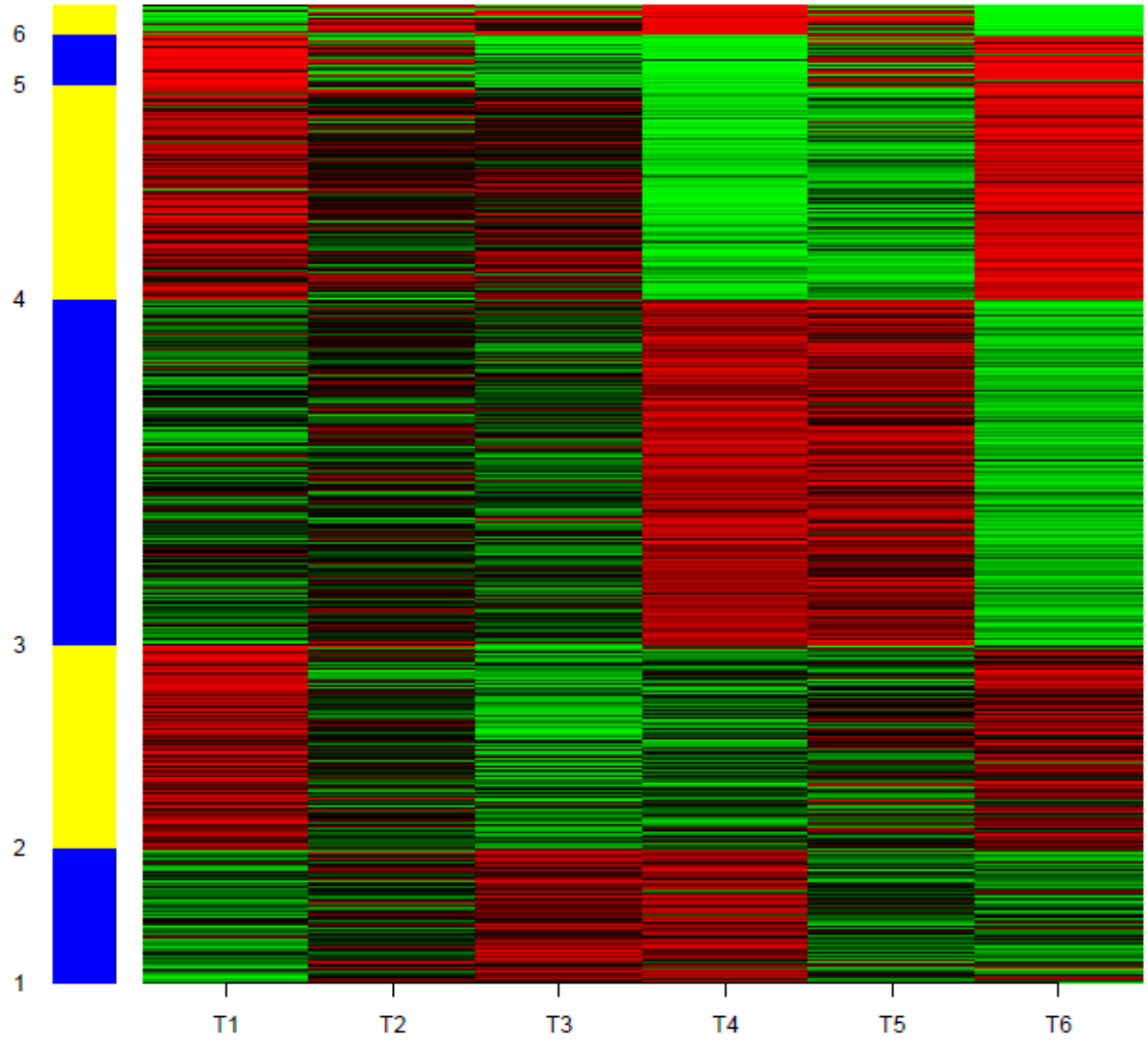
K-means clustering

Six clusters found



K-means clustering

Six clusters found

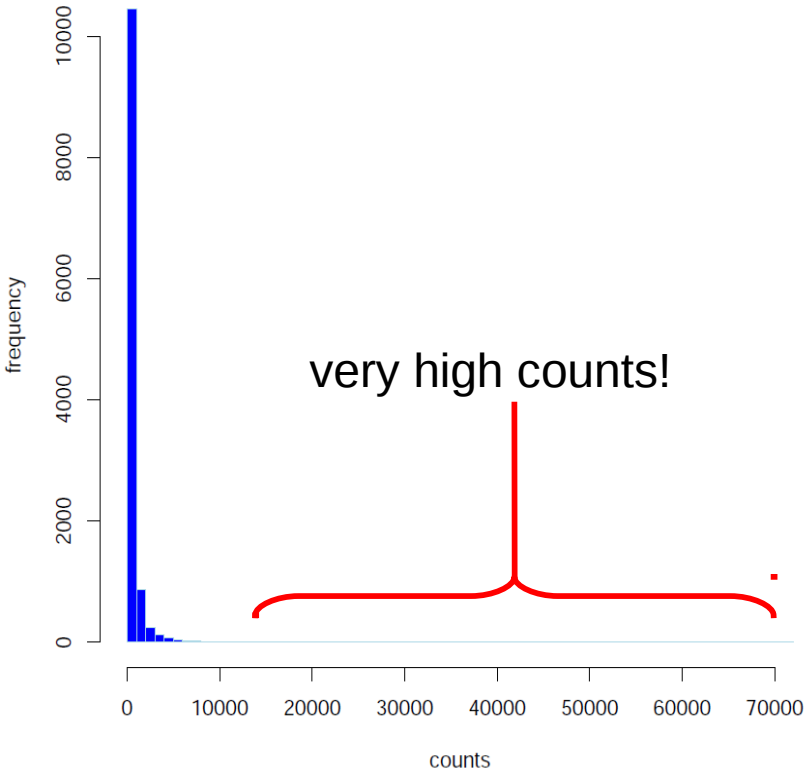


K-means clustering

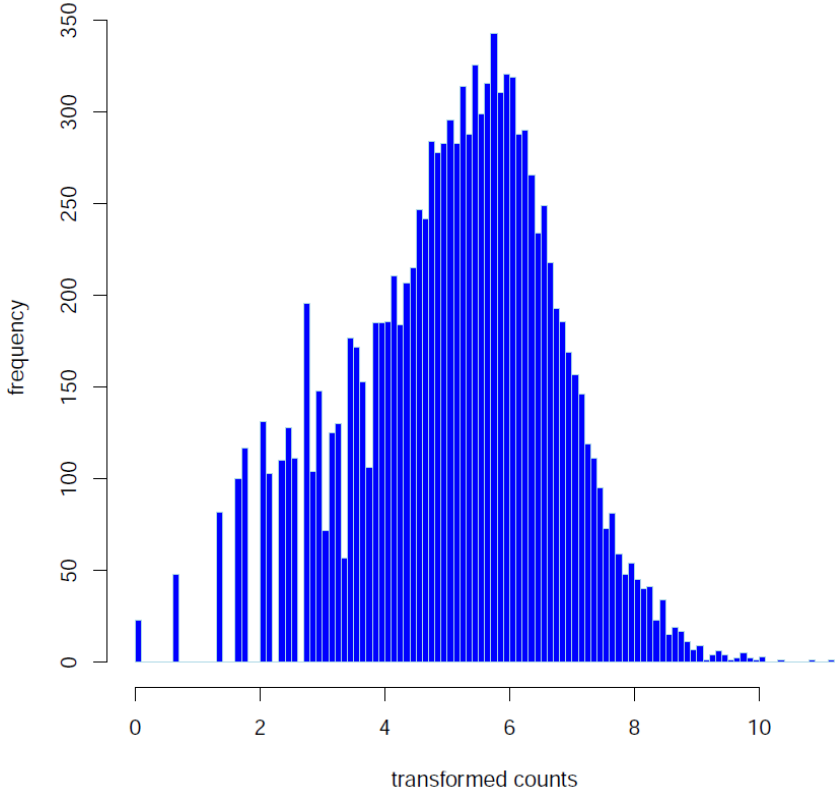
To transform or not?

Transformation of the data or use tailor-made methods.

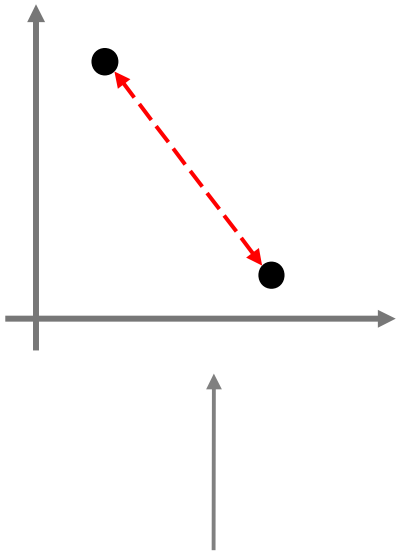
Histogram counts



Histogram transformed counts



K-means clustering



To transform or not?

No transformation and standard method.

What if ...

Expression data

	s_1	s_2
g_1	75	92
g_2	120	274
g_3	3	8
g_4	8854	6267

Euclidean distance:

$$D(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sqrt{\sum_{j=1}^{p} (x_{i_1,j} - x_{i_2,j})^2}$$

Euclidean distance:

$[(75 - 92)^2 + (120 - 274)^2 + (3 - 8)^2 + \underline{(8854 - 6267)^2}]^{1/2}$

dominates other terms

K-means clustering

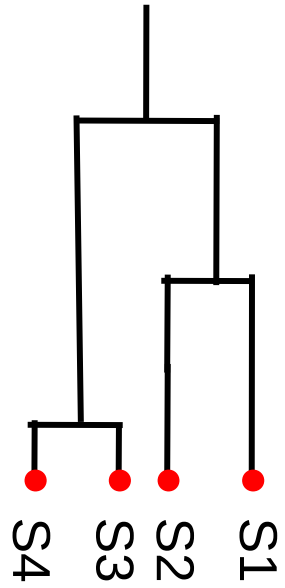
To transform or not?

No transformation and standard method.

What if ...

Expression data

	S1	S2	S3	S4
g1	75	92	87	63
g2	120	274	167	199
g3	3	8	12	24
g4	8854	6267	228	78



Very high count determine the clustering.

Above: S1 and S2 always cluster together.

K-means clustering

To transform or not?

What to do?

→ no transformation with RNAseq tailored method
(which? has it proven itself?)

→ no transformation with rank-based method
(distance measure based on Spearman correlation)

→ transformation and standard method
(which? what is effect of transformation?)

K-means clustering

Effect of transformation

Expression data

	s_1	s_2
g_1	75	92
g_2	120	274
g_3	3	8
g_4	8854	6267

log(x+1)

Expression data

	s_1	s_2
g_1	4.34	4.53
g_2	4.80	5.62
g_3	1.39	2.20
g_4	9.09	8.74

Largest contribution:

→ 4th gene,

→ much larger than rest.

Largest contribution:

→ 2th gene,

→ comparable to rest.

K-means clustering

Effect of transformation

Different clustering, but also more uniform cluster size.

