Group testing: global tests, holistic approaches

Ulrich Mansmann, Manuela Hummel
IBE, Medical School, University of Munich

Content of the lecture

Biological relevant information may rather be encoded in groups and not predominantly in the expression of single 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
- **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
  - Relevance of specific gene group for biological phenomena

Gene groups

- **Pathways**
  networks of interacting genes (KEGG, cMAP, BioCarta)
- **Gene Ontology**
  biological process, molecular function, cellular component
- **Regions in the genome**
- **Signatures for classification**
- **Groups defined by literature search**
- **…**

Strategies for group testing

- **Differential gene expression:**
  - Dividing genes into two groups: differentially expressed yes/no is artificial
  - p-value correction methods don't really do what we want
  - Categories enter by *gene set enrichment* methods (e.g. Fisher-test approach, GSEA)
  - Identification of categories with many differentially expressed genes

- **Holistic approach:**
  - Define interesting categories
  - Find categories that include differentially expressed genes
  - Find categories of genes where there are potentially small but coordinated changes in gene expression

Fisher-test approach

In comparison with the whole population of genes, is the category of interest enriched with genes found to be differentially expressed?

	in category	not in category
differential		
not differential		

- Fisher-test
- Hypergeometric distribution

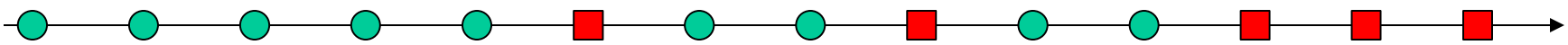
Gene set enrichment (I)

Problem:

Two groups of genes have to be compared with respect to gene expression *or* differential gene expression:
Is the gene expression in gene group A different from that in gene group B?
or: Is one of the groups enriched with differentially expressed genes?
Important: Genes in both groups are different!

Basic idea:

Order the genes with respect to the expression value *or* a measure for differential expression. If there is a difference between groups with respect to expression levels *or* differential expression, groups 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.



