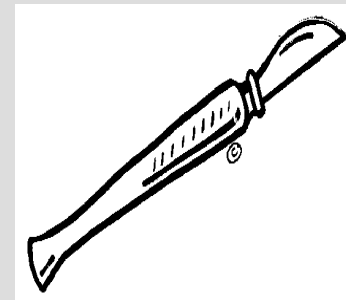

Diagnosis using computers

One disease



Three therapies



Clinical Studies

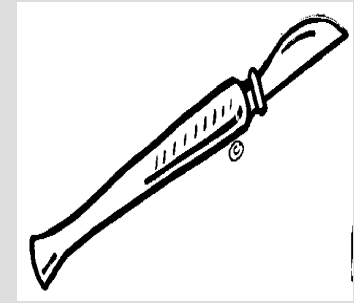
In average



75%



55%



35%

Success

Three subtypes of the disease



A



B



C



A



B



C



100%

60%

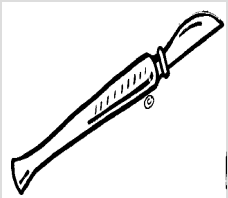
65%



40%

40%

85%



10%

90%

5%



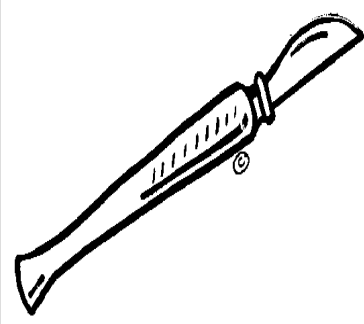
A



100%



B



90%

91,7%

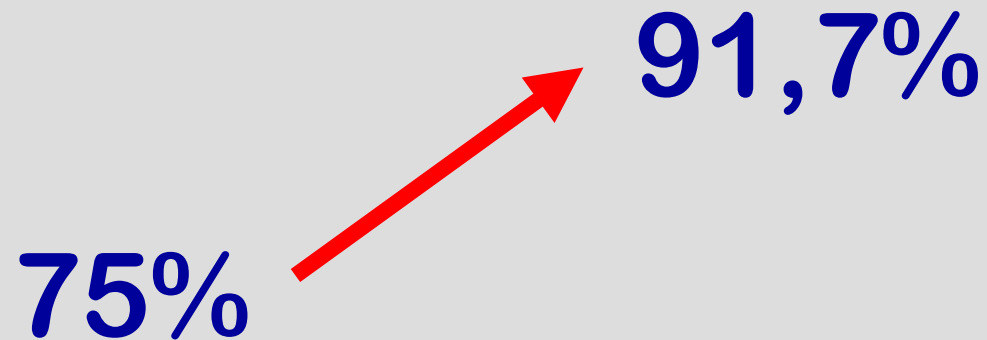


C



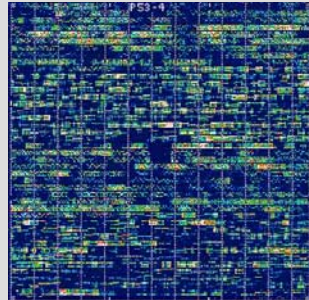
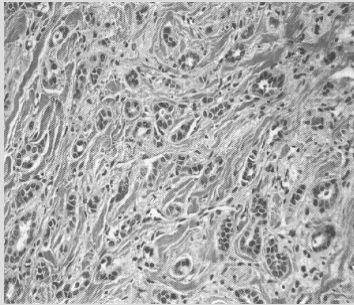
85%

Therapeutic success improved because of the refined diagnosis



Without developing any new therapies

DNA Chip



Tissue

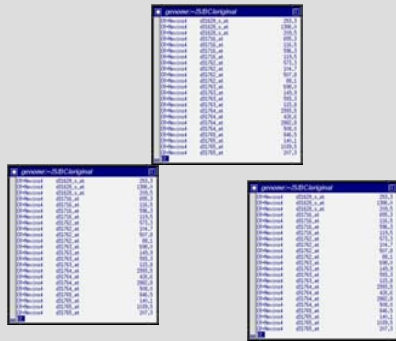
genome:~/ISIBC/original		
ER+Nevins4	d31628_s_at	253.3
ER+Nevins4	d31628_s_at	1386.0
ER+Nevins4	d31628_s_at	209.5
ER+Nevins4	d31716_at	655.3
ER+Nevins4	d31716_at	116.5
ER+Nevins4	d31716_at	596.3
ER+Nevins4	d31716_at	119.5
ER+Nevins4	d31762_at	573.3
ER+Nevins4	d31762_at	104.7
ER+Nevins4	d31762_at	507.8
ER+Nevins4	d31762_at	88.1
ER+Nevins4	d31763_at	698.0
ER+Nevins4	d31763_at	149.9
ER+Nevins4	d31763_at	593.3
ER+Nevins4	d31763_at	115.8
ER+Nevins4	d31764_at	2993.5
ER+Nevins4	d31764_at	426.6
ER+Nevins4	d31764_at	2882.8
ER+Nevins4	d31764_at	508.0
ER+Nevins4	d31765_at	846.5
ER+Nevins4	d31765_at	140.1
ER+Nevins4	d31765_at	1039.5
ER+Nevins4	d31765_at	207.3

Expression
profile

The setup:

100 patients in each arm

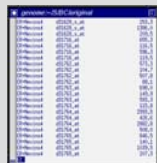
30.000 genes on the chip



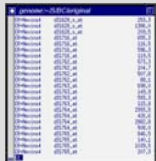
Microarray data table A showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



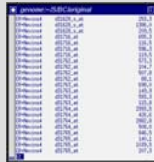
A



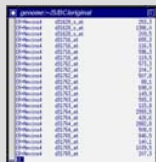
Microarray data table A showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



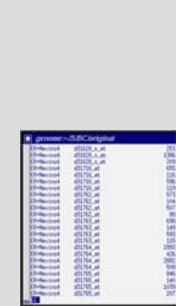
Microarray data table A showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



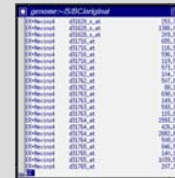
Microarray data table A showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



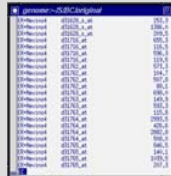
Microarray data table A showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



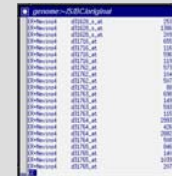
Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



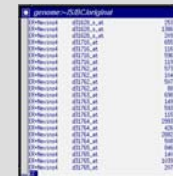
Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



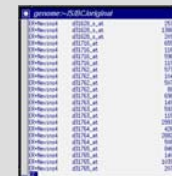
B



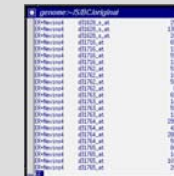
Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



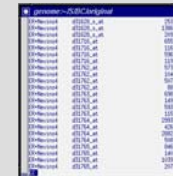
Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.



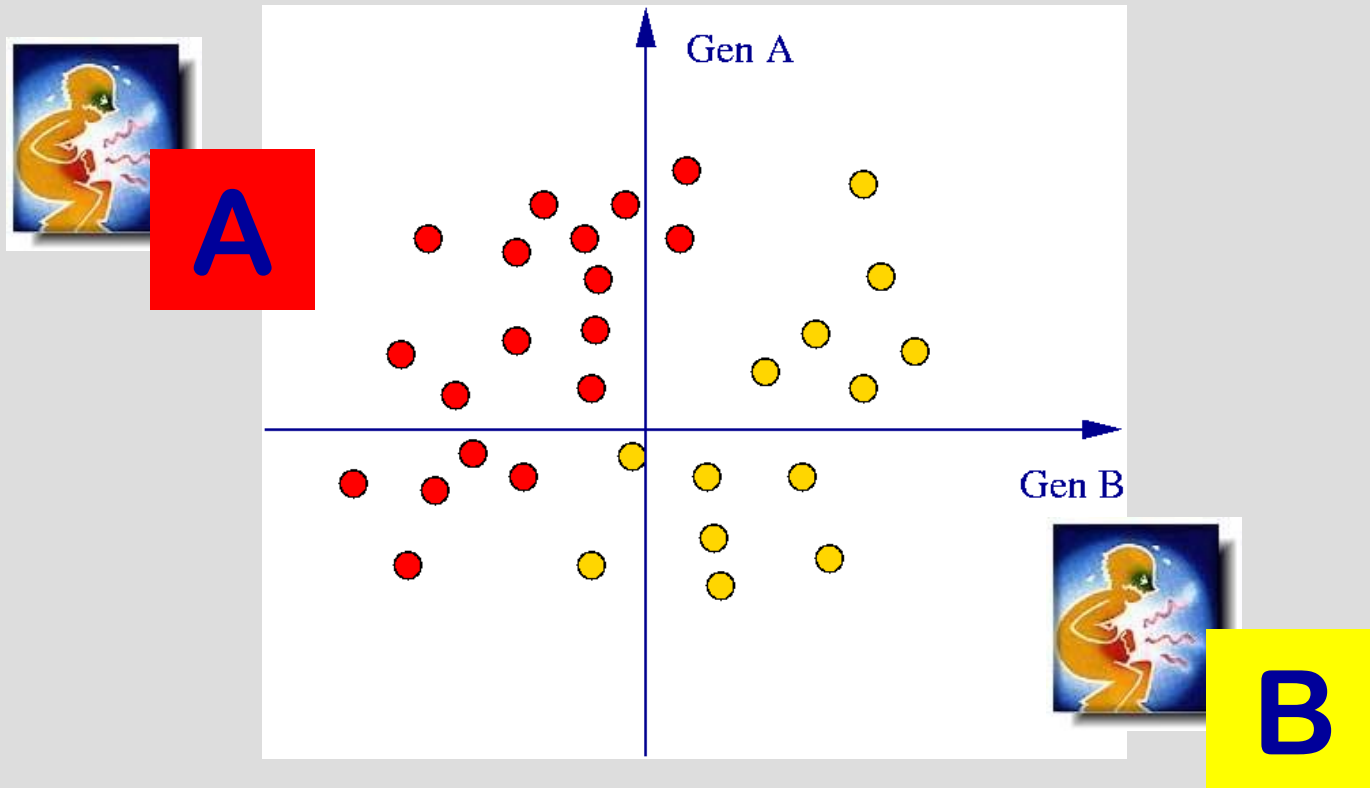
Microarray data table B showing gene expression levels for 30,000 genes across 100 patients. The table has 30,000 rows and 100 columns. The first column lists gene IDs, and the subsequent columns represent individual patients. The data is organized into a grid of 100 columns and 30,000 rows.