Cluster size (no transformation)

	C1	C2	C3	C4	C5	C6
K=2	19	131	.	.	.	.
K=3	18	5	127	.	.	.
K=4	24	4	15	107	.	.
K=5	4	5	97	20	24	.
K=6	4	97	24	20	1	4

Cluster size (log-transformation)

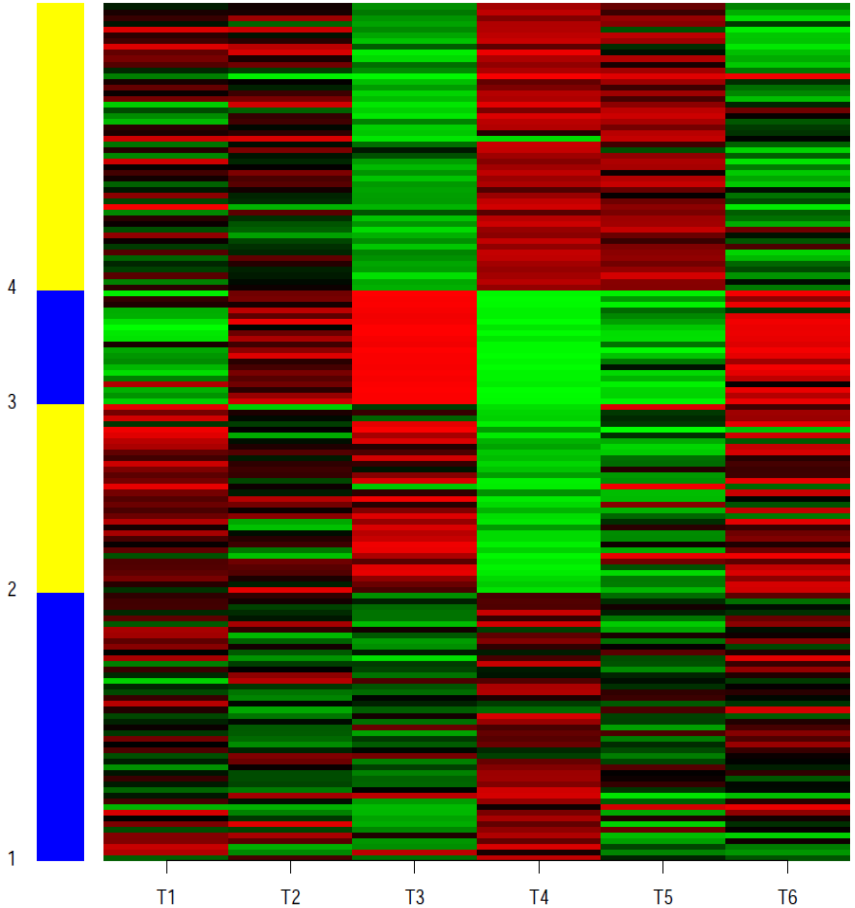
	C1	C2	C3	C4	C5	C6
K=2	49	101	.	.	.	.
K=3	44	88	18	.	.	.
K=4	60	36	40	14	.	.
K=5	36	60	4	37	13	.
K=6	1	36	35	17	4	57

K-means clustering

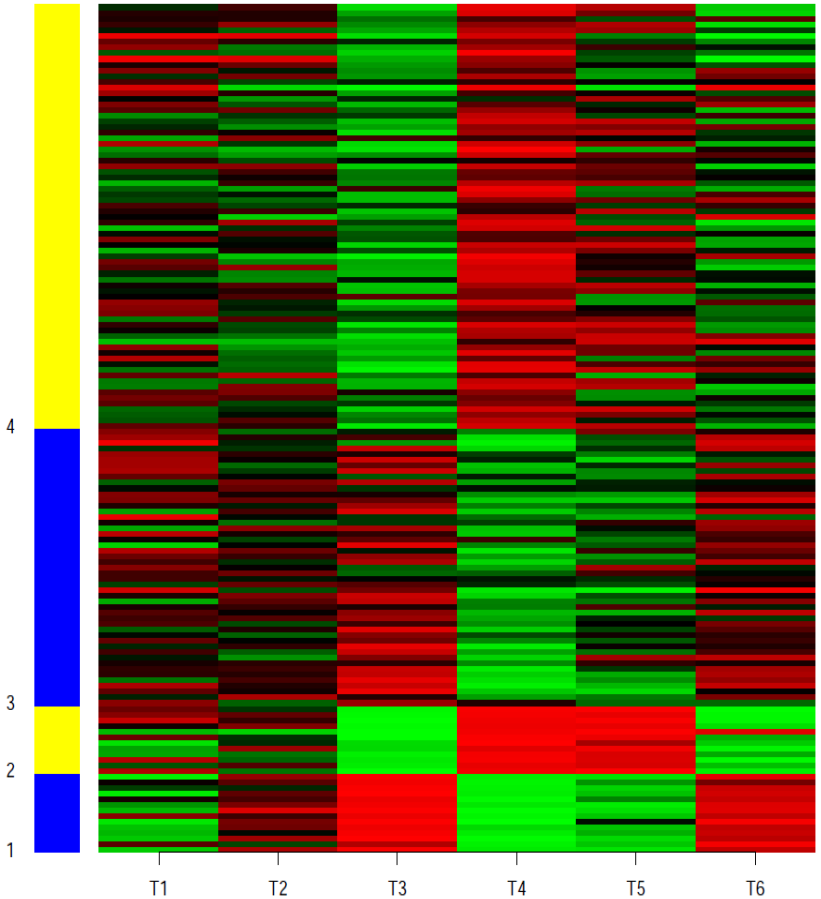
Effect of transformation

Different transformation, different clustering.

Log transform



Square root transform



Principal component analysis (PCA)

Principal component analysis

Objective of principal component analysis

Data reduction by representing the data as simply as possible while minimizing loss of information to make interpretation easier.

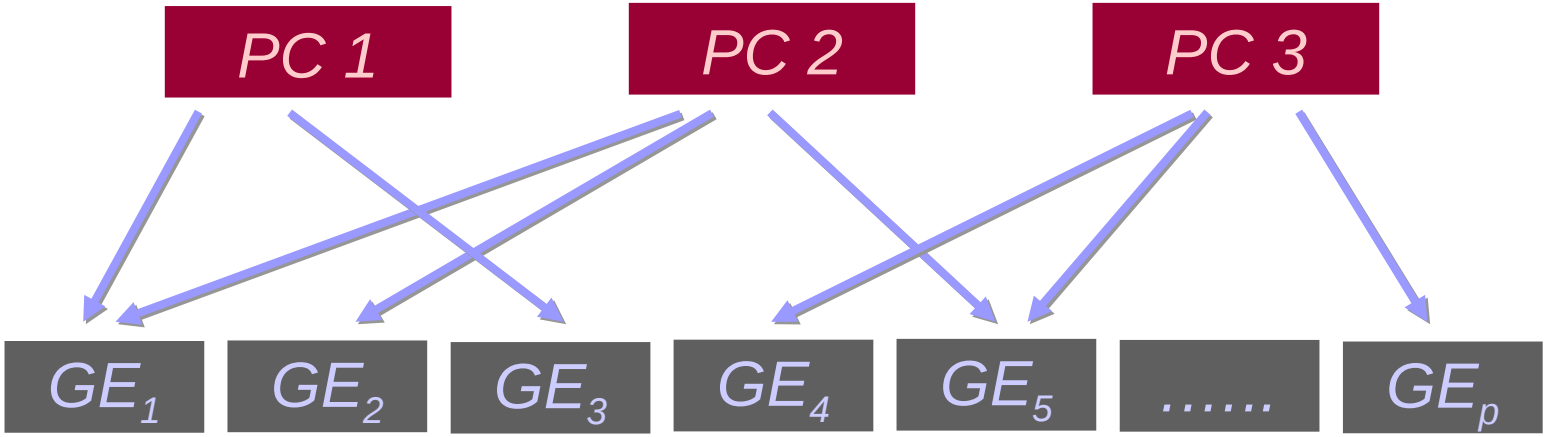
Gene expression matrix

	$Sample_1$	$Sample_2$	……	$Sample_n$
Feature 1	0.05	2.77	……	0.45
Feature 2	-2.93	0.36	……	-0.87
⋮	⋮	⋮	⋮	⋮
Feature p	1.64	-1.10	……	0.26

Principal component analysis

We assume that in a tumor:

- Not 1000 independent things happening
- Only a few underlying events affecting all variables
- Variables are highly correlated.



Principal component analysis

original data

	id1	id2	id3	id4
gene1	-0.219	-0.427	0.264	-5.749
gene2	-0.083	-7.216	-3.864	0.774
gene3	-1.083	-1.146	-1.225	-2.361
gene4	0.043	-0.463	0.127	-0.395
gene5	1.165	0.862	1.160	1.239
gene6	-0.296	0.286	-0.696	-1.197
gene7	1.762	1.075	1.462	1.160
gene8	0.463	0.212	0.494	0.582
...	...	...	...	...

principal component analysis →

low-dim. representation

	id1	id2	id3	id4
pc1	-0.726	-0.590	-1.259	-0.680
pc2	1.198	-0.061	2.215	0.434
pc3	-6.305	-3.013	1.646	-7.037

principal components ≈
“sample characterization”

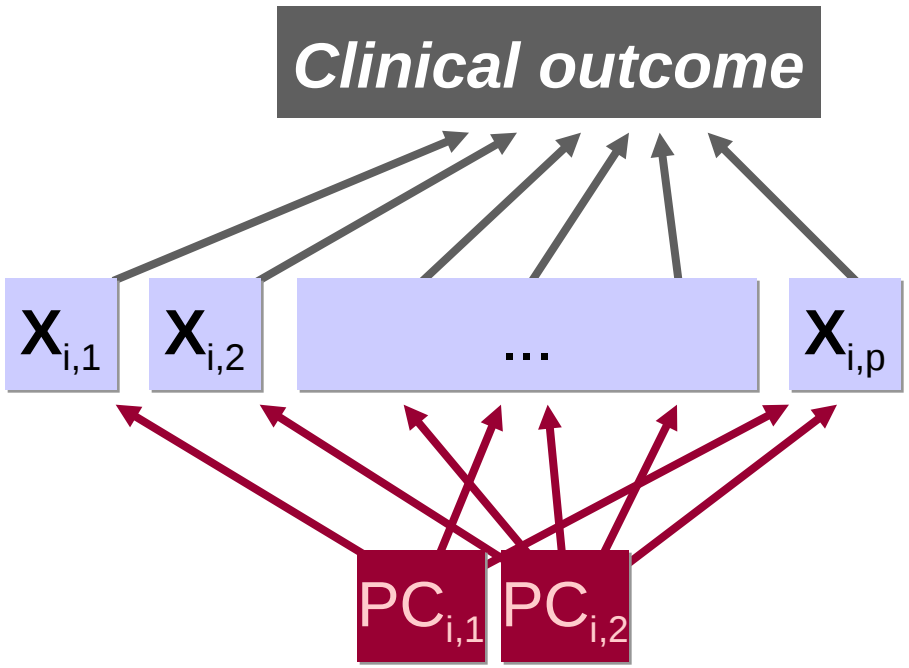
from characterization
to observation

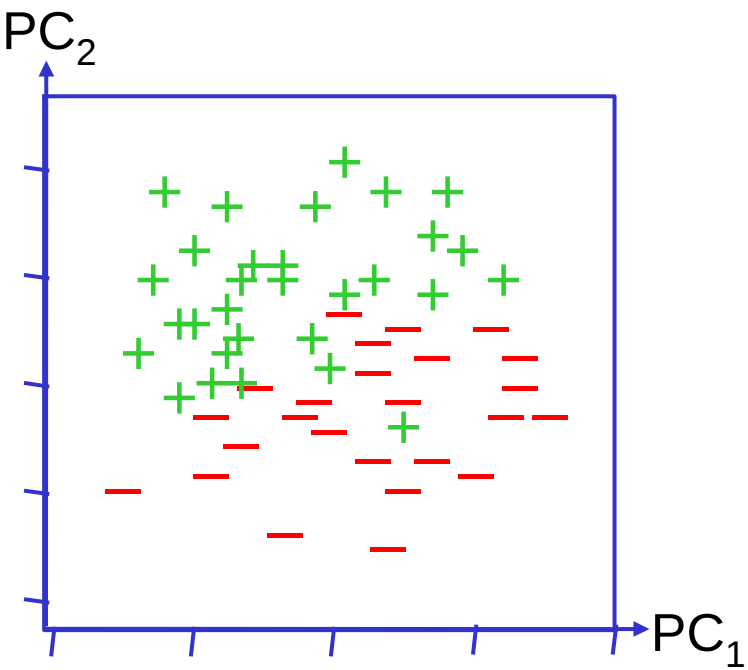
```
gene1  = b11 * pc1 + b12 * pc2 + b13 * pc3 (+ error)
gene2  = b21 * pc1 + b22 * pc2 + b23 * pc3 (+ error)
...    = ...
```

b's are component
contributions to observations

Principal component analysis

In a ***PC-plot*** the values of the PCs of each sample is plotted. The samples are labeled in accordance with a clinical parameter.

