

---

# Group testing: global tests

---

Ulrich Mansmann  
IBE, Medical School, University of Munich

---

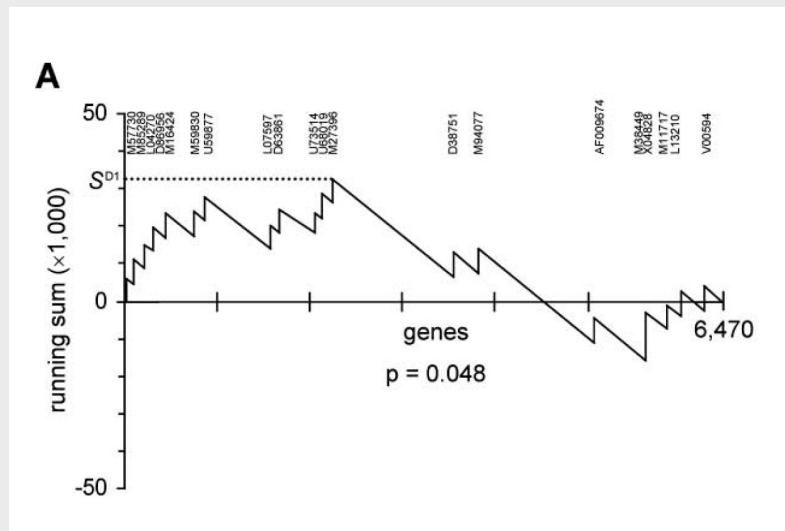
## Content of the lecture

---

- **How to assess the relevance of groups of genes:**
  - outstanding gene expression in a specific group compared to other genes;
  - differential gene expression not of single genes but over a specific group of genes.
- **How to define gene-groups:**
  - exploratory research produces functional groups and genomic signatures: confirm the relevance of the specific group.
  - Bioinformatic algorithms can be used to define pathways and functional groups

## Example I: Cyclin D1 Action

- Lamb J et al. (2003) *A mechanism of Cyclin D1 Action Encoded in the Patterns of Gene Expression in Human Cancer*, Cell, 114: 323-334
- Cyclin D1 expression signature: cyclin D1 target gene set.
- Cyclin D1 activity in Human Tumors: Does the cyclin D1 target gene set play a prominent role in different tumor entities? Being present as highly expressed genes.



---

## The ideas behind the analysis

---

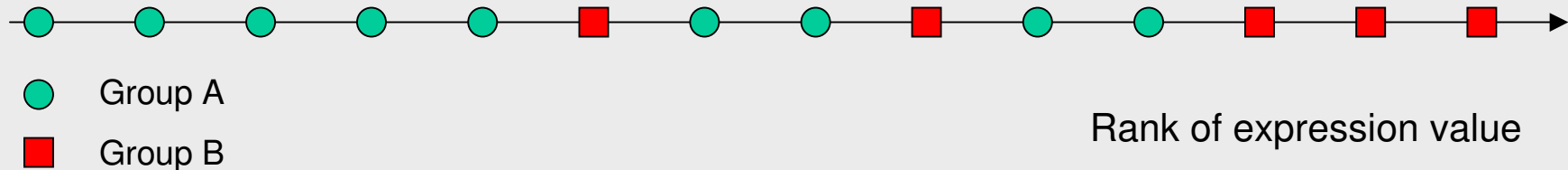
### Problem:

Two groups of genes have to be compared with respect to gene expression: Is the gene expression in gene group A different from the expression in gene group B. **Important: Genes in both groups are different!**