**Are there any differences
between the gene
expression profiles of type
A patients and type B
patients?**

**30.000 genes are a lot.
That's too complex to start
with**

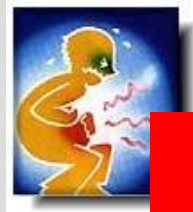
**Let's start with considering
only two genes:
gene A und gene B**

In this situation we can see that ...

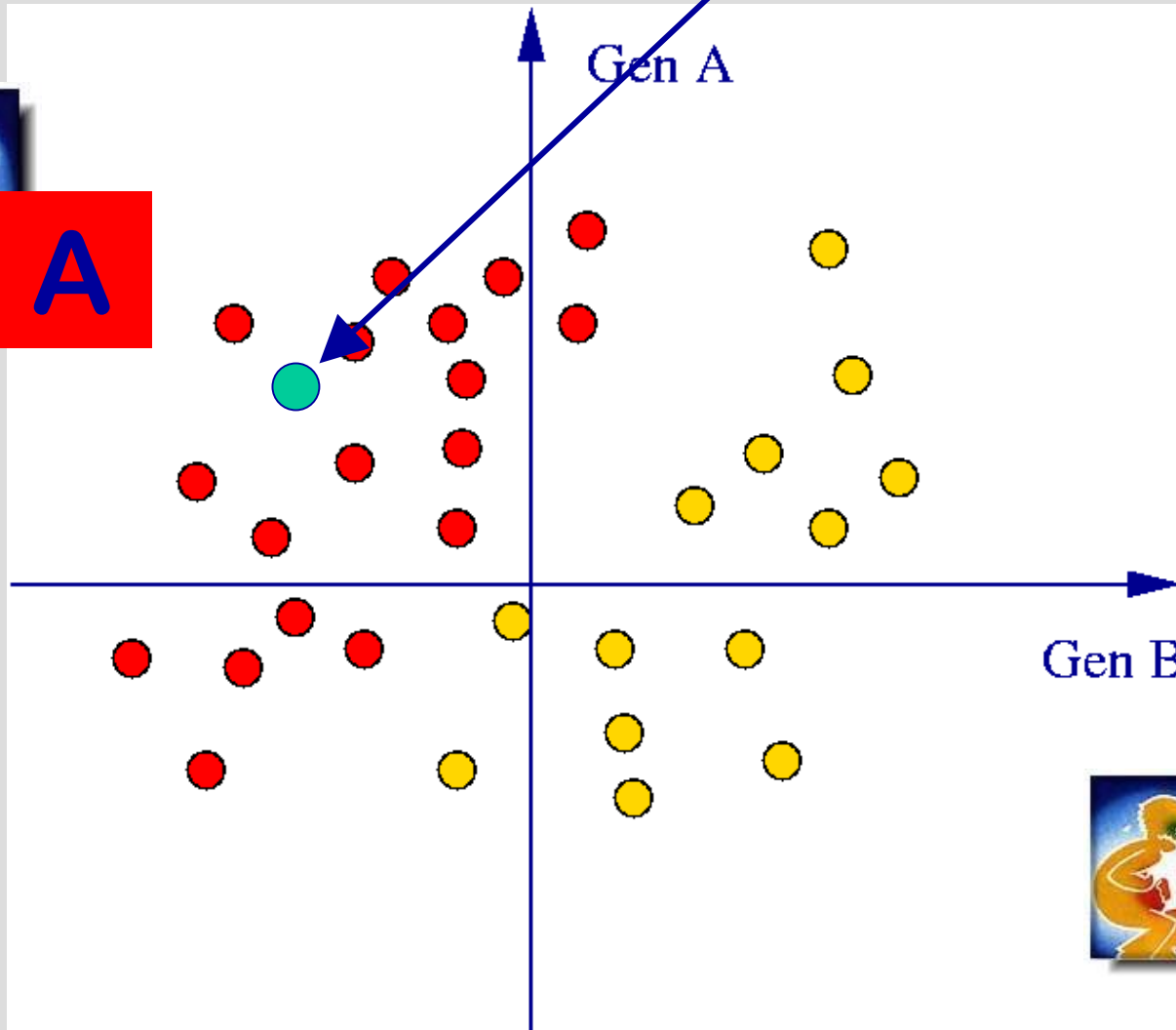


... there is a difference.

A new patient



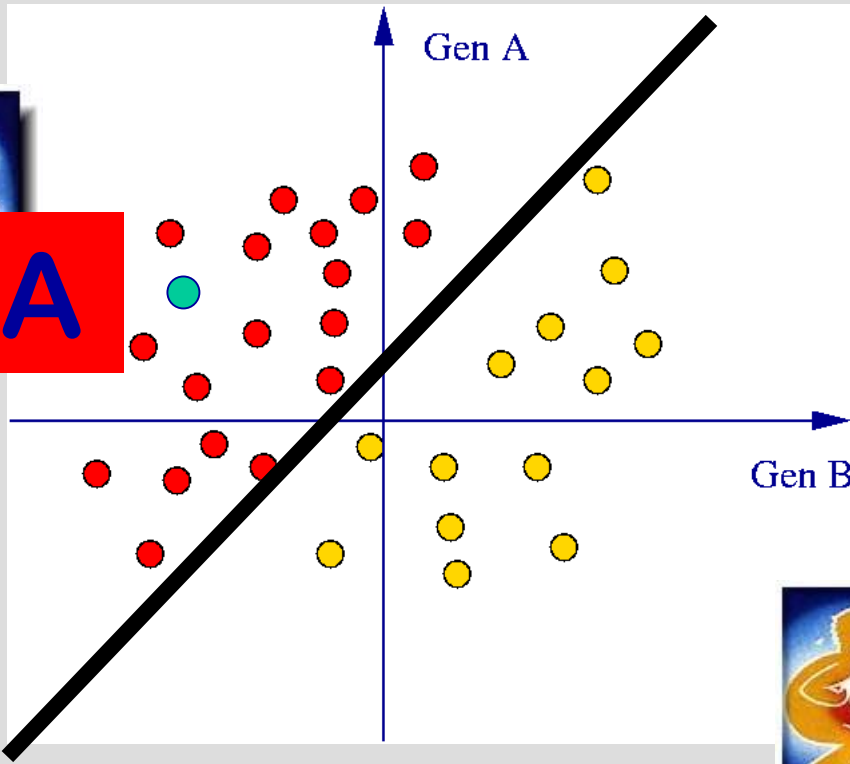
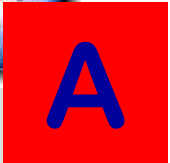
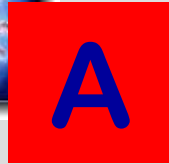
A