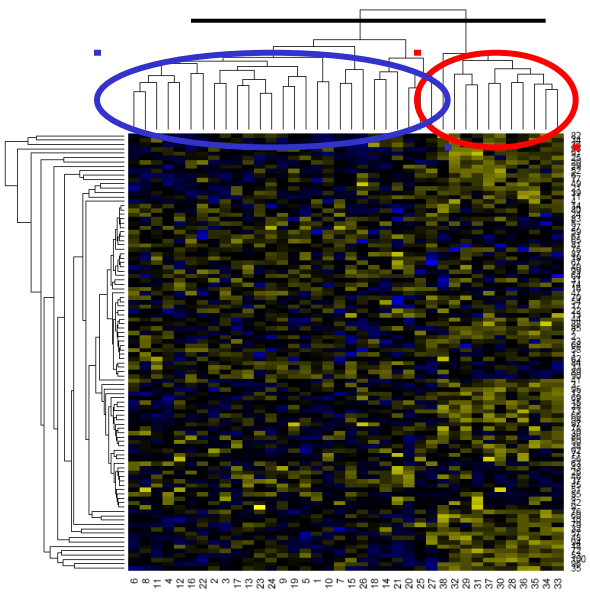


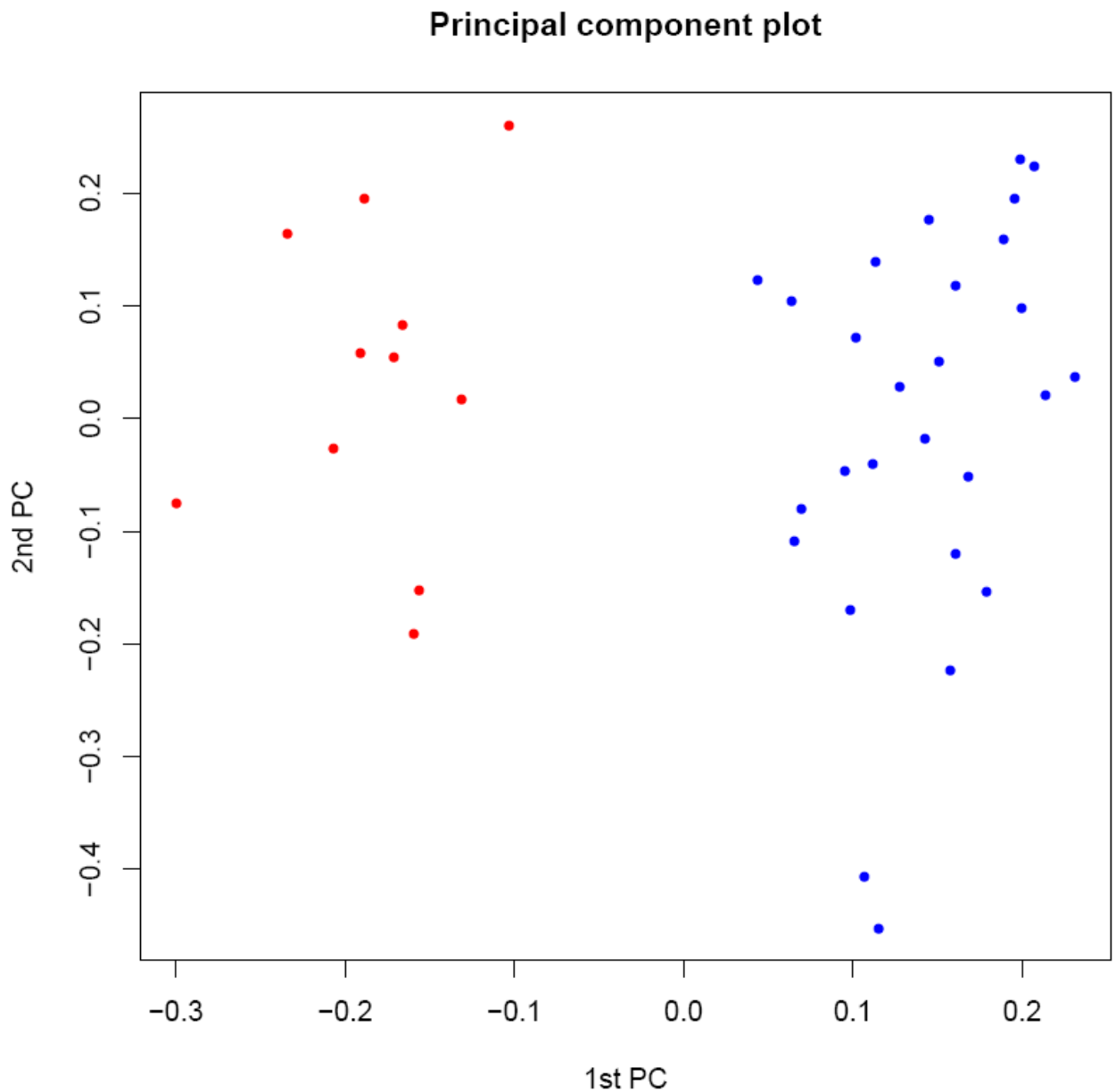


If different values of the clinical parameter occupy different parts of the plot, it seems possible to separate the samples using PCs.

Principal component analysis

The PC-plot can be used to confirm hierarchical clustering results.

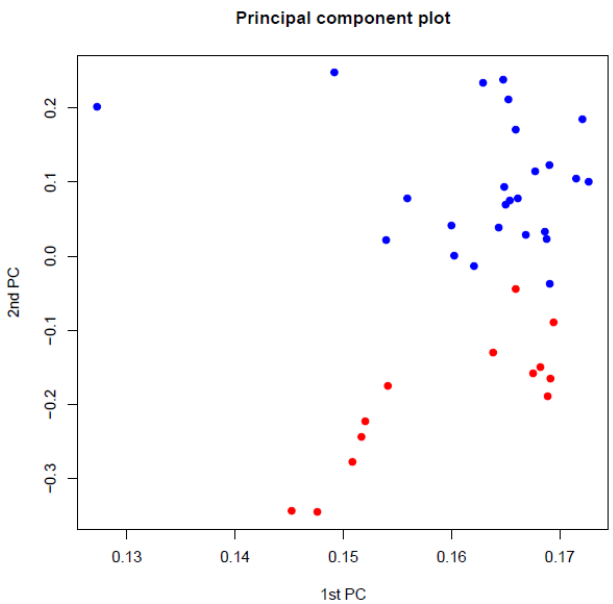


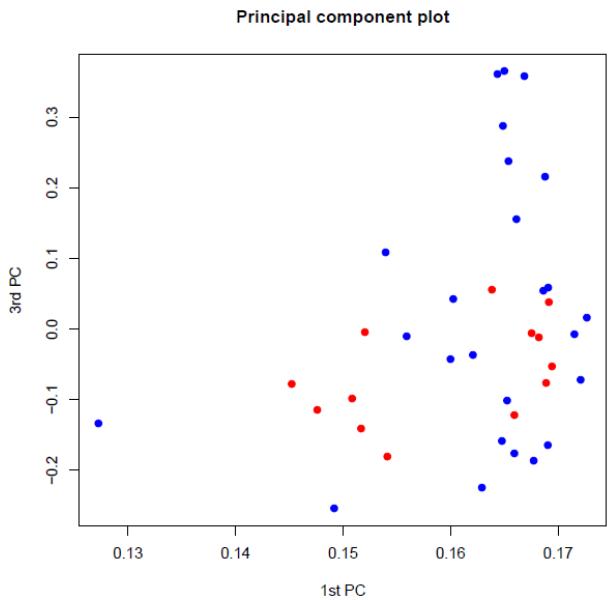


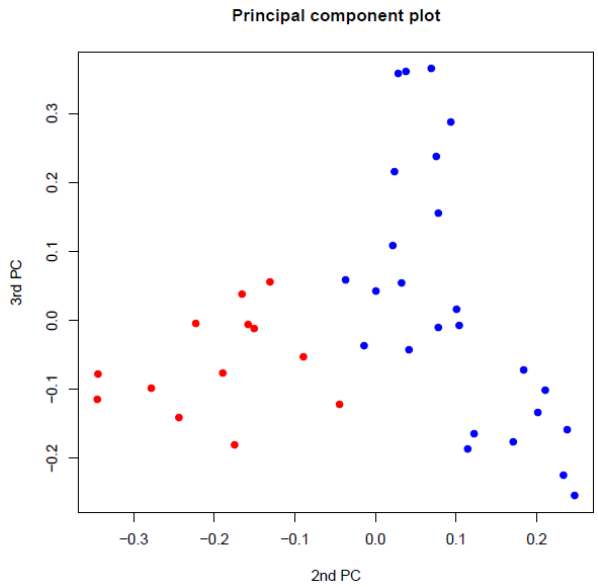
Principal component analysis

Golub data

SVD-plots for 2 clusters as found by hierarchical clustering





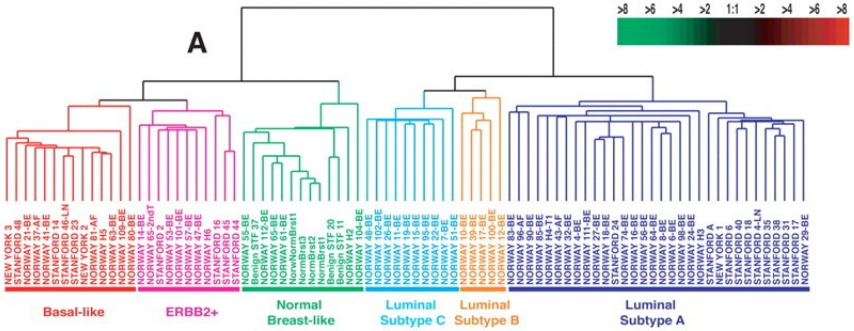


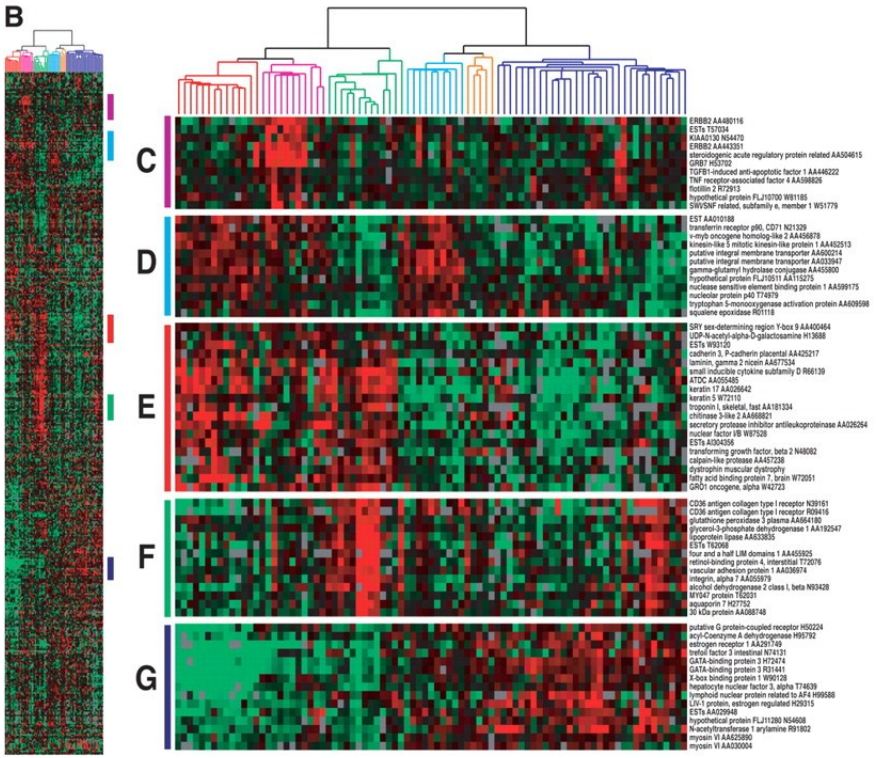
A published example

A published example

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]





Using 78 breast cancer profiles, five subtypes are distinguished:

- Basal
- ERBB2
- Luminal A
- Luminal B
- Normal

A published example

Let us have a closer look.

Microarrays

"The cDNA microarrays … contained … 8102 genes."