### Basic idea:

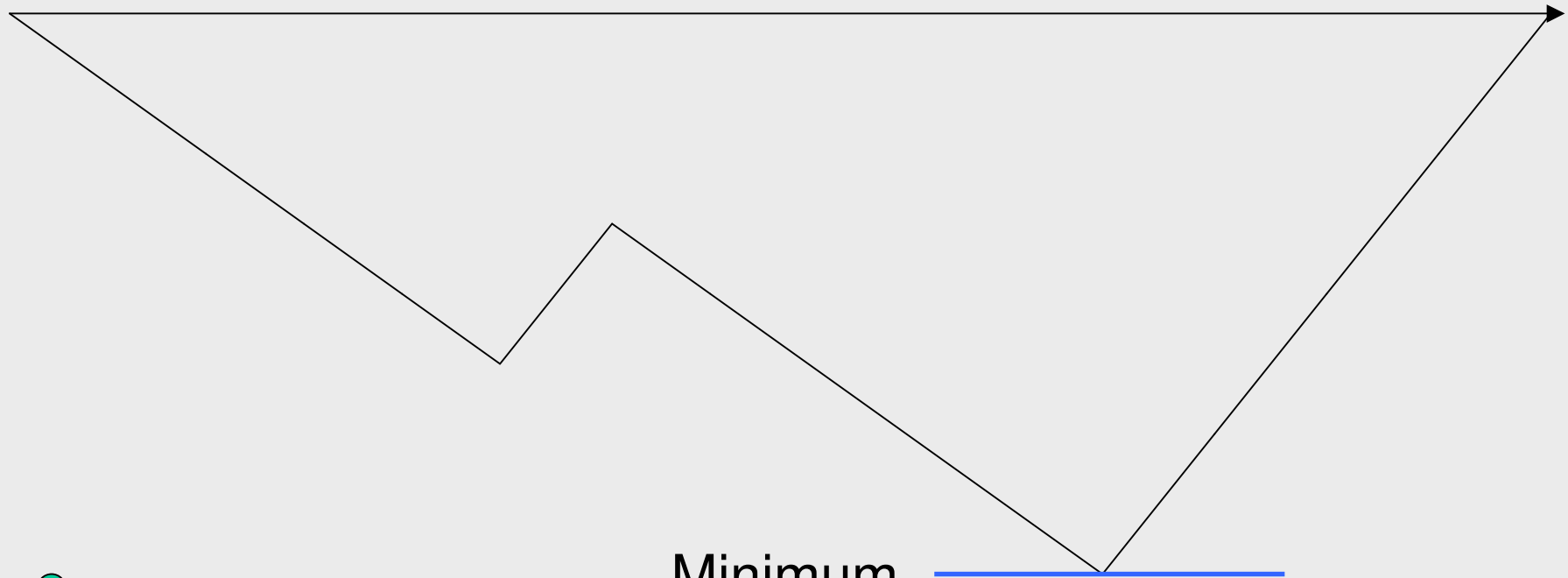
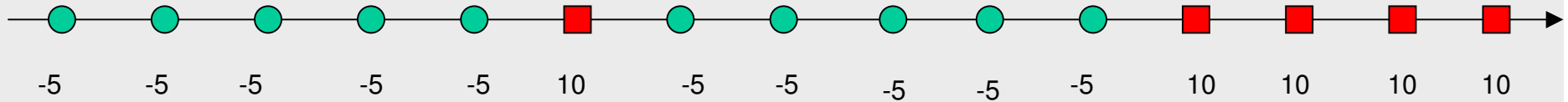
$n_A$  genes in group A,  $n_B$  genes in group B

Order the genes with respect to the expression value. If there is a difference in expression level between both groups, the expression values will be separated. **The position of a value in group A will have the tendency to be in general high or low.** In case of no difference, the values will be nicely mixed.



# The ideas behind the analysis

Genes ordered by rank of expression



- Group A ( $n_A=10$ )
- Group B ( $n_B=5$ )

Minimum

Is the minimum extreme with respect to random group mixing?

Group testing

---

## The algorithm formalized

---

### Basic idea:

- $n_A$  genes in group A,  $n_B$  genes in group B.
- Order the genes with respect to expression values.
- Create a vector  $vv$  of  $(n_A+n_B)$  components with value  $-n_B$  at each position where a value from group A is sitting and with value  $n_A$  at each position where a value from group B is sitting.
- Calculate  $yy = \text{cumsum}(vv)$ .
- Draw a line starting at  $(0,0)$  through points  $(i, yy[i])$ . The line will end in  $(n_A+n_B, 0)$  because  $(-n_B) \cdot n_A + n_A \cdot n_B = 0$ .
- Look at  $M_{vv} = \max\{|\min(yy)|, \max(yy)\}$  which will be large in case of a good separation between both groups.
- Permute the vector  $vv$  to get  $vv^*$ , calculate  $yy^*$  and  $M_{vv^*}$ . Use permutation to calculate the distribution of  $M_{vv}$  under the Null hypothesis, determine the permutation based p-value:  $p_{\text{perm}} = \#\{M_{vv^*} \geq M_{vv}\} / \# \text{ permutations}$ .