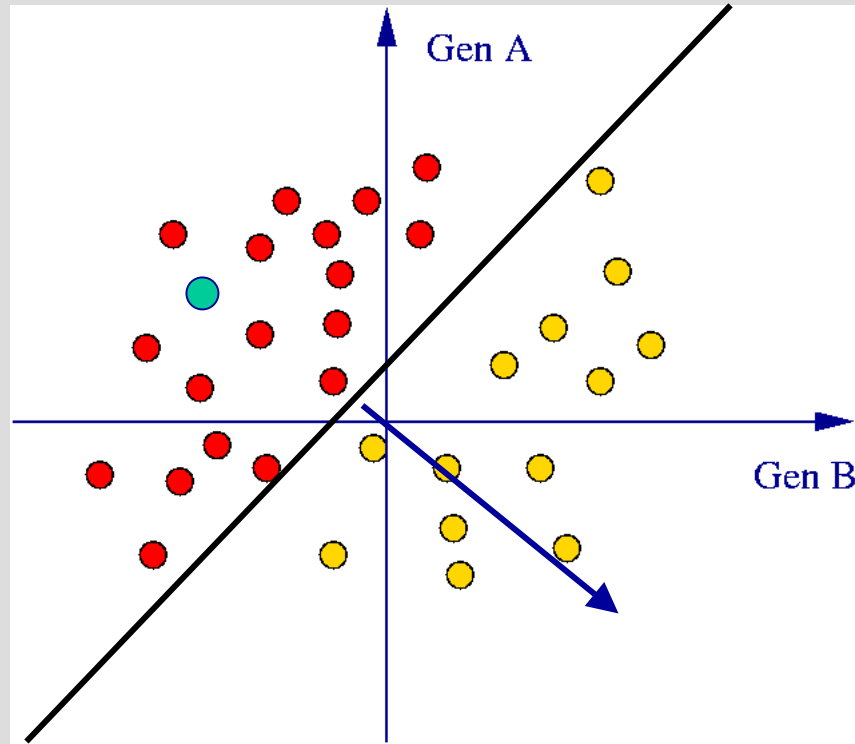
B



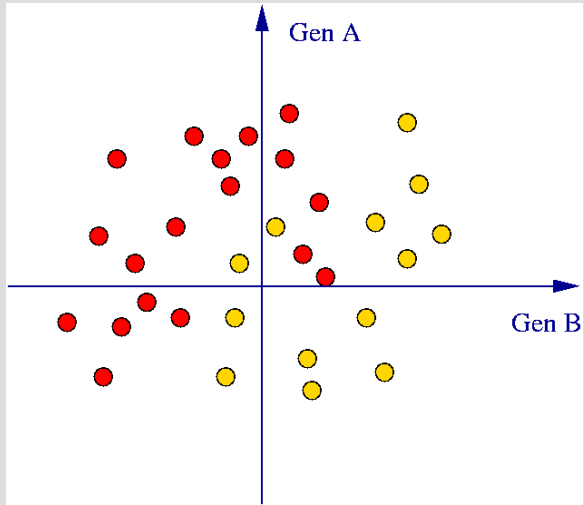
The new patient



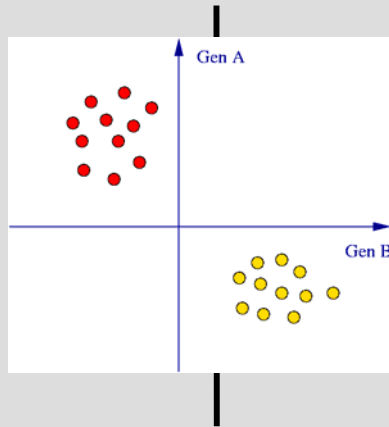
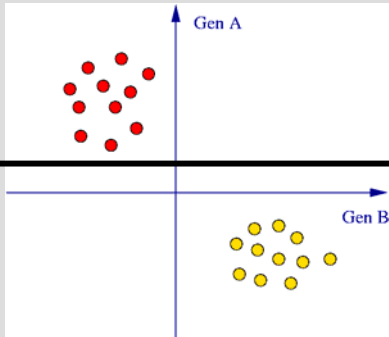
Here everything is clear.



**Unfortunately, expression data is different.
What can go wrong?**

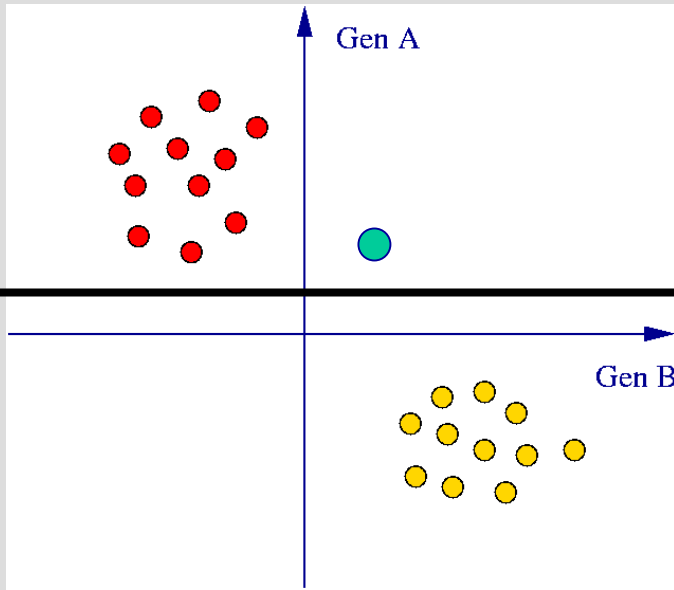


Problem 1:
No separating line

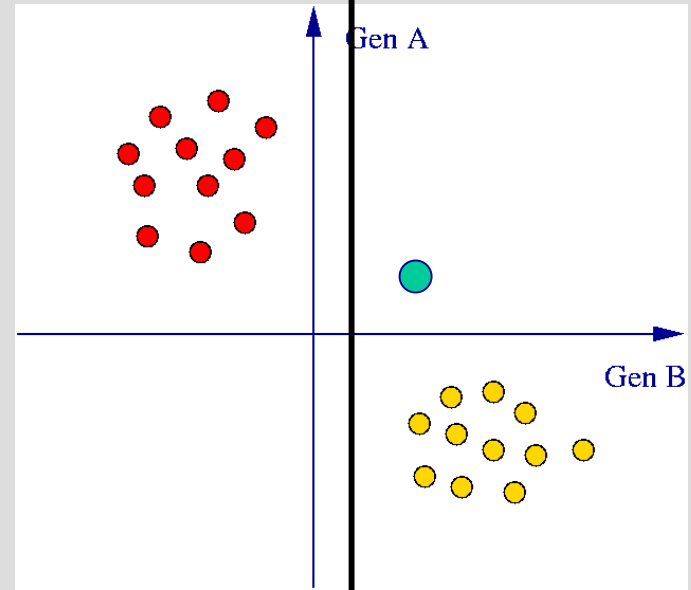


Problem 2:
To many separating lines

New patient ?

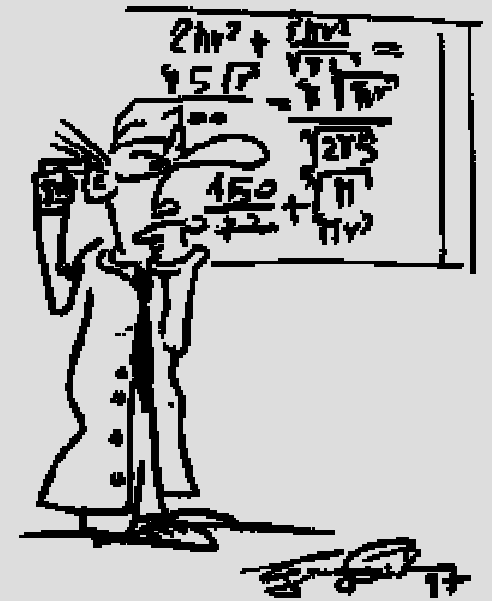
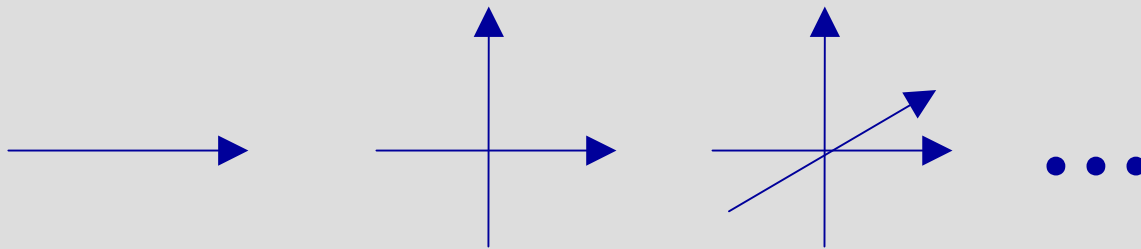


A

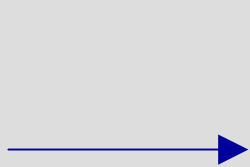


B

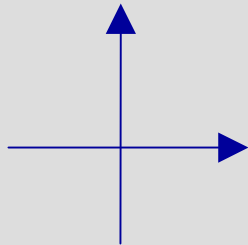
In praxis we look at thousands of genes, generally more genes than patients



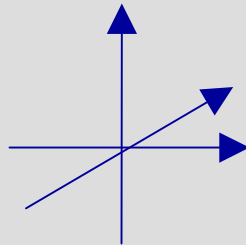
An in 30000 dimensional spaces different laws apply



1



2



3

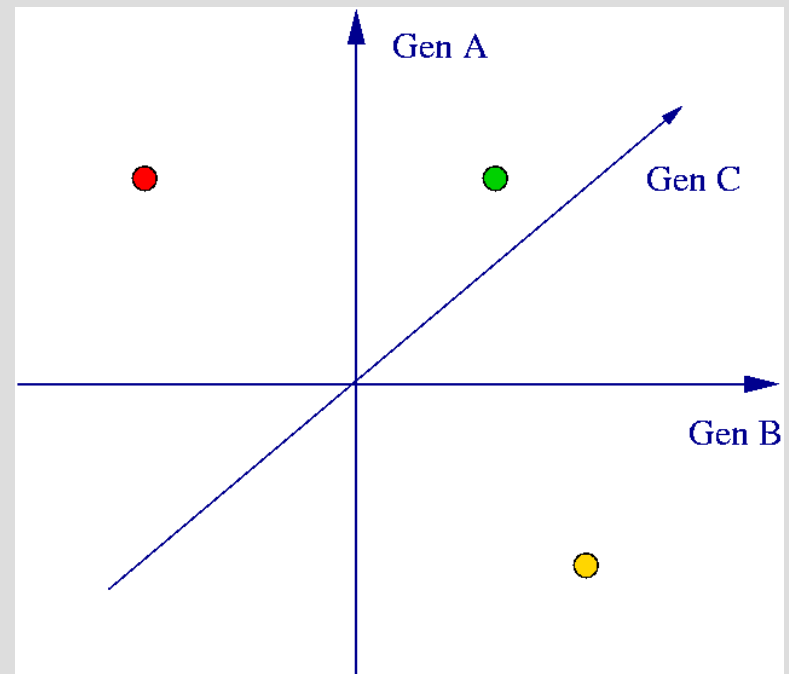
...