Features

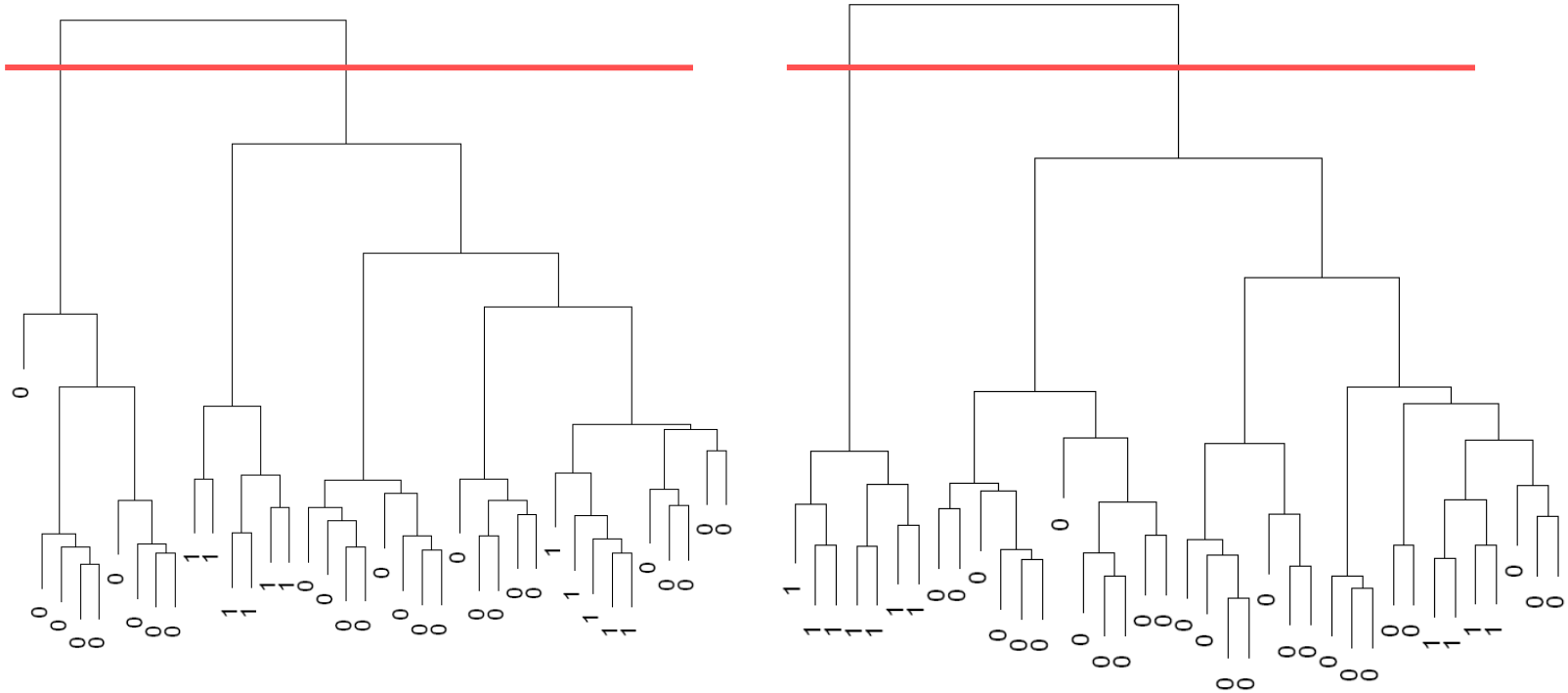
"Using the intrinsic gene set of 456 cDNA clones, selected to optimally identify the intrinsic characteristics of breast tumors …"

Clustering

"Average-linkage hierarchical clustering was applied …"

A published example

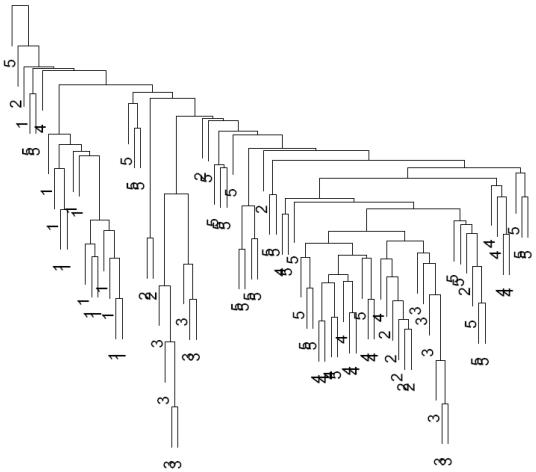
What is the effect of gene selection?
Two different gene subsets of the Golub data:



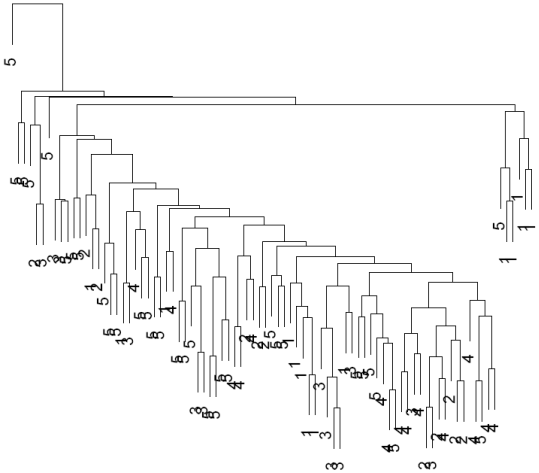
```
Contingency table of two clusterings:
                 1   2
               1 22   7
               2  9   0
```

A published example

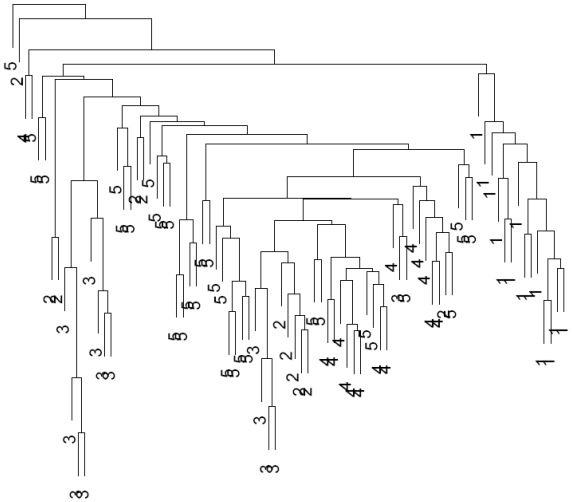
Which distance measure is used in the clustering?
Let’s try a few.



Link: average
Dist: Euclidean



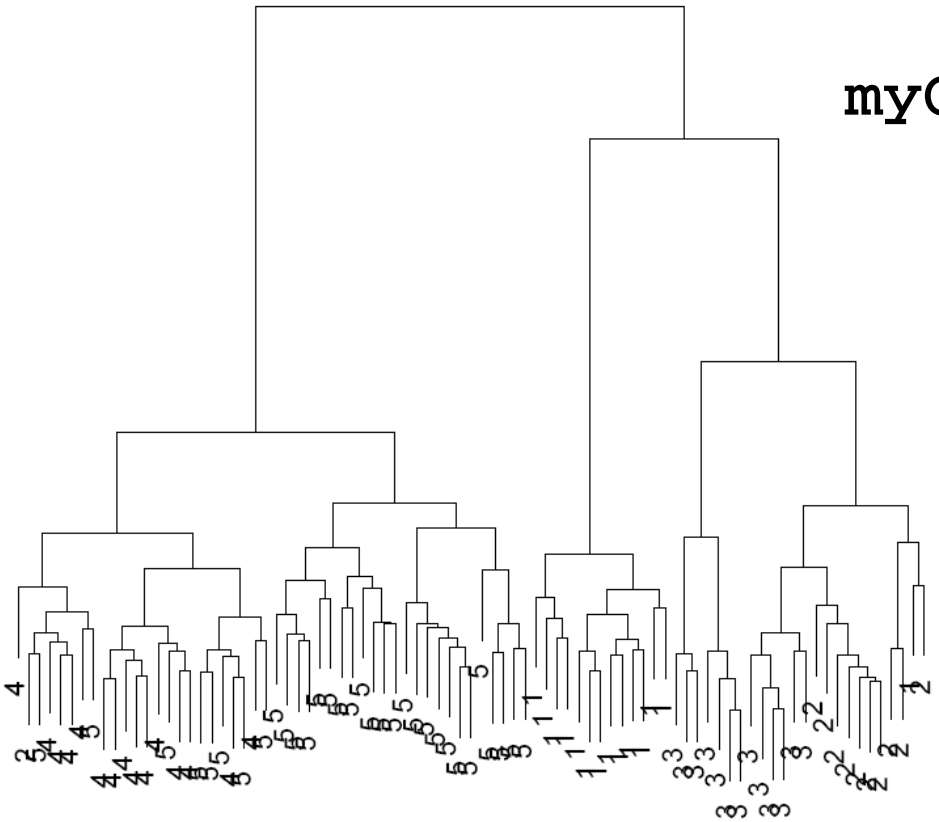
Link: average
Dist: maximum



Link: average
Dist: Manhattan

A published example

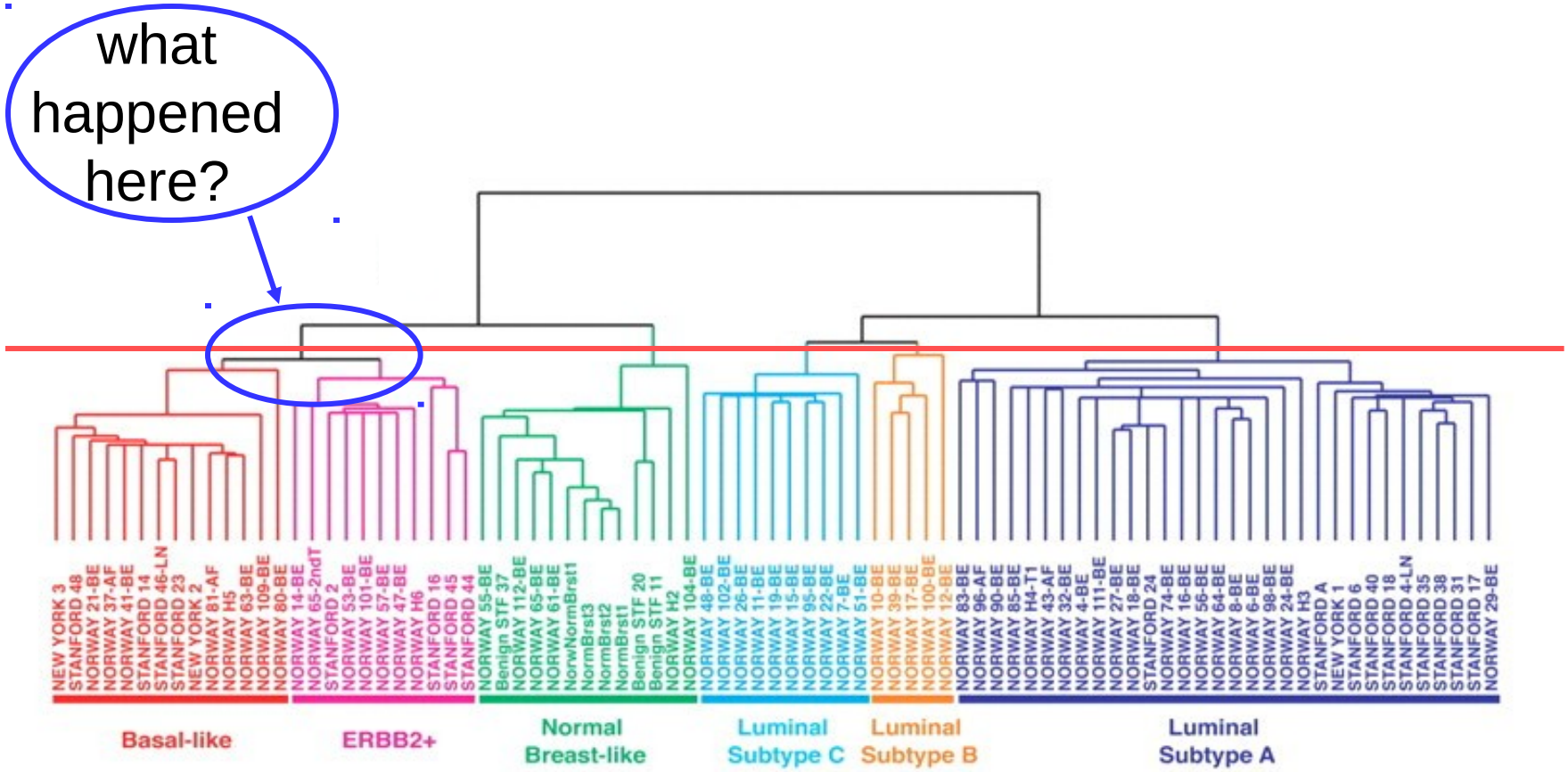
When using another type of linkage:



	sorlielabels				
myClust	1	2	3	4	5
1	13	0	0	0	0
2	1	10	6	0	0
3	0	1	0	15	8
4	0	0	7	0	0
5	0	0	0	0	24