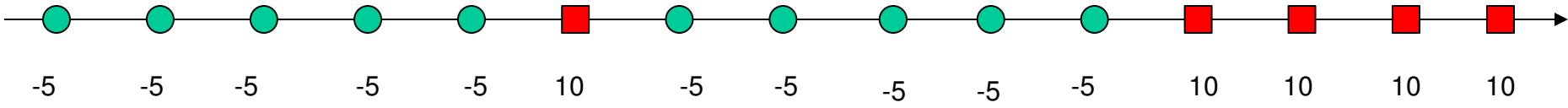
Group A

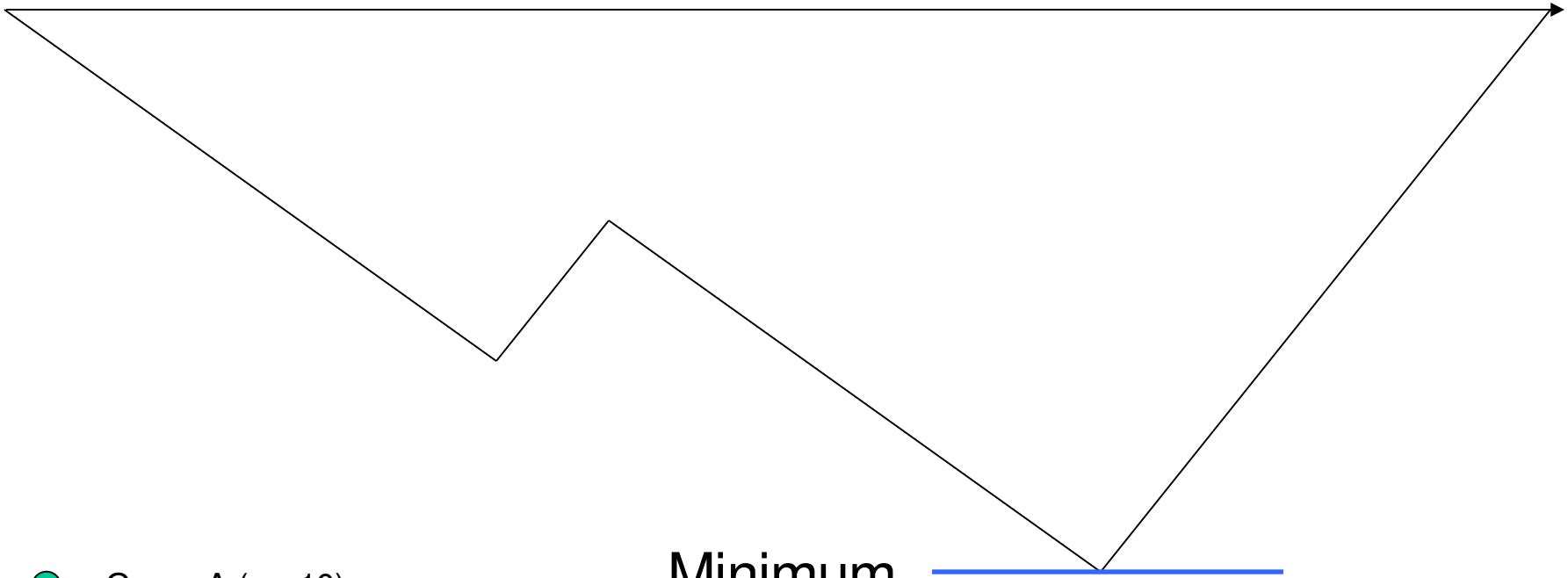
Group B

Rank of expression value *or* measure for differential expression

Gene set enrichment (II)

Genes ordered by rank of expression *or* by a measure for differential expression





Group A (n_A=10)

Group B (n_B=5)

Minimum

Is the minimum extreme with respect to random group mixing?

The algorithm formalized

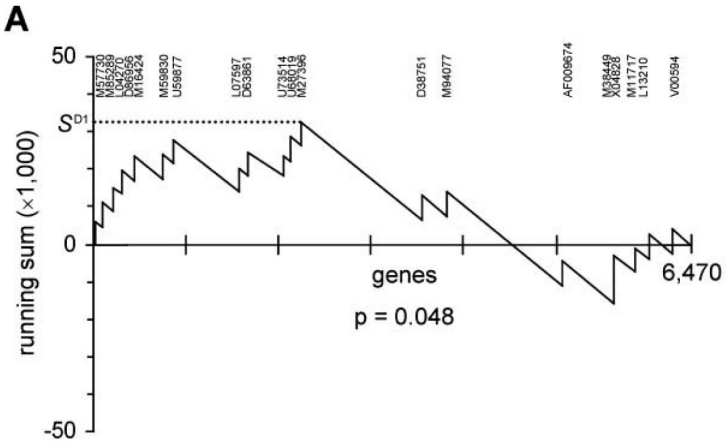
Basic idea:

- n_A genes in group A, n_B genes in group B.
- Order genes with respect to expression values *or* a measure for differential expression
- Create a score 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 the cumulative sum yy = 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 the most extreme value of the cumulative sum $M_{vv} = \max\{|\min(yy)|,\max(yy)\}$ which will be large in case of a good separation between both groups
- Permutation test:
  - 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: $p_{perm} = \#\{M_{vv^*} \geq M_{vv}\} / \#$ permutations.
  - *or:* Permute phenotype labels and compute vv*, yy* and M_{yy^*} for each permutation

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 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
- Group A: Cyclin D1 expression signature: cyclin D1 target gene set.

  Group B: all o



Example II: Colon Cancer

Study:
18 patients with UICC II colon cancer, 18 with UICC III colon cancer,
snap-frozen material, laser microdissection,
HG-U133A-arrays, 22.283 probesets representing ~18.000 genes

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

Analysis:
Use Gene Set Enrichment approach described before and rank genes according to a measure for differential expression
(absolute values of t-statistics)