30000

- **Problem 1 never exists!**
- **Problem 2 exists almost always!**

Spent a minute thinking about this
in three dimensions

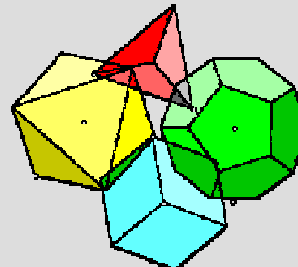
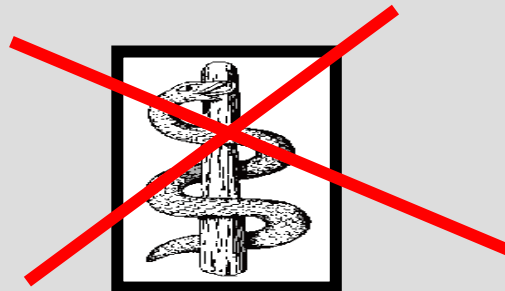
Ok, there are three genes, two
patients with known diagnosis, one
patient of unknown diagnosis, and
separating planes instead of lines



OK! If all points fall onto one line it does not always work. However, for measured values this is very unlikely and never happens in praxis.

From the data alone we can not decide which genes are important for the diagnosis, nor can we give a reliable diagnosis for a new patient

This has little to do medicine. It is a geometrical problem.



In summary:

If you find a separating signature, it does not mean (yet) that you have a nice publication ...

... in most cases it means nothing.



Wait! Believe me!

There are meaningful differences in gene expression. And these must be reflected on the chips.



Ok,OK...

On the one hand we know that there are completely meaningless signatures and on the other hand we know that there must be real disorder in the gene expression of certain genes in diseased tissues.



What are strategies for finding meaningful signatures?

Later we will discuss 2 possible approaches

1. Gene selection followed by linear discriminant analysis, and the PAM program
2. Support Vector Machines

What is the basis for this methods?

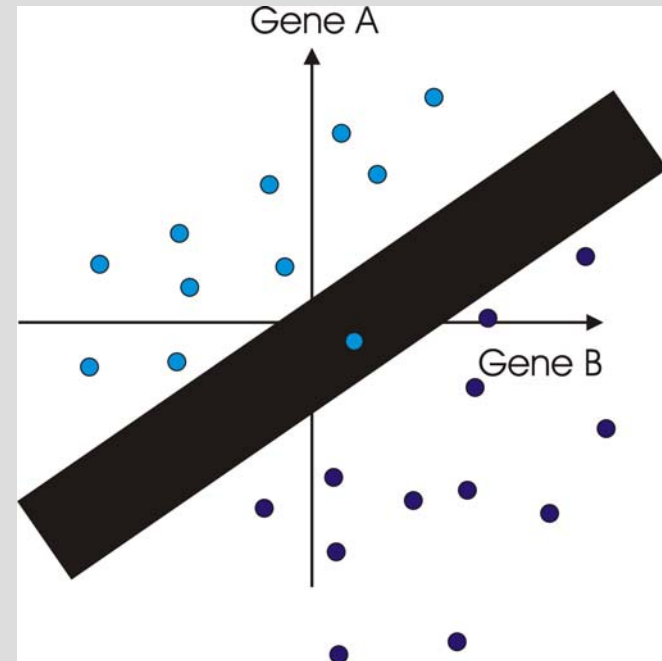
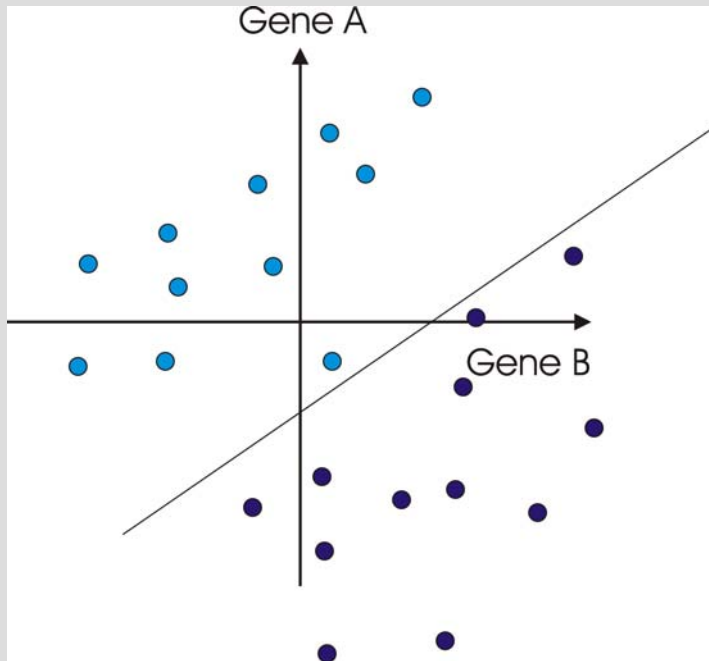


Gene selection

When considering all possible linear planes for separating the patient groups, we always find one that perfectly fits, without a biological reason for this.

When considering only planes that depend on maximally 20 genes it is not guaranteed that we find a well fitting signature. If in spite of this it does exist, chances are good that it reflects transcriptional disorder.

Support Vector Machines



Fat planes: With an infinitely thin plane the data can always be separated correctly, but not necessarily with a fat one.

Again if a large margin separation exists, chances are good that we found something relevant.

Large Margin Classifiers

Both gene selection and Support Vector Machines confine the set of a priori possible signatures. However, using different strategies.

Gene selection wants a small number of genes in the signature (**sparse model**)

SVMs want some minimal distance between data points and the separating plane (**large margin models**)

There is more than you could do ...

Learning Theory

Ridge Regression, LASSO, Kernel based methods, additive Models, classification trees, bagging, boosting, neural nets, relevance vector machines, nearest-neighbors, transduction etc. etc.

Let us start with something simple:

Consider a single gene

a_1, \dots, a_{100} expression levels in group a

b_1, \dots, b_{100} expression levels in group b

$$\bar{a} = \frac{1}{100} (a_1 + \dots + a_{100})$$

$$\bar{b} = \frac{1}{100} (b_1 + \dots + b_{100})$$

c expression level of a patient
with unknown diagnosis

Compare $|c - \bar{a}|$ and $|c - \bar{b}|$

Diagnosis : a if $|c - \bar{a}| < |c - \bar{b}|$

b if $|c - \bar{a}| \geq |c - \bar{b}|$

Both groups are summarized by the mean gene expression in this

Diagnosis is according to the closest mean

Consider two genes:

$a_{1,1}, \dots, a_{1,100}, a_{2,1}, \dots, a_{2,100}$ group a

$b_{1,1}, \dots, b_{1,100}, b_{2,1}, \dots, b_{2,100}$ group b

$$\bar{a} = (\bar{a}_1, \bar{a}_2)$$

$$\bar{b} = (\bar{b}_1, \bar{b}_2)$$

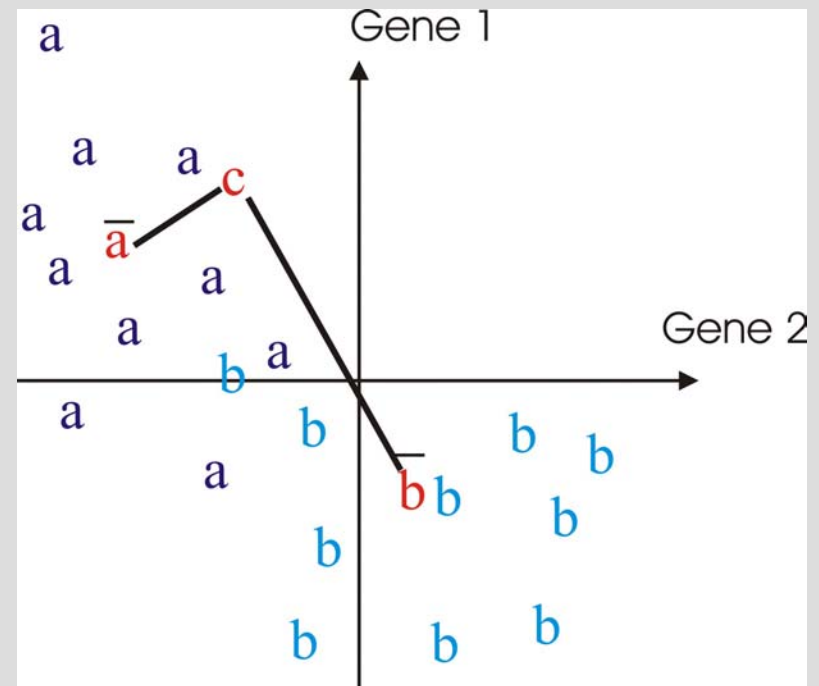
$c = (c_1, c_2)$ Patient without diagnosis