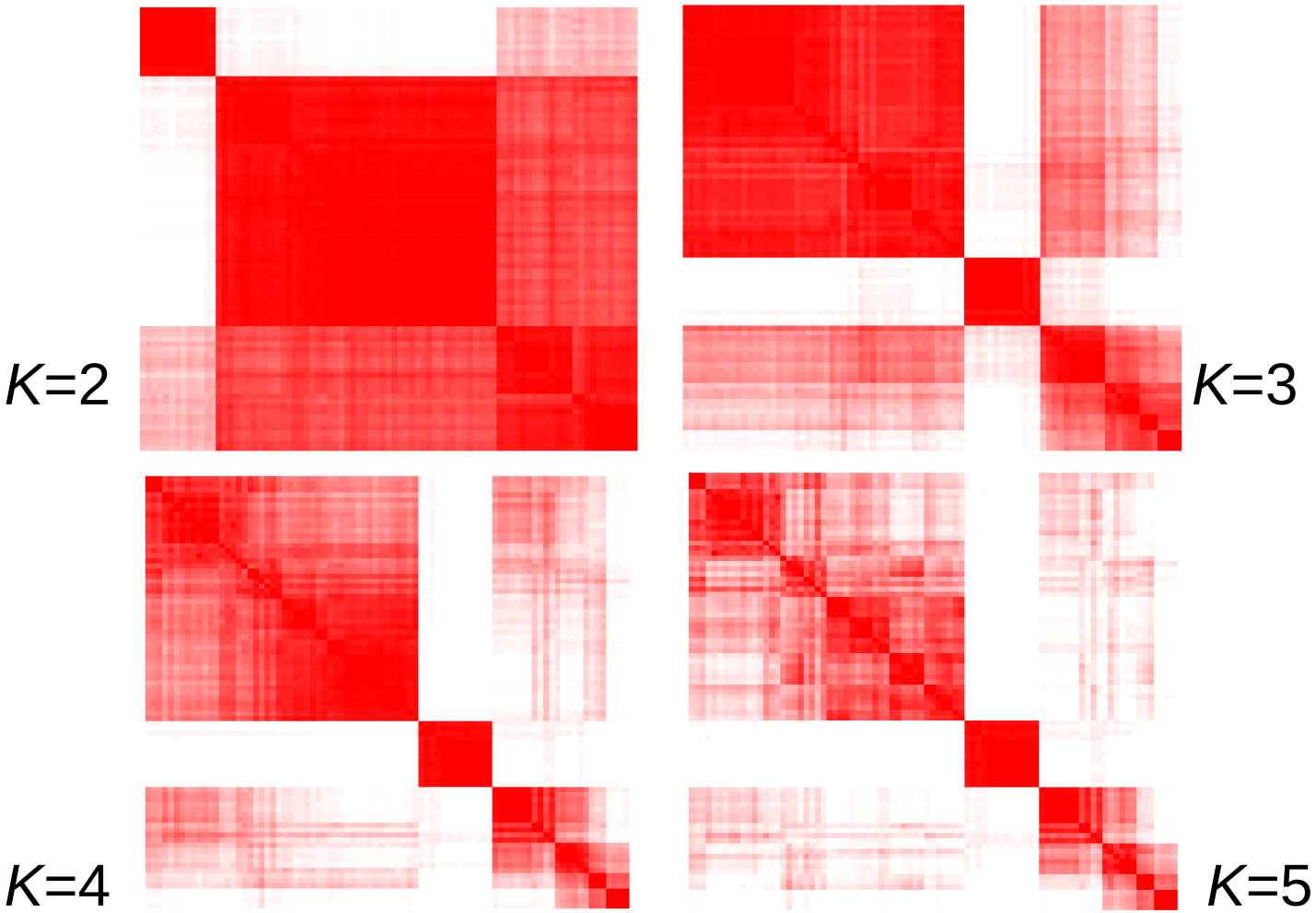
A published example

Having chosen the number of clusters somehow, a ruler is used to cut the dendrogram.



A published example

How many subtypes are there? Concensus clustering.



A published example

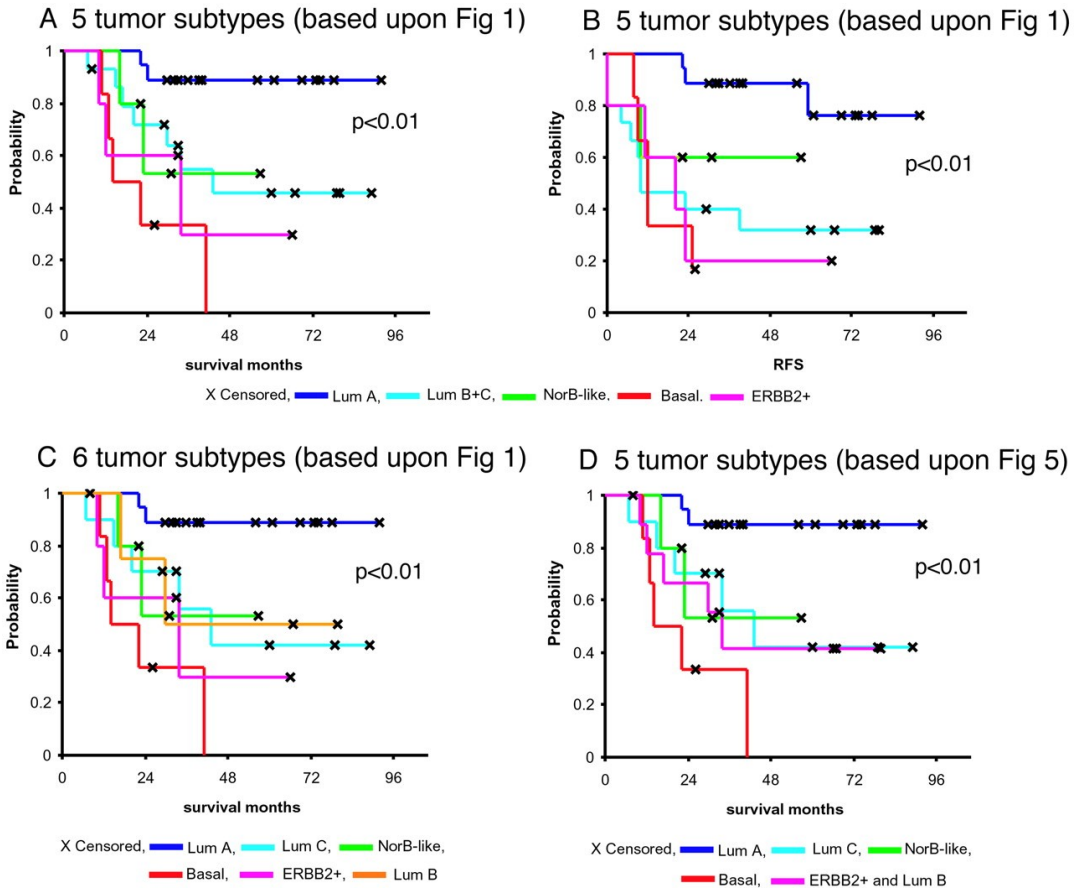
Do you want to try yourself?

```
> library(hybridHclust)
> library(marray)
> data(sorlie)
> data(sorlielabels)
```

.. and off you go!

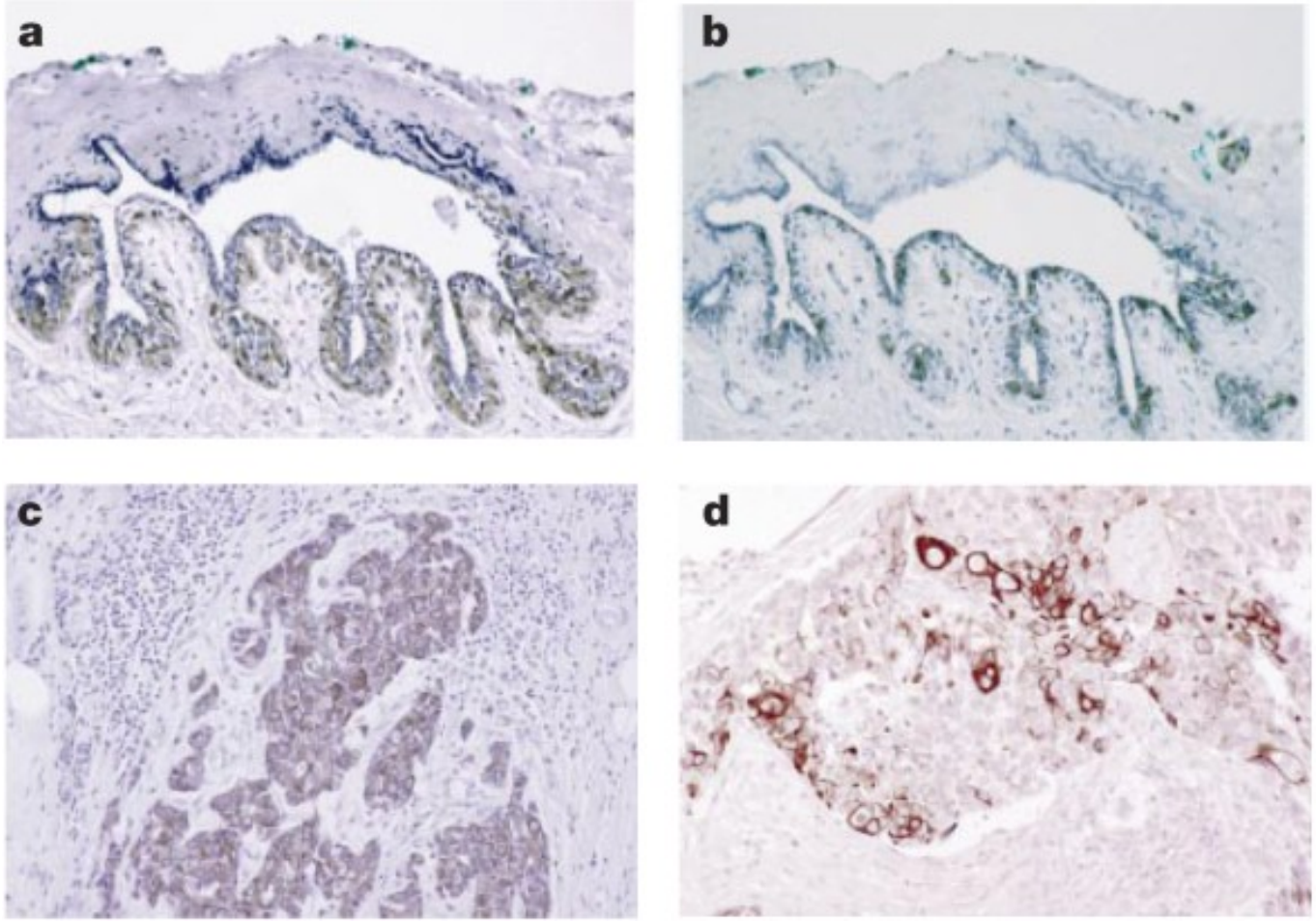
A published example

Why do people believe these breast cancer subtypes?
1) Subtypes exhibit different clinical outcome.



A published example

Why do people believe these breast cancer subtypes?
2) Exhibit different morphology.

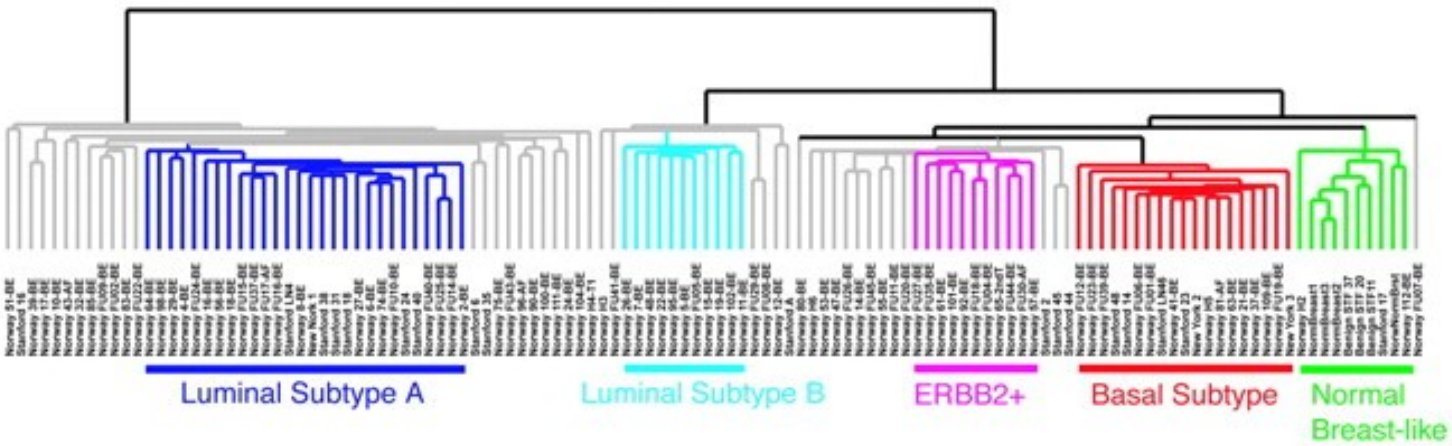


A published example

Why do people believe these breast cancer subtypes?
3) Subtypes have been confirmed.