---

## Example II: Colon Cancer

---

**Study:** 18 patients with UICC II colon cancer, 18 patients with UICC III colon cancer, HG-U133A, 22,283 probesets representing ~18,000 genes. Snap-frozen material, laser microdissection.

**Question 1:** Are there specific cancer related pathways with a more distinct differential gene expression between UICC II/III?

---

## Gene set enrichment – Colon cancer

---

1407 probe sets are studied which belong to 9 cancer specific pathways.

androgen_receptor_signalling	122
apoptosis	245
cell_cycle_control	51
notch_delta_signalling	50
p53_signalling	45
ras_signalling	316
tgf_beta_signalling	100
tight_junction_signalling	425
wnt_signalling	214



---

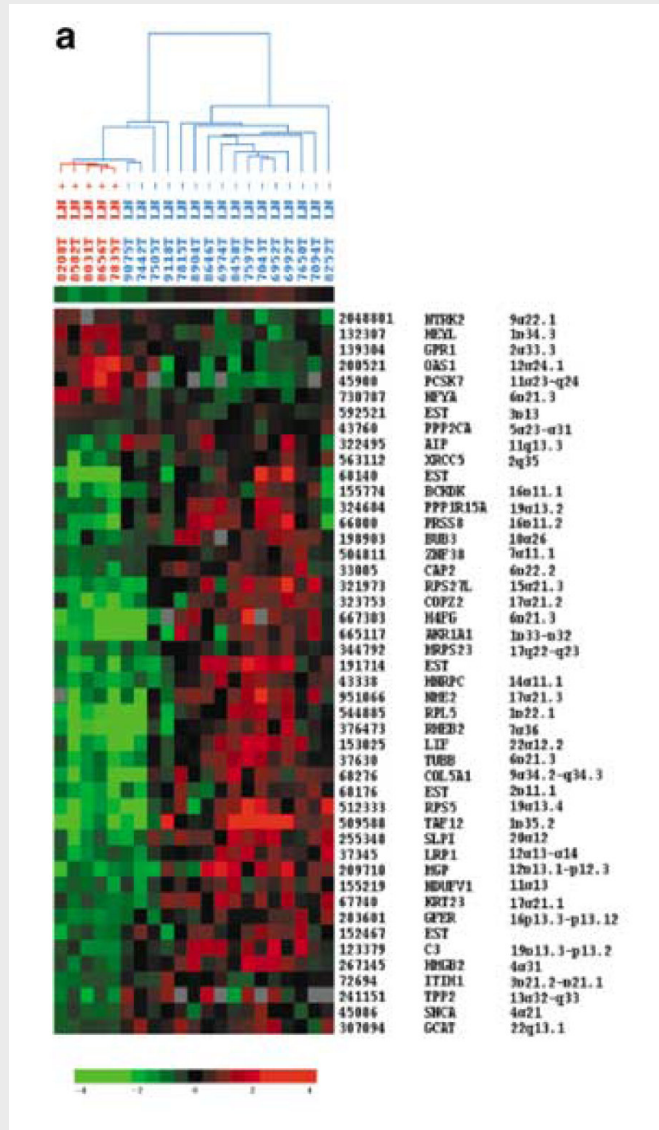
## Gene set enrichment – Colon cancer

---

	group.A	group.B	$M_{yy}$	p.value
androgen_receptor_signaling	118	1289	6983	0.0568
Apoptosis	238	1169	17801	0.7438
cell_cycle_control	51	1356	10413	0.3616
notch_delta_signalling	50	1357	9010	0.6492
p53_signalling	45	1362	12390	0.0924
ras_signalling	311	1096	15486	0.6252
tgf_beta_signaling	100	1307	22615	0.0128
tight_junction_signaling	406	1001	15456	0.4414
wnt_signaling	214	1193	16318	0.8432

Restriction of the analysis to genes in cancer specific pathways

## Example III: Lymph node metastases



Bertucci F et al. (2004) *Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters*, Oncogene 23, 1377–1391

Bertucci et al present a gene signature consisting of 46 genes which is claimed to be able to discriminate between LN- and LN+ colorectal cancer.