Compare: $d_a = (\bar{a}_1 - c_1)^2 + (\bar{a}_2 - c_2)^2$ and

$$d_b = (\bar{b}_1 - c_1)^2 + (\bar{b}_2 - c_2)^2$$

Diagnosis: a if $d_a < d_b$

b else



Many (N) genes:

$a_{i,j}$ Gene i in Patient j from group a

$b_{i,j}$ Gene i in Patient j from group b

$$\bar{a} = (\bar{a}_1, \dots, \bar{a}_N)$$

$$\bar{b} = (\bar{b}_1, \dots, \bar{b}_N)$$

c_1, \dots, c_N Patient without diagnosis

Compare distances to the centroids :

$$d_a = \sum_{i=1}^N (\bar{a}_i - c_i)^2$$

$$d_b = \sum_{i=1}^N (\bar{b}_i - c_i)^2$$

Diagnosis : a if $d_a < d_b$

b else

Nearest Centroid Method

(Plain Vanilla)

Patient groups are
modelled separately by
centroids

Diagnosis is according
to the nearest centroid
in euclidean distance

$a_{i,j}$ gene i in patient j from group a

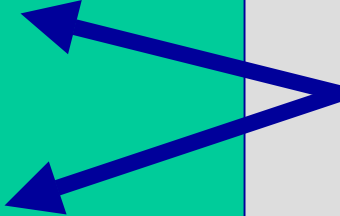
$b_{i,j}$ gene i in patient j from group b

$$d_a = \sum_{i=1}^N (\bar{a}_i - c_i)^2$$

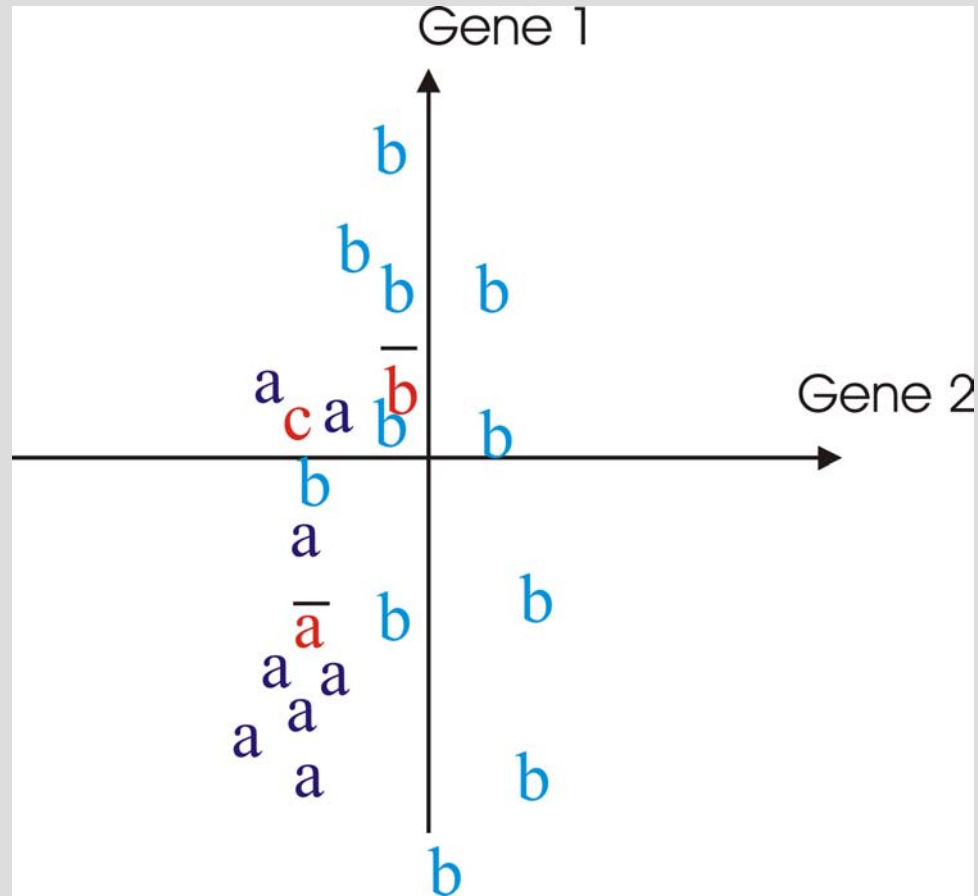
$$d_b = \sum_{i=1}^N (\bar{b}_i - c_i)^2$$

Diagnosis : a if $d_a < d_b$
b else

All N genes
contribute equally
to the diagnosis ...

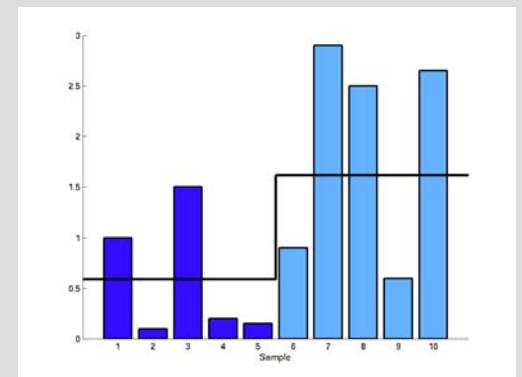
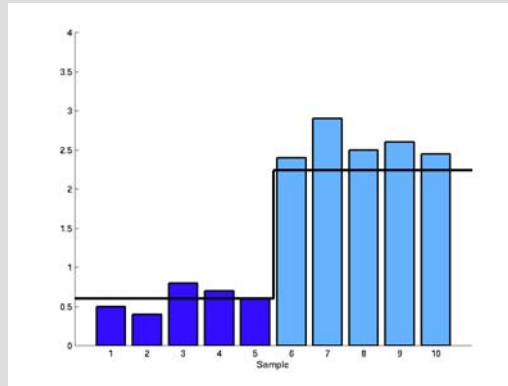
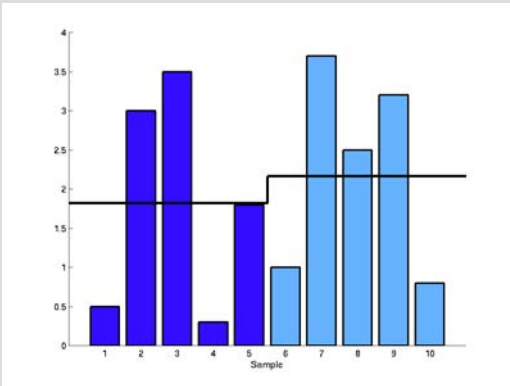


... that is a problem



Genes with a small „variance“ should get more weight than genes with high variance

$$d_a = \sum_{i=1}^N w_i (\bar{a}_i - c_i)^2 \quad d_b = \sum_{i=1}^N w_i (\bar{b}_i - c_i)^2$$



Use the pooled within class variance ... instead of the overall variance

The variances need to be estimated

$$\sigma_i^2 = \frac{1}{n-2} \sum_{j=1}^{n/2} (a_{i,j} - \bar{a}_i)^2 + (b_{i,j} - \bar{b}_i)^2$$

pooled in class variance

In our case :

$$n = 200$$

→ SAM

$$w_i = (\sigma_i + \sigma_0)^2$$

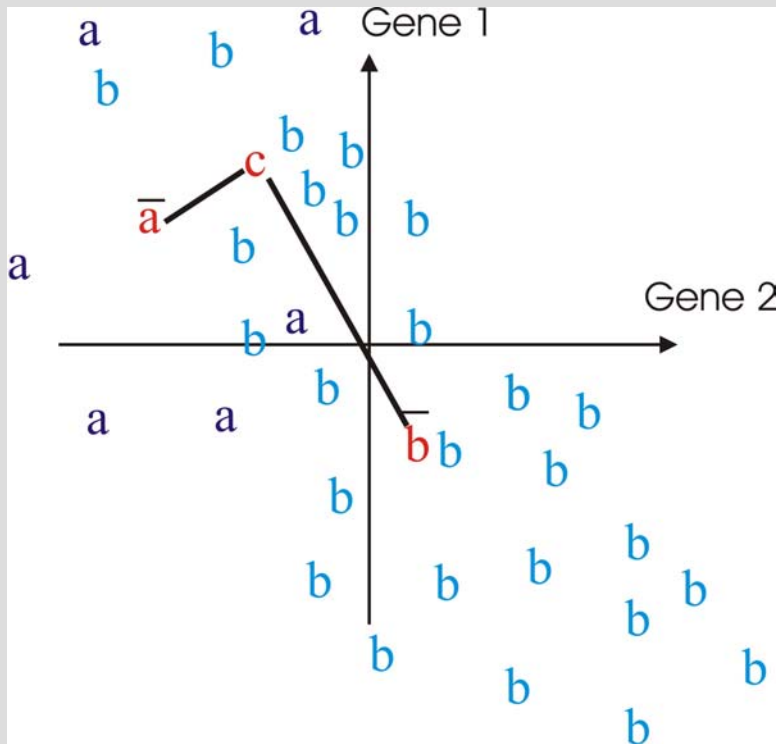
$$\sigma_0^2 = \text{median}(\sigma_1^2, \dots, \sigma_N^2)$$

The estimated variance is not the true variance. It can be higher or lower. If a small variance is underestimated σ_i^2

can be very small and w_i is unnaturally high.

While this is a rare event for a fixed gene it happens quite often if we are looking for 30000 genes

Is c an a or a b?