Repeated observation of breast tumor subtypes in independent gene expression data sets

Therese Sørlie*, Robert Tibshirani†, Joel Parker‡, Trevor Hastie§, J. S. Marron¶, Andrew Nobel¶, Shibing Deng‖, Hilde Johnsen**, Robert Pesich*, Stephanie Geisler††, Janos Demeter*, Charles M. Perou‡,‡‡, Per E. Lønning††, Patrick O. Brown§§, Anne-Lise Børresen-Dale**, and David Botstein*¶¶



A published example

Medio 2012, the story continues …

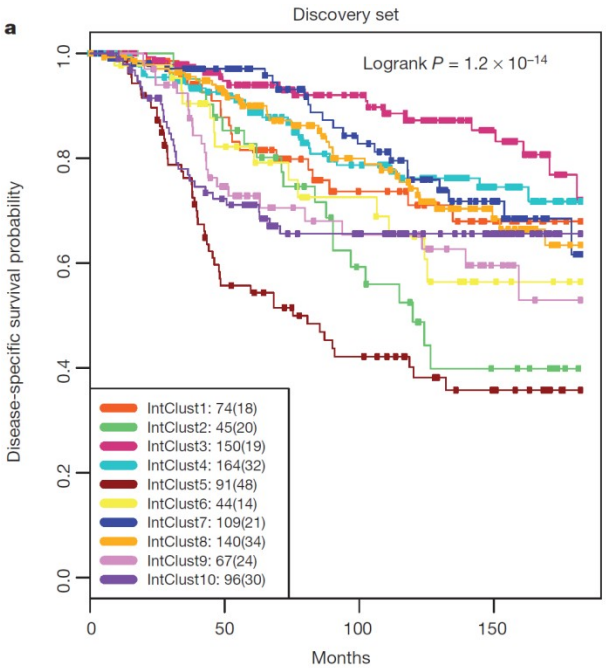
ARTICLE

doi:10.1038/nature10983

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis[1,2]†*, Sohrab P. Shah[3,4]*, Suet-Feung Chin[1,2]*, Gulisa Turashv
Doug Speed[2,5]†, Andy G. Lynch[1,2], Shamith Samarajiwa[1,2], Yinyin Yuan[1,2], Ste
Ali Bashashati[3], Roslin Russell[2], Steven McKinney[3,4], METABRIC Group‡, Anit
Gordon Wishart[8], Sarah Pinder[9], Peter Watson[3,4,10], Florian Markowetz[1,2], Lei
Anne-Lise Børresen-Dale[6,12], James D. Brenton[2,13], Simon Tavaré[1,2,5,14], Carlo

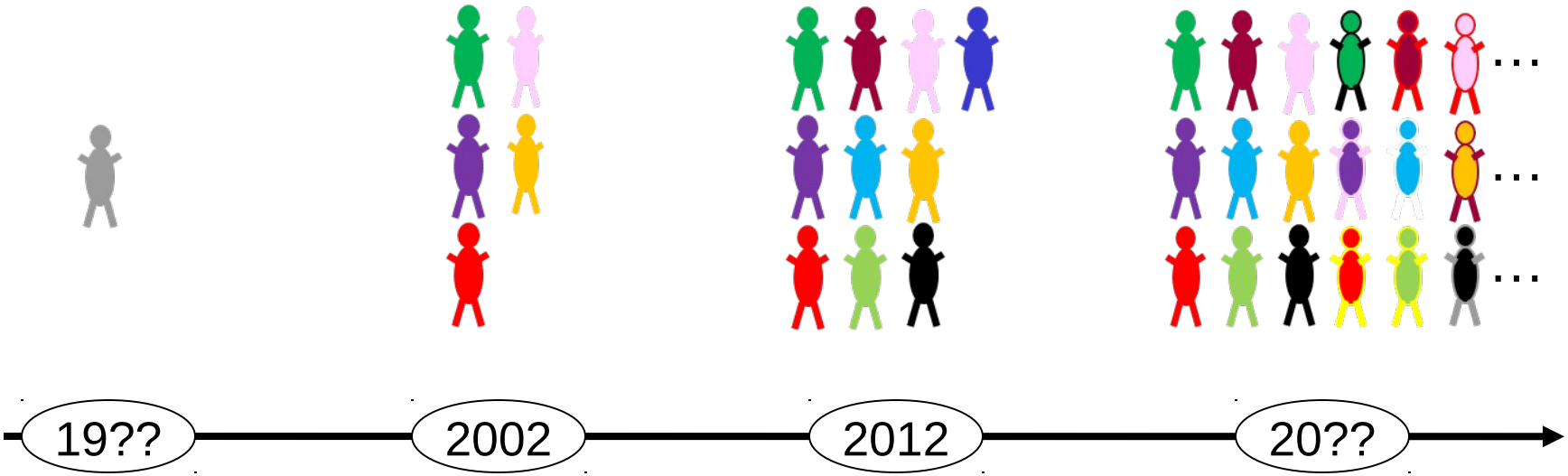
Inclusion of more molecular information suggests the existence of 10 subgroups.



A published example

How many subgroups really exist?

Genetically, everybody is unique. Thus …



8 billion?

…

personalized medicine
to the max?

Hierarchical clustering of aCGH data

Hierarchical clustering of aCGH data

Normalized or segmented DNA copy number data are mostly used for clustering, with standard techniques.

Clustering DNA copy number using calls and call probabilities requires special ***similarities*** (or distances) between objects being clustered.

Calls	
1.	Agreement
2.	Concordance

Call probabilities	
1.	KLdiv
2.	Ordinal

Hierarchical clustering of aCGH data

Agreement

The DNA copy number of a probe of two samples is in ***agreement*** if they are identical.

The ***agreement similarity*** between sample i_1 and i_2 is the proportion of probes with identical DNA copy number call.

	S_1	S_2
probe 1	-1	-1
probe 2	-1	0
probe 3	1	1
⋮	⋮	⋮
probe p-1	1	1
probe p	0	0

Hierarchical clustering of aCGH data

Concordance

The DNA copy number of a pair of probes of two samples are in ***concordance*** if they agree on which probe has the largest DNA copy number.

The ***concordance similarity*** between samples i_1 and i_2 is the proportion of probe pairs which DNA copy number calls are concordant.

Hierarchical clustering of aCGH data

Pairs of probes that are in concordance

	S_1	S_2
probe 1	-1	-1
probe 2	-1	0
probe 3	1	1
⋮	⋮	⋮
probe p-1	1	1
probe p	0	0

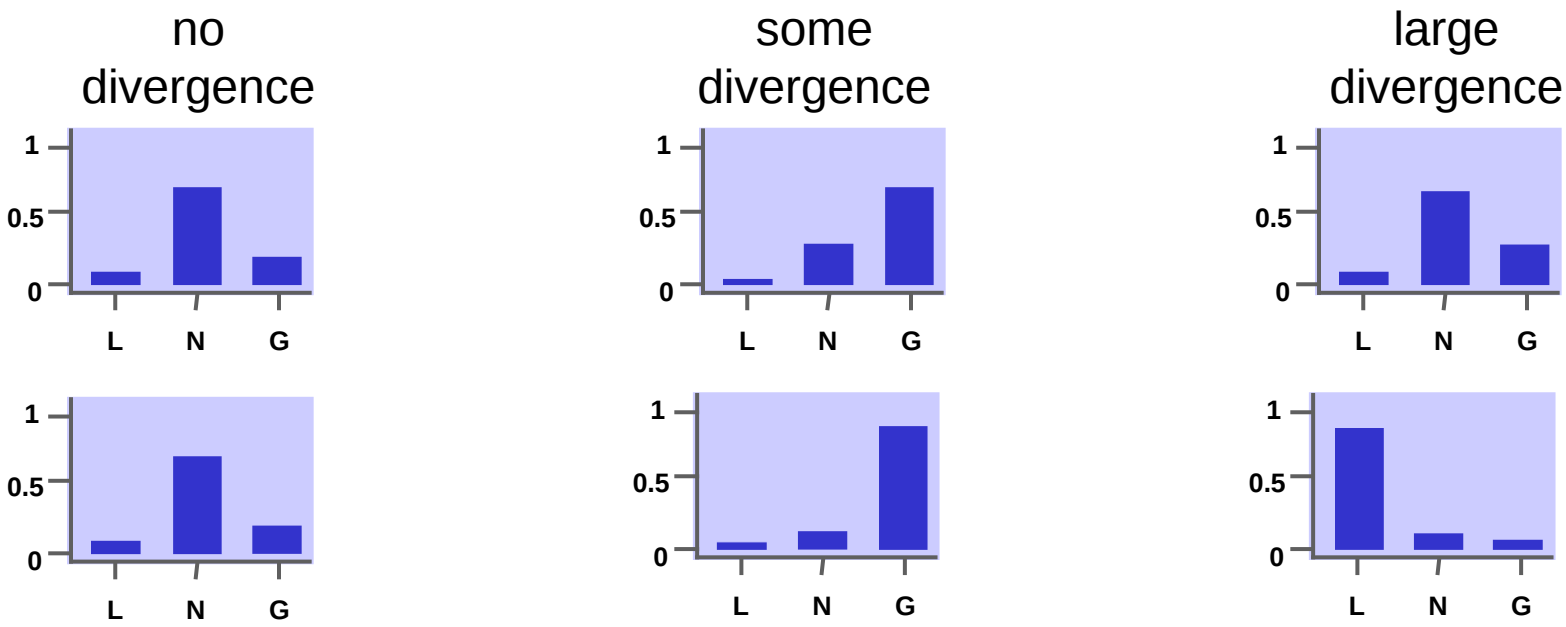
Pairs of probes that are in *dis*-concordance

	S_1	S_2
probe 1	-1	-1
probe 2	-1	0
probe 3	1	1
⋮	⋮	⋮
probe p-1	1	1
probe p	0	0

Hierarchical clustering of aCGH data

KLdiv

The Kullback-Leibler divergence is a measure for the difference between two probability distributions. KLdiv sums the call probability divergences of all probes.

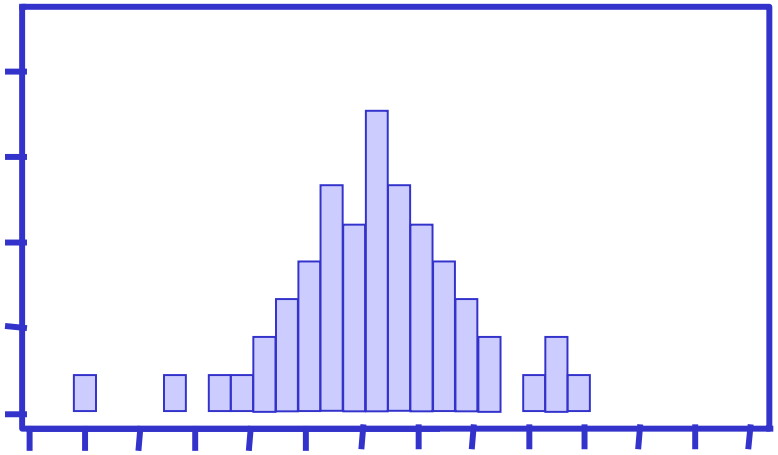


Hierarchical clustering of aCGH data

Ordinal

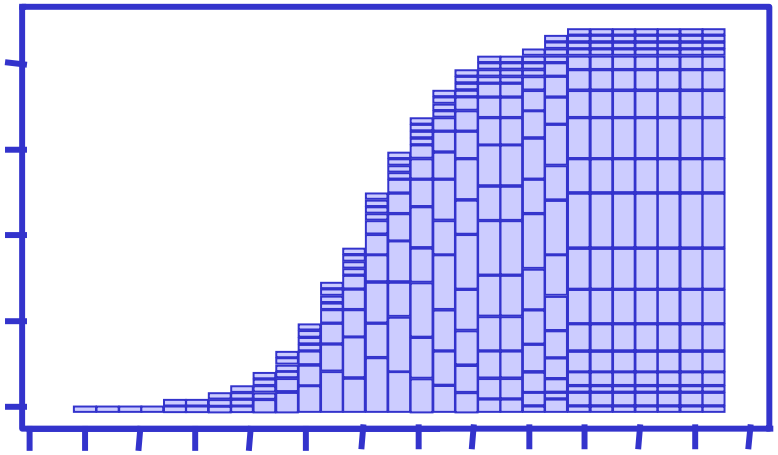
Calculate for each probe the area between the cumulative call probability distributions of the two samples. The ordinal distance is the average area over the probes.