Is it possible to prove with new data that the signature has discriminative value. Can we reject the Nullhypothesis

$$P[Y|X] = P[Y].$$

Y : LN+/LN-

X: expression pattern of 46 genes

Group testing

---

## Goeman's Global Test

---

- Test if global expression pattern of a group of genes is significantly related to some outcome of interest (groups, continuous phenotype).
- If this relationship exists, then the knowledge of gene expression helps to improve the prediction of the phenotype of interest. If the prediction can not improved by knowing the gene expression then there will not be differential gene expression.

- Test statistic:

$$Q \sim (Y-\mu)'R (Y-\mu)$$

$$\sim \sum [X_i'(Y-\mu)]^2 \quad \text{sum over genes of the pathway}$$

$$\sim \sum \sum R_{ij}(Y_i-\mu) (Y_j-\mu) \quad \text{sum over subjects}$$

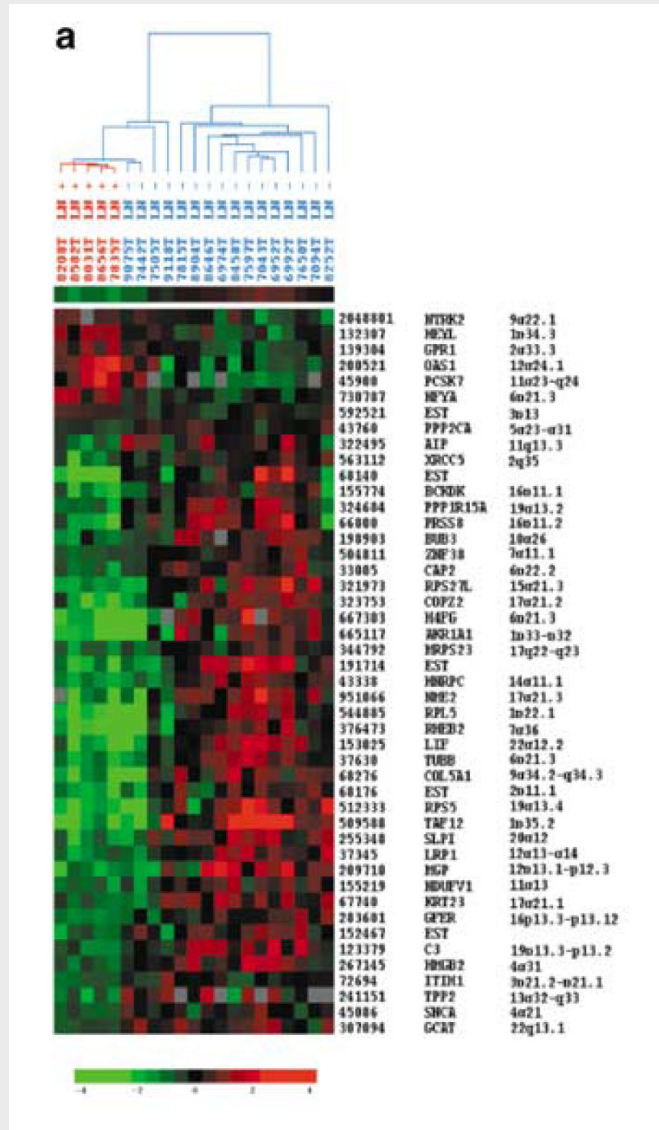
$\mu$ : Mean of phenotype,

$X_{mi}$  Expression for gene  $m$  in subject  $i$

$R$  :  $X'X$  matrix of correlations between gene expression of subjects

Goeman JJ. Et al. (2003) *A global test for groups of genes: Testing association with a clinical outcome*, Bioinformatics, 20:93-99; Bioconductor package: *globaltest*

## Example III: Lymph node metastases



Test is not significant ( $p=0.43$ )

No clear answer on the predictive power of the signature.

No evidence for a difference is not evidence for no difference!

Question of power

Reasons for a non-significance: bad experiment or ...?

---

## Example II: Colon Cancer

---

**Study:** 18 patients with UICC II colon cancer, 18 patients with UICC III colon cancer, HG-U133A, 22,283 probesets representing ~18,000 genes. Snap-frozen material, laser microdissection.

**Question 2:** Is there differential gene expression in the p53 signalling pathway between UICC II and UICC III colon cancer?