Is closer to the a centroid but there much more b than a samples

If this reflects the true population, than c should be classified as b

Baseline correction

π_a = relative size of group a
i.e. relative frequency of type a
samples in the study, or expert
knowledge

$$\pi_b = 1 - \pi_a$$

$$d_a(c) = \sum_{i=1}^N \frac{(\bar{a}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_a$$

$$d_b(c) = \sum_{i=1}^N \frac{(\bar{b}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_b$$

Discriminant Score

distance to the centroid

$$d_a(c) = \sum_{i=1}^N \frac{(\bar{a}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_a$$

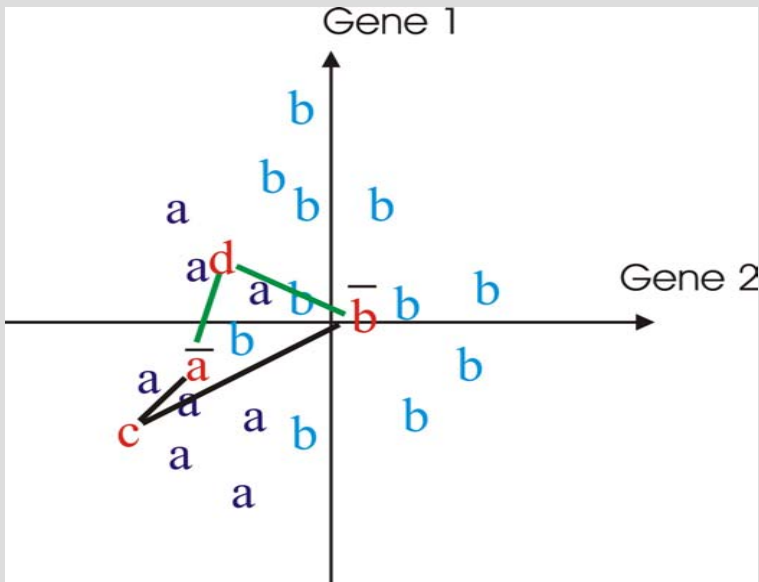
baseline correction

$$d_b(c) = \sum_{i=1}^N \frac{(\bar{b}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2 \log \pi_b$$

pooled within class variance

variance regularization parameter

Classification probabilities



Both c and d are diagnosed as group a

But for d that was a close decision

$$\text{Prob} [Group(c) = a] = \frac{e^{-\frac{1}{2}d_a(c)}}{e^{-\frac{1}{2}d_a(c)} + e^{-\frac{1}{2}d_b(c)}}$$

$$\text{Prob} [Group(c) = b] = 1 - \text{Prob} [Group(c) = a]$$

Putting things into context

$d_a(c) = d_b(c)$ is a linear plane

We are still using all the 30000 genes

→ Overfitting problem

The plane is not necessarily optimal in terms of separation

This might be an advantage or a disadvantage

Variable selection

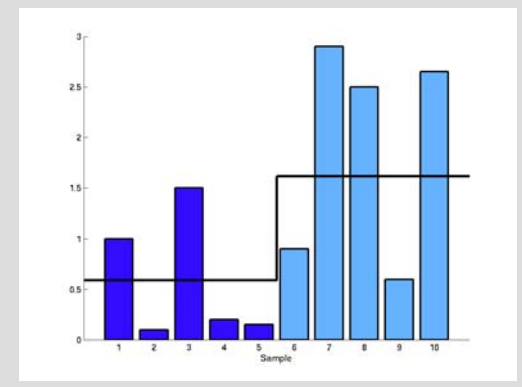
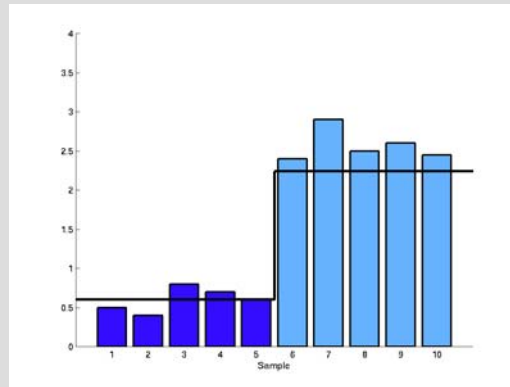
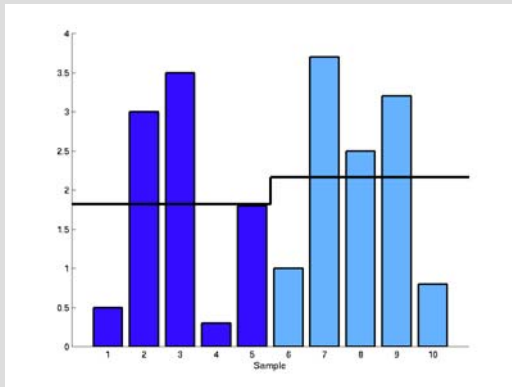
30000 genes are to many

They may cause overfitting

They introduce noise ... there weights are low ... but if there are many ...

They can not all matter

→ Choose genes:



Choose the genes with the highest weights
regularized t-score a la SAM

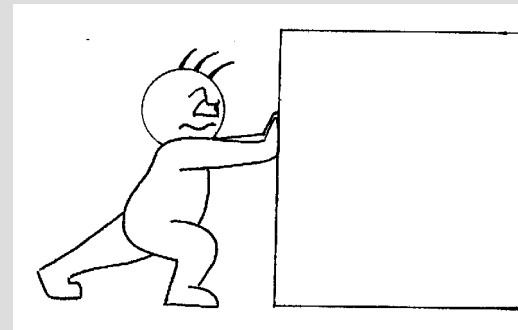
Hard thresholding vs. soft thresholding

Lets say we pick the top 100 genes

Gene Nr. 100 is in but gene Nr. 101 is not,

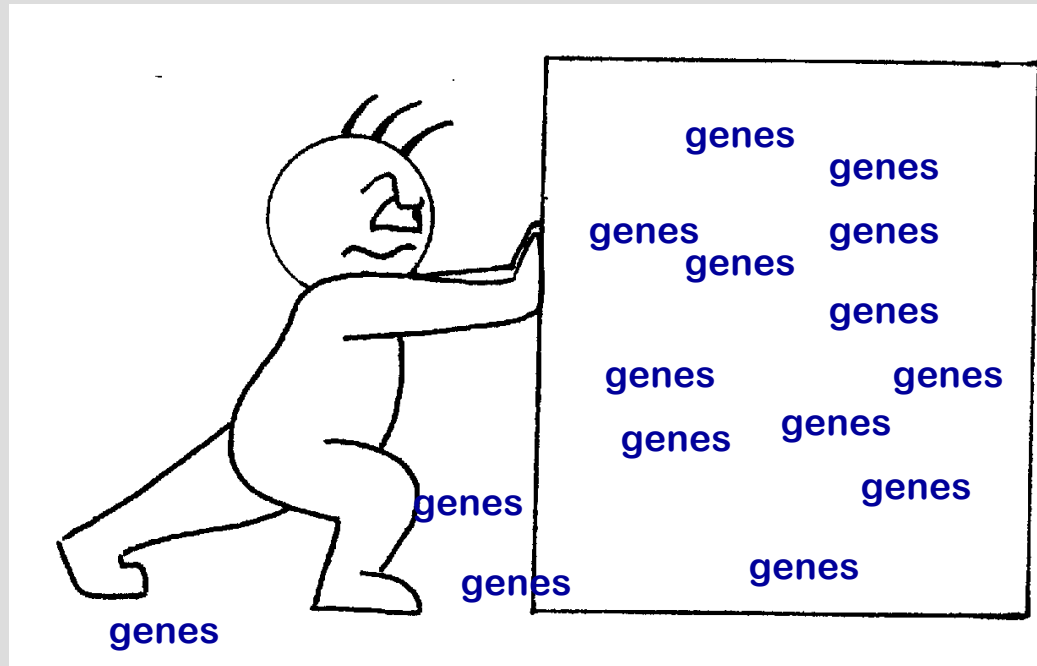
however, both genes are almost equally informative

If you want to get rid of genes you can chop them off
or slowly push them out



The shrunken centroid method and the PAM program

Tibshirani et al 2002



Idea

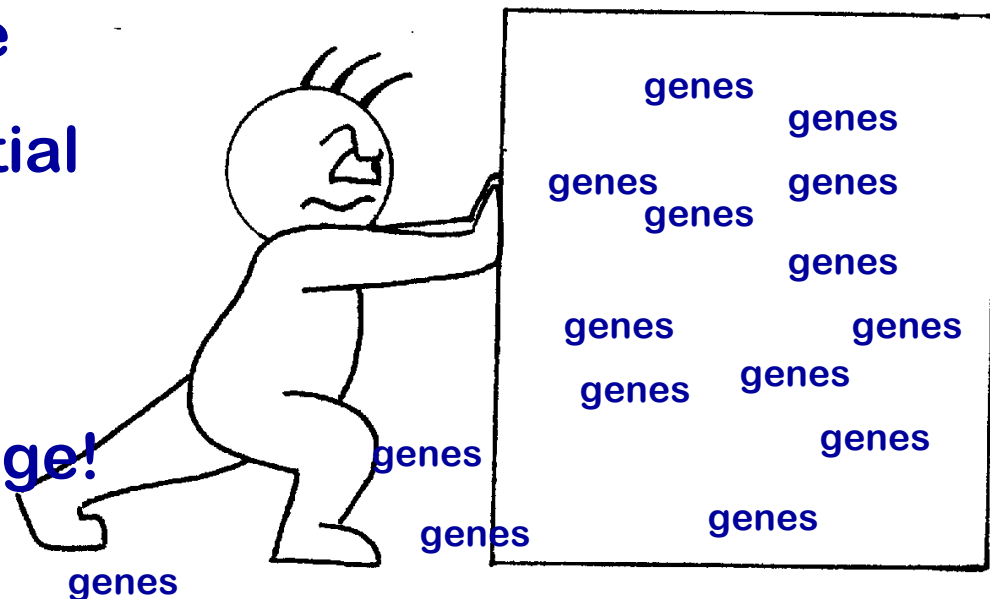
Genes with high weights are influential for diagnosis