Probability distribution



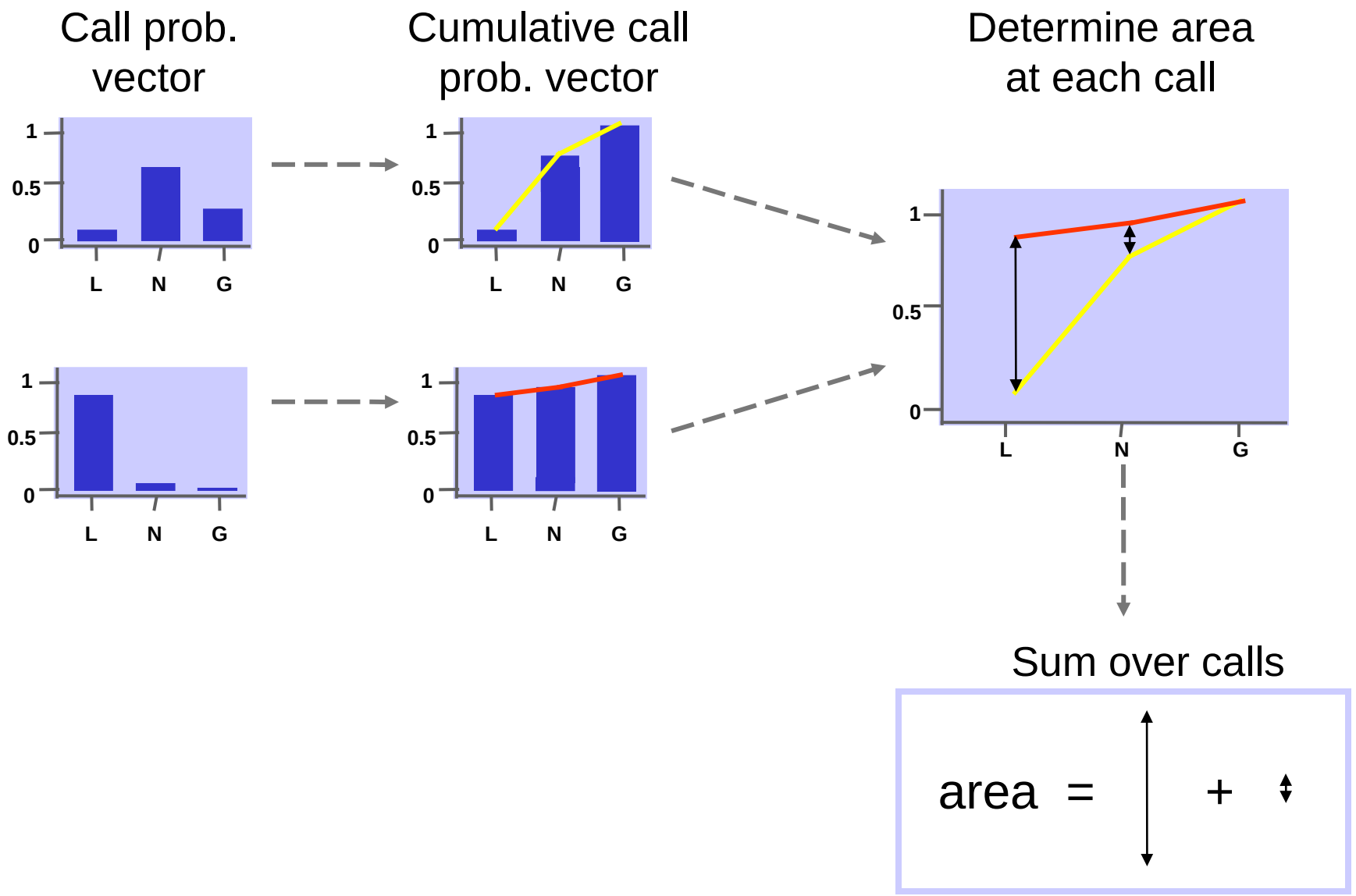
Call probability vector

Cumulative prob. distribution



Cumulative call prob. vector

Hierarchical clustering of aCGH data



Hierarchical clustering of aCGH data

Cervical cancer

Research question

Do samples of various stages of cervical cancer separate on the basis of their DNA copy number profiles?

Experimental design

DNA copy number profiles of 50 samples.

Data

DNA copy number data compactifies to 769 regions.

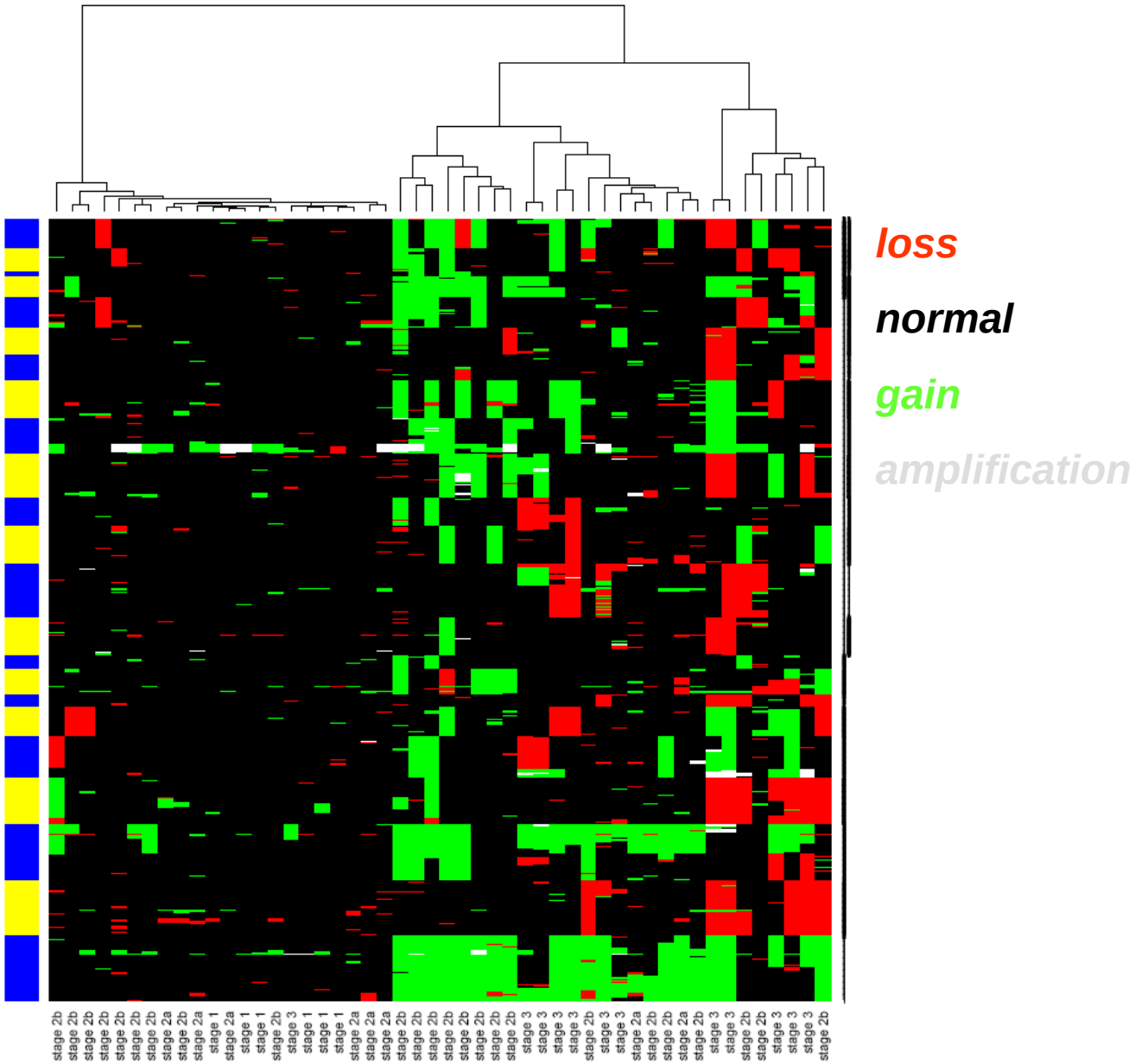
Conclusion

????

Hierarchical clustering of aCGH data

How many clusters are there?

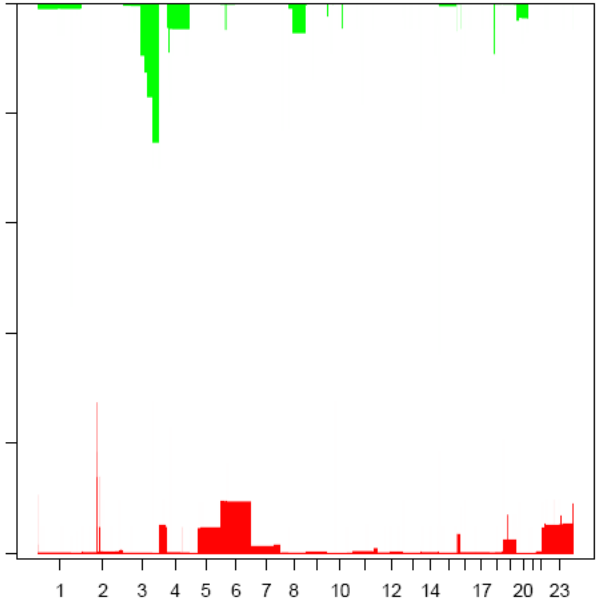
Can you interpret them from the heatmap?