---

## Goeman's Global Test – Example II

---

- Test for differential gene expression in *p53 signalling* pathway  
45 probesets

- Global Test result:

45 out of 45 genes used; 36 samples

p value = 0.0114

based on 10000 permutations

Test statistic  $Q = 11.78$

with expectation  $EQ = 5.466$

and standard deviation  $sdQ = 2.152$  under the null hypothesis

- Informative plots:

Sample plot: how good fits a sample to its phenotype

Checkerboard: Correlation between samples

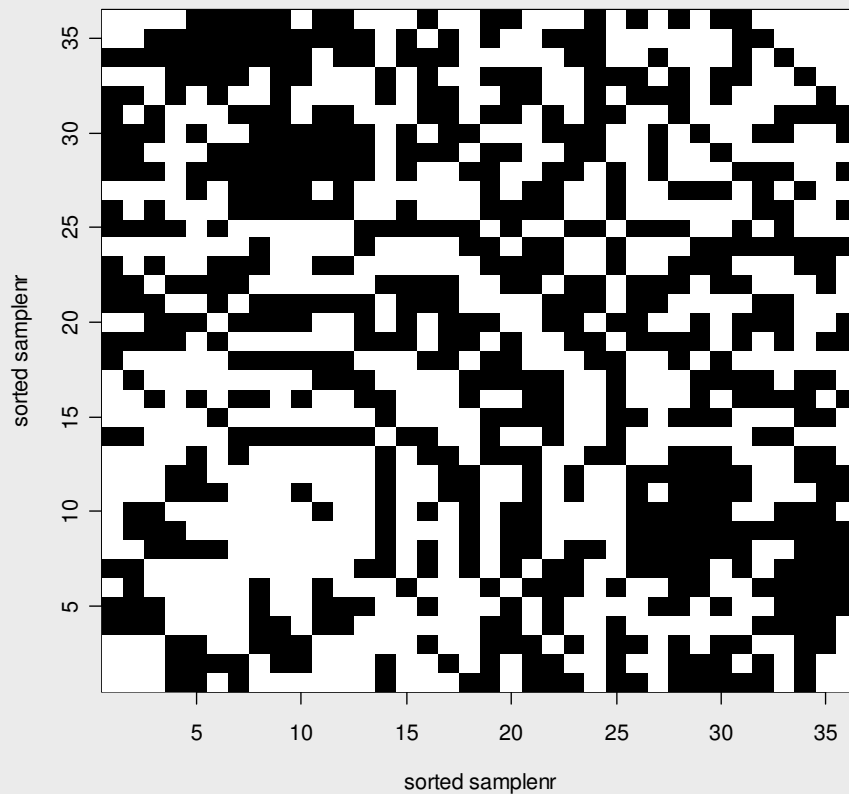
Gene plot: Influence of single genes to test statistics