Genes with lower weights are less influential for diagnosis

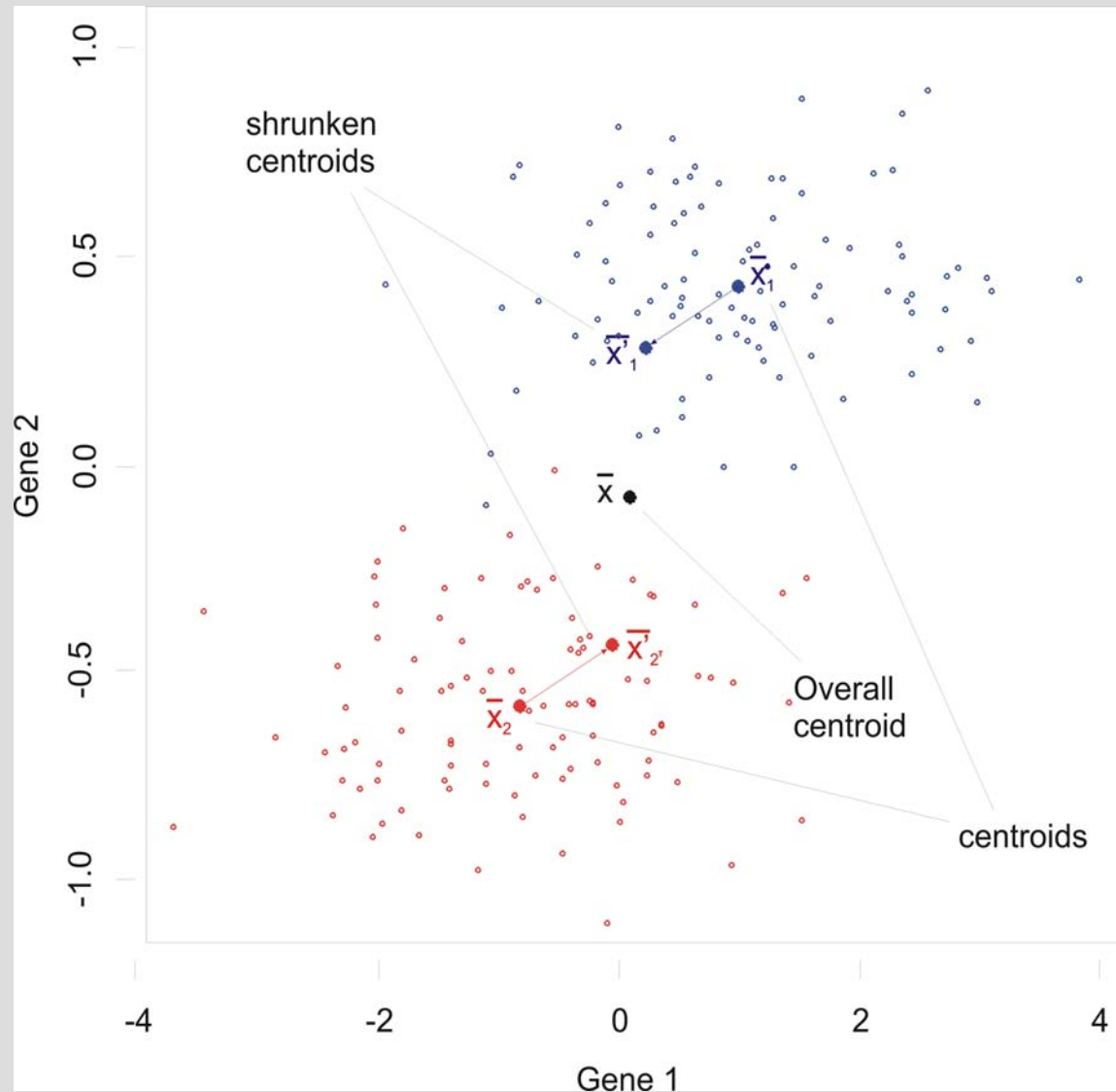
Genes that are excluded can not be influential for diagnosis at all

Before you exclude a gene totally from analysis make it continuously less influential for the diagnosis

How? By centroid shrinkage!



Centroid shrinkage



Notation

\bar{a}_i mean of gene i in group a

\bar{b}_i mean of gene i in group b

\bar{x}_i mean of gene i using all data

Let

$$D_{i,a} = \frac{\bar{a}_i - \bar{x}_i}{m_a(\sigma_i + \sigma_0)}, \quad m_a = \sqrt{1/n_a + 1/n}$$