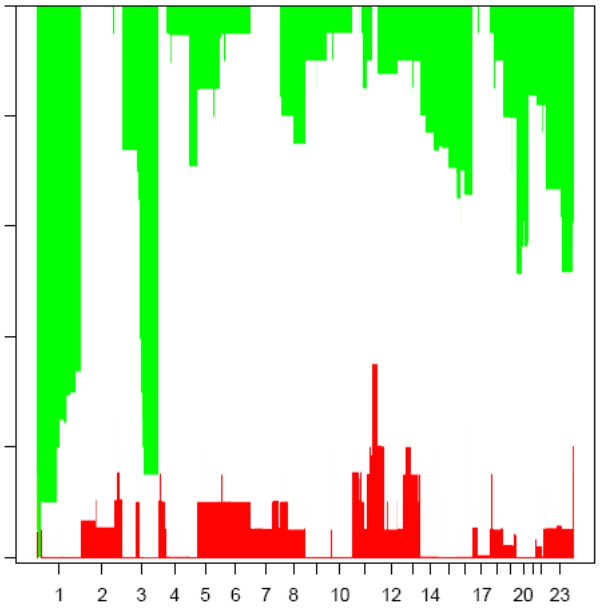
Hierarchical clustering of aCGH data

CGHcall summary plots of three clusters

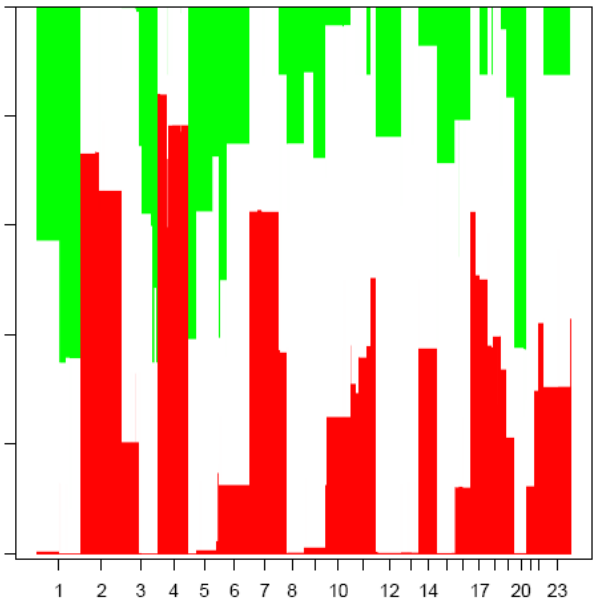
Cluster 1



Cluster 2

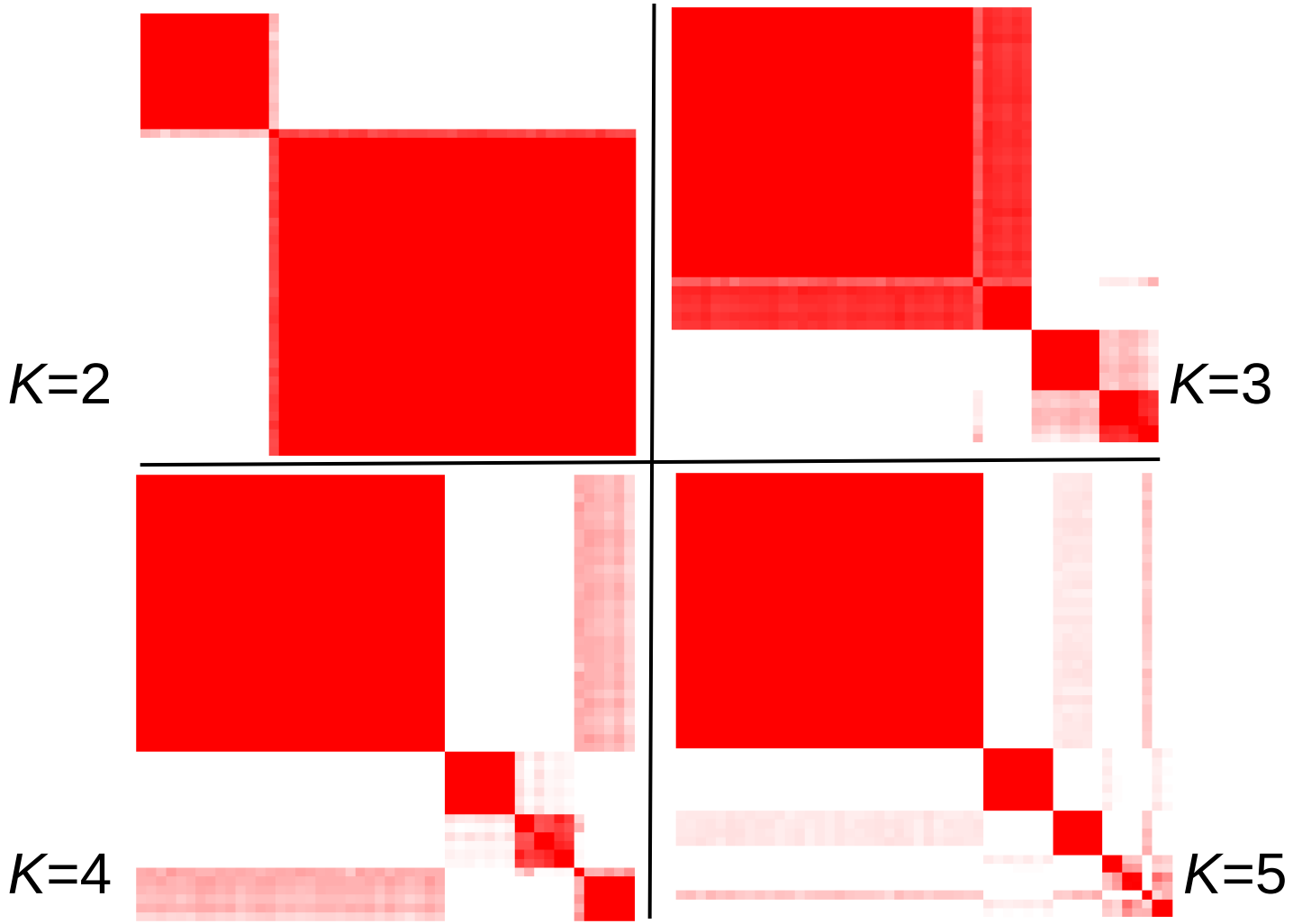


Cluster 3



Hierarchical clustering of aCGH data

Consensus clustering for the cervical cancer CN data



Hierarchical clustering of aCGH data

Assume there are two or three clusters.
Are these clinically interpretable?
Link clustering to tumor stage.

two clusters

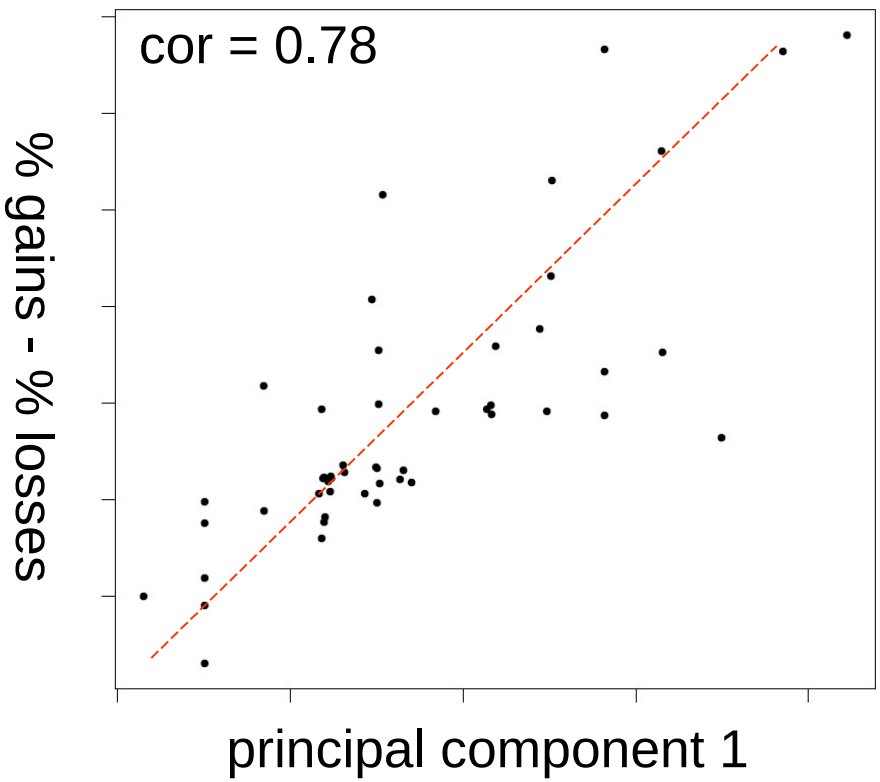
	cluster	
	1	2
stage 1	6	0
stage 2a	6	2
stage 2b	9	15
stage 3	1	11

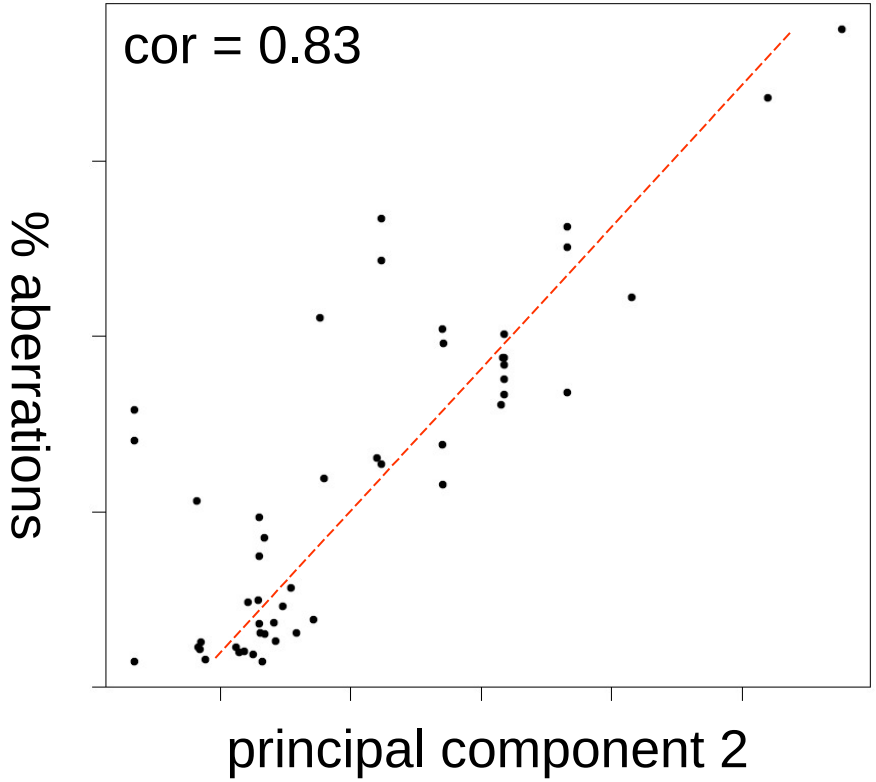
three clusters

	cluster		
	1	2	3
stage 1	6	0	0
stage 2a	6	2	0
stage 2b	9	12	3
stage 3	1	6	5

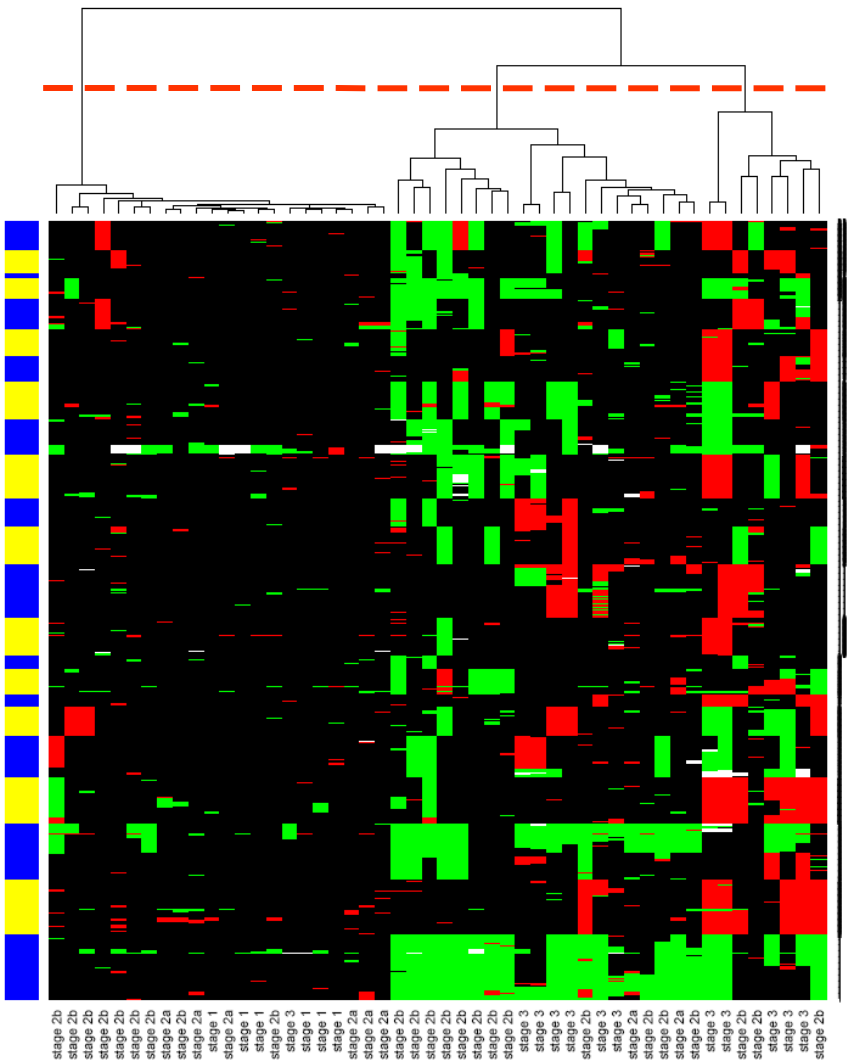
Hierarchical clustering of aCGH data

Principal component analysis for DNA copy number data
(beyond scope of the course)

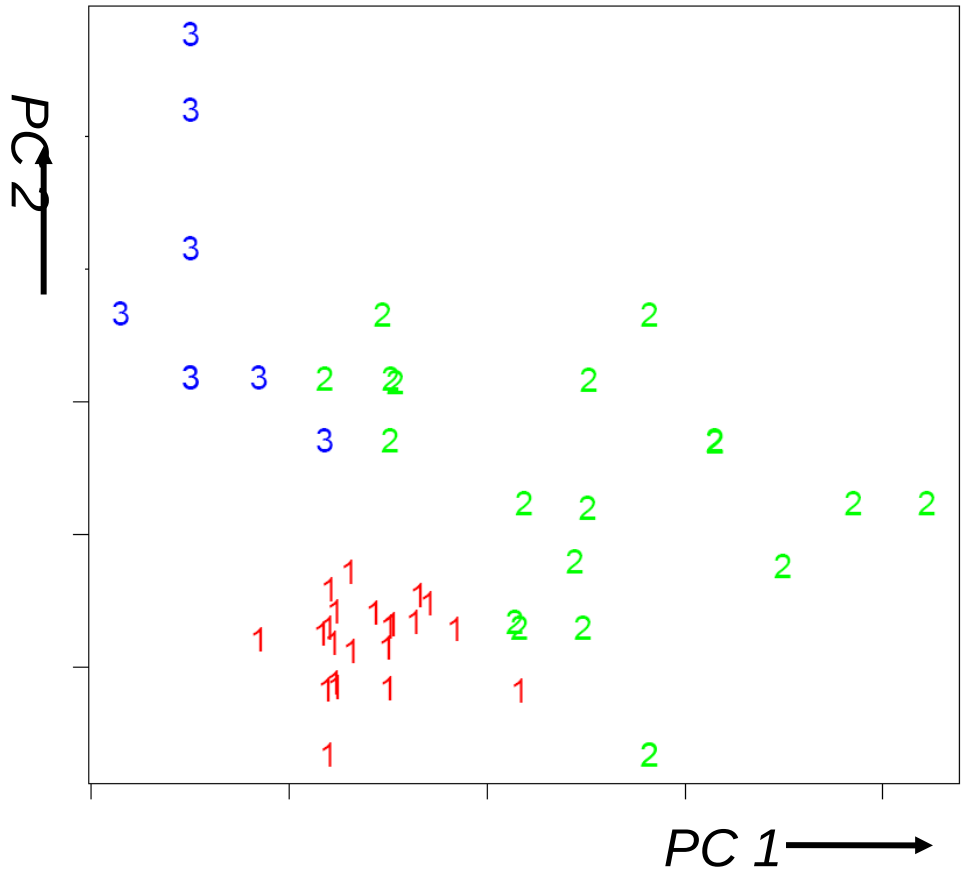




Hierarchical clustering of aCGH data



Found clustering in PCA-plot



Hierarchical clustering of aCGH data

Cervical cancer

Conclusion

We may hypothesize that:

- early tumor stage samples separate from the higher stage tumor samples by the number of aberrations.
- the later tumor stages appear to separate into two groups, along the lines of the loss-gain contrast.