---

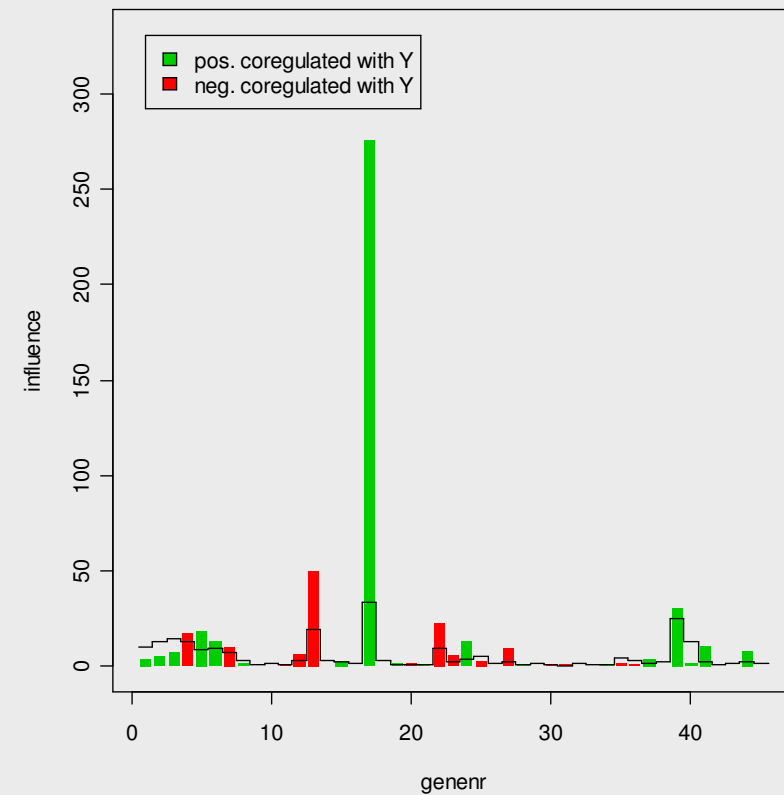
## Goeman's Global Test – Example II

---

$$\sum \sum R_{ij}(Y_i - \mu)(Y_j - \mu)$$



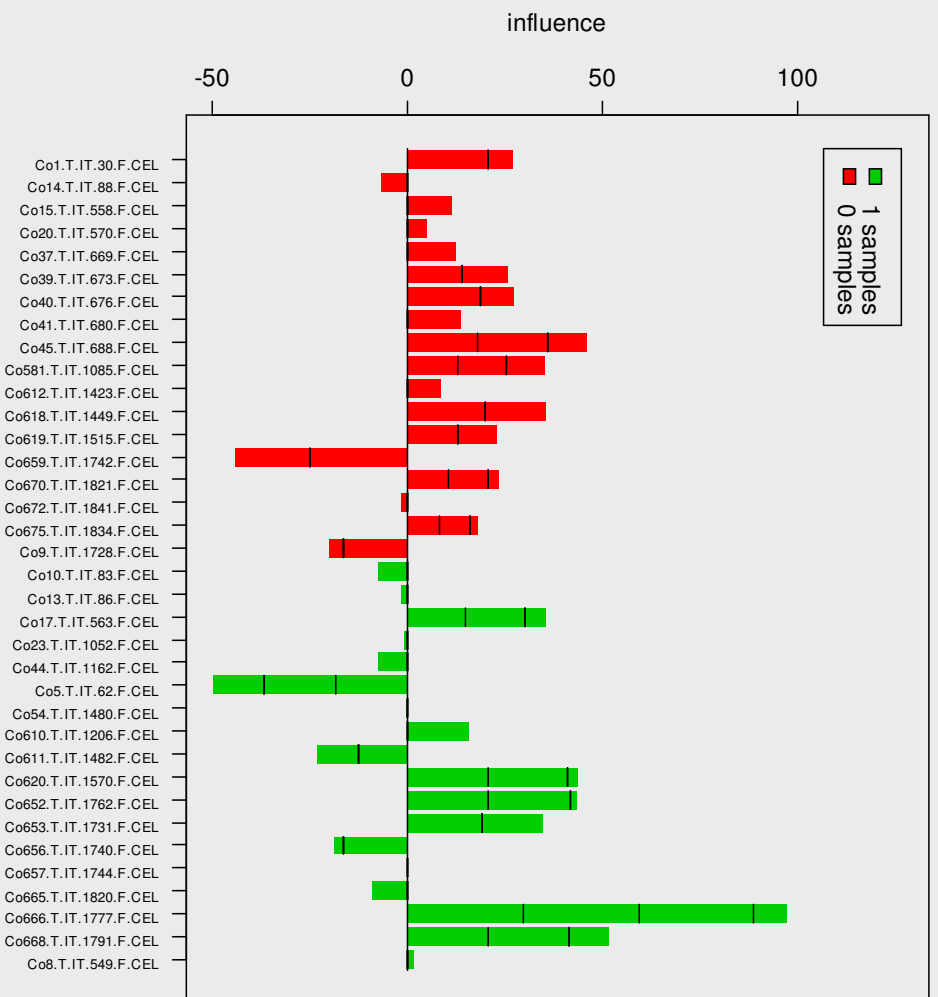
$$\sum [X_i'(Y - \mu)]^2$$



Values dichotomized around median

Group testing

# Goeman's Global Test – Example II





---

## Two questions about groups of genes

---

Question 1: Two groups of genes have to be compared with respect to gene expression: Is the gene expression in gene group A different from the expression in gene group B.

Genes of group A

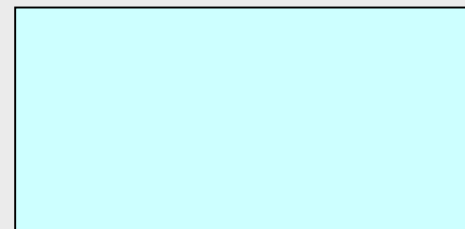
Genes of group B

Question 2: Is there differential gene expression between different biological entities not in terms of single genes but with respect to a defined group of genes.

Entity I

Entity II

Well defined  
group of genes



Group testing