$$D_{i,b} = \dots$$

or

$$\bar{a}_i = \bar{x}_i + m_a(\sigma_i + \sigma_0) D_{i,a}$$

$$\bar{b}_i = \dots$$

group centroid

overall centroid

scaling factor

$$\bar{a}_i = \bar{x}_i + m_a (\sigma_i + \sigma_0) D_{i,a}$$

offset

$$\bar{a}_i = \bar{x}_i + m_a (\sigma_i + \sigma_0) D'_{i,a}$$

shrunk offset

$$D'_{i,a} = \text{sign}(D_{i,a}) (|D_{i,a}| - \Delta)_+$$

shrinkage parameter

$(\dots)_+ = \text{truncation at zero}$

The amount of shrinkage is controlled by Delta

Little shrinkage many genes are still contributing to the centroids

High shrinkage only few genes are still in the analysis

The amount of shrinkage can be determined by

cross validation ... we will discuss this later

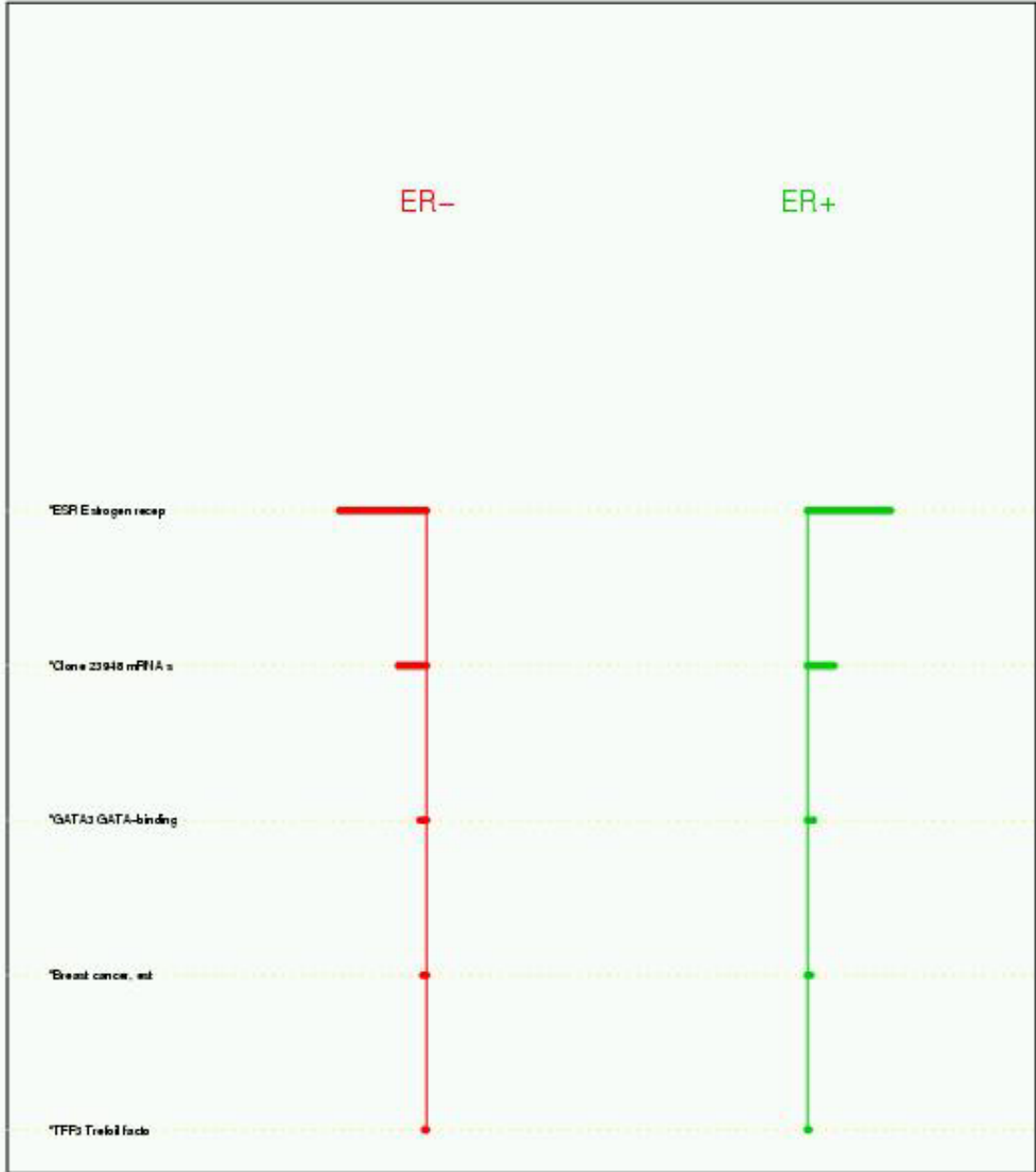


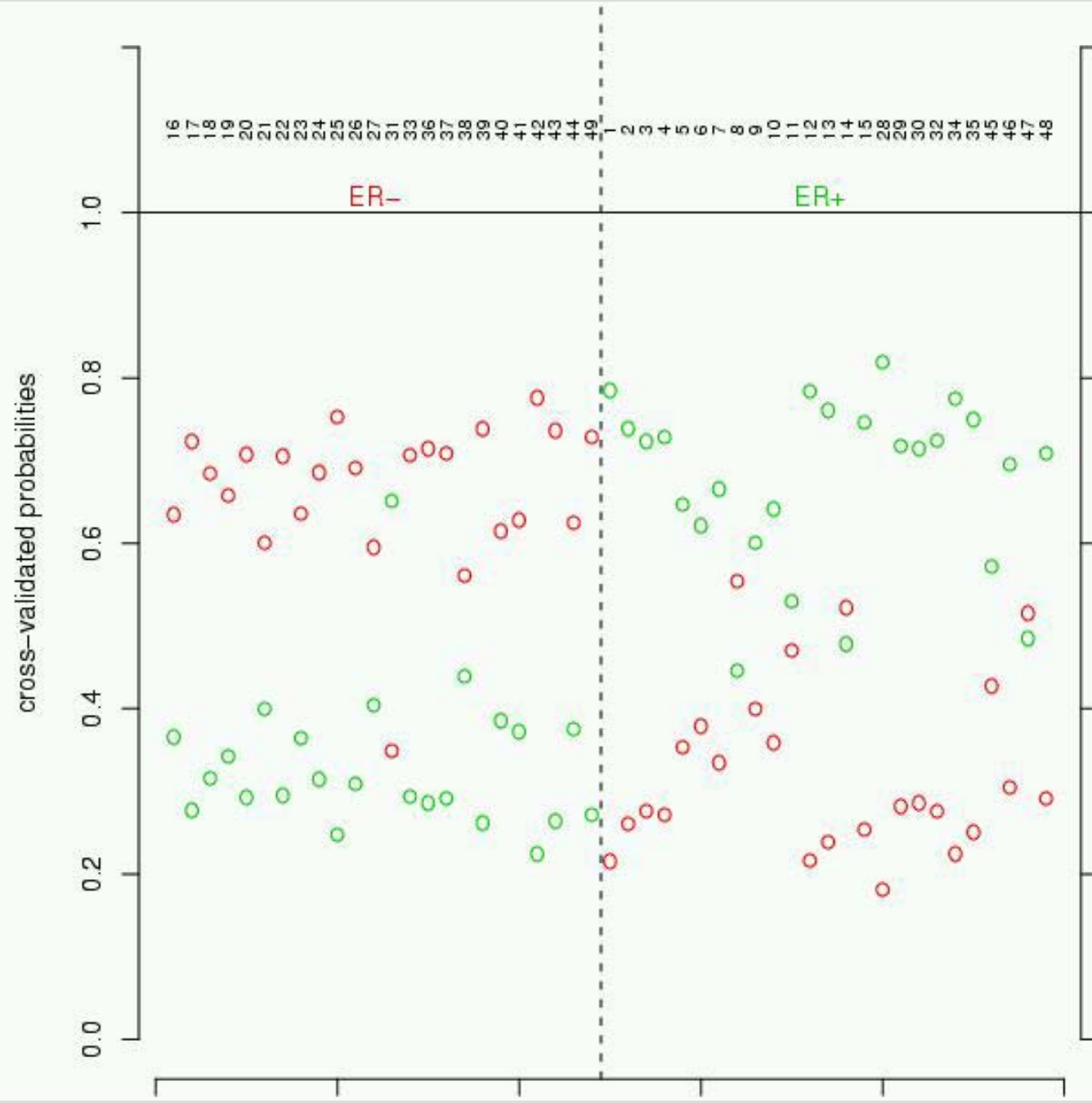
Estrogen Receptor Status

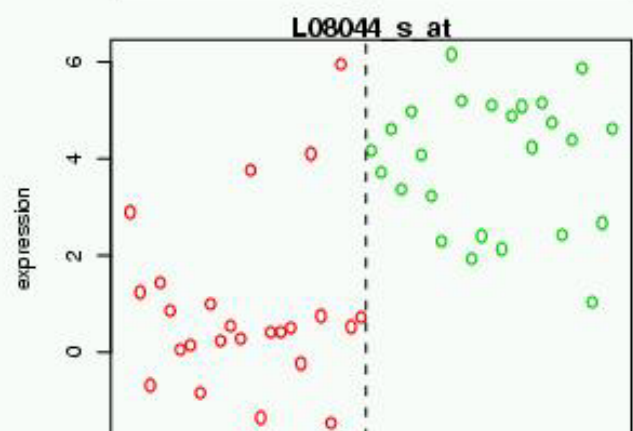
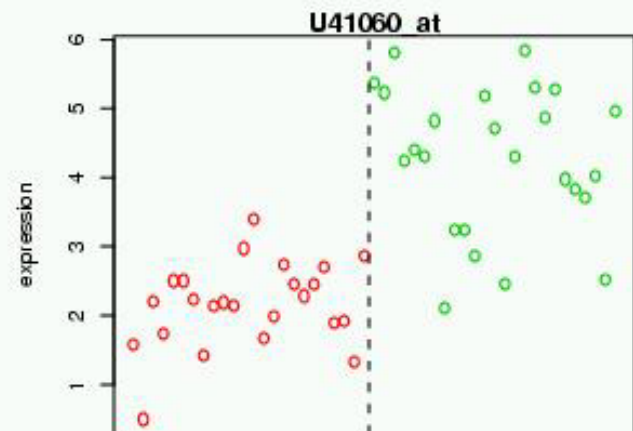
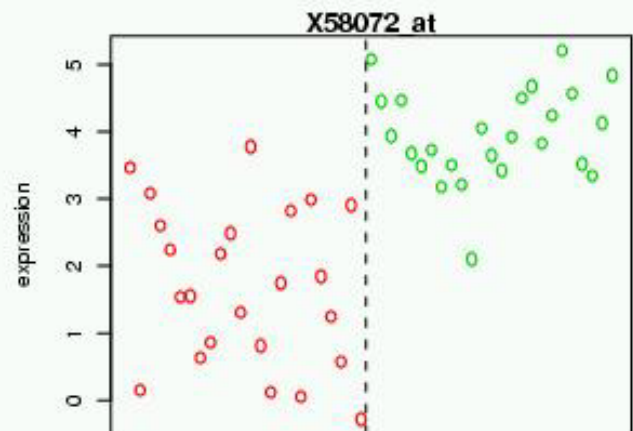
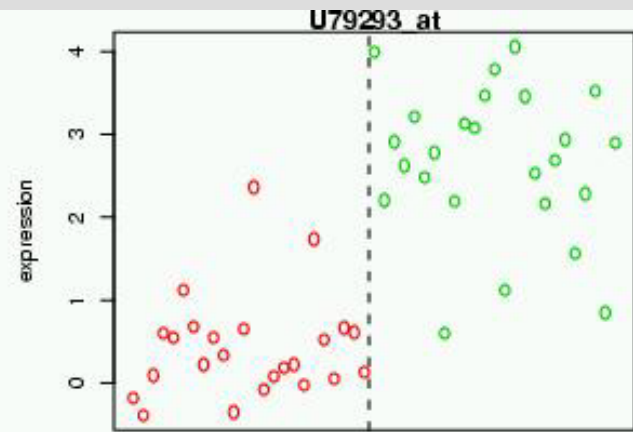
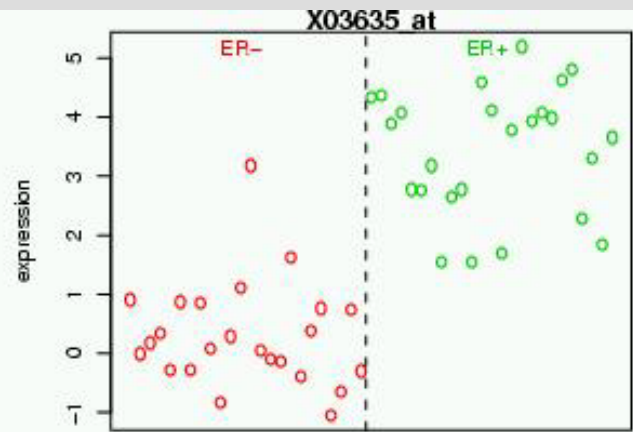
- 7000 genes
- 49 breast tumors
- 25 ER+
- 24 ER-

ER-

ER+







Devices of regularization used by PAM

- Gene selection

- Shrinkage

- Gene selection by screening (no wrapping)

- The weight of a gene only depends on the gene and not on its interaction with others

- Use of a baseline depending on the population size of the groups ... more information in addition to the expression data

Questions



Coffee