Gene set enrichment – Colon cancer

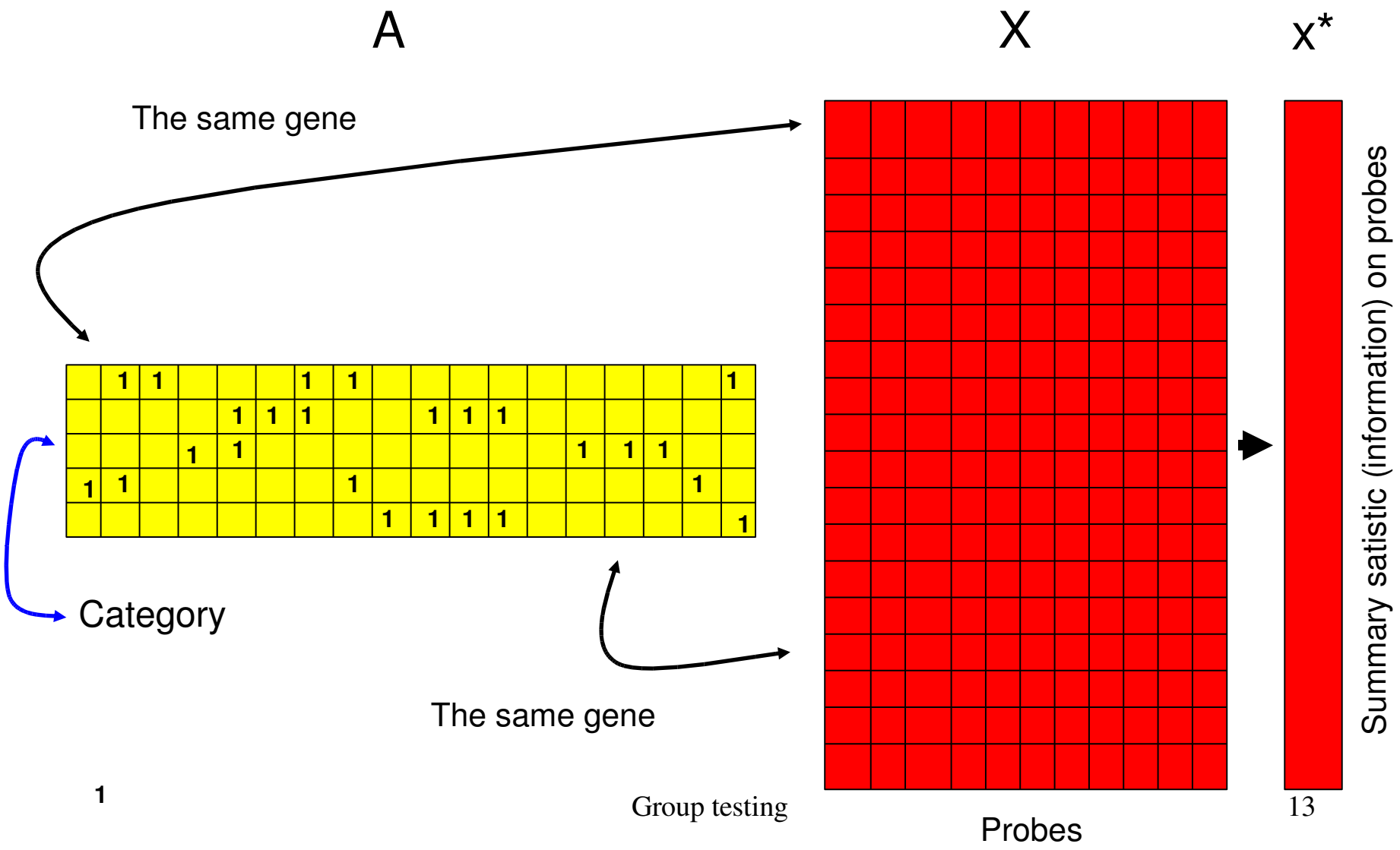
1407 probe sets are studied which belong to 9 cancer specific pathways

	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

Gentleman's categories (I)

- A set of categories is merely a grouping of genes (entities)
- The groups do not need to be exhaustive or disjoint
- The mapping from a set of entities (genes) to a set of categories can be represented as a bipartite graph:
  one set of nodes are the genes
  the other are the categories
- This mapping can be presented by an incidence matrix A (CxG)
  C: Number of categories
  G: Number of genes
- The elements of A: A[i,j] = 1 if gene j is in category I, else 0
- Row sums: Numbers of genes in each category
- Column sums: Number of categories a gene is in

Gentleman's categories (II)



Gentleman's categories (III)

- $z = A \bullet X$ or $z = A \bullet x^*$
- z is a vector of length C, represents *per category* sum,
  we are interested in large or small z's
- x^* could be the vector of gene-wise t-statistics between two groups,
  so we look for differential gene expression
- H_0: no difference between their means
- Components of x^* are approximately N(0,1)
- The elements of $z = A \bullet x^*$ are sums of N(0,1)
  [unfortunately not independent summands]
- Permutation test:
  - Permute the columns of A. This is the same as permuting the gene labels (the labels or rows of X and x^*)
  - *or:* Permute the sample labels (the columns of X) and recompute x^*

Gentleman's categories (IV)

- **Comparisons:**

  within category comparison: for a given category, is the observed test statistic unusual?

  overall comparison: are any of the observed category statistics unusually large or small with respect to the entire reference distribution?

- **Note:** The approach is inherently multivariate, one data set gives G test statistics and these are transformed to yield C z_i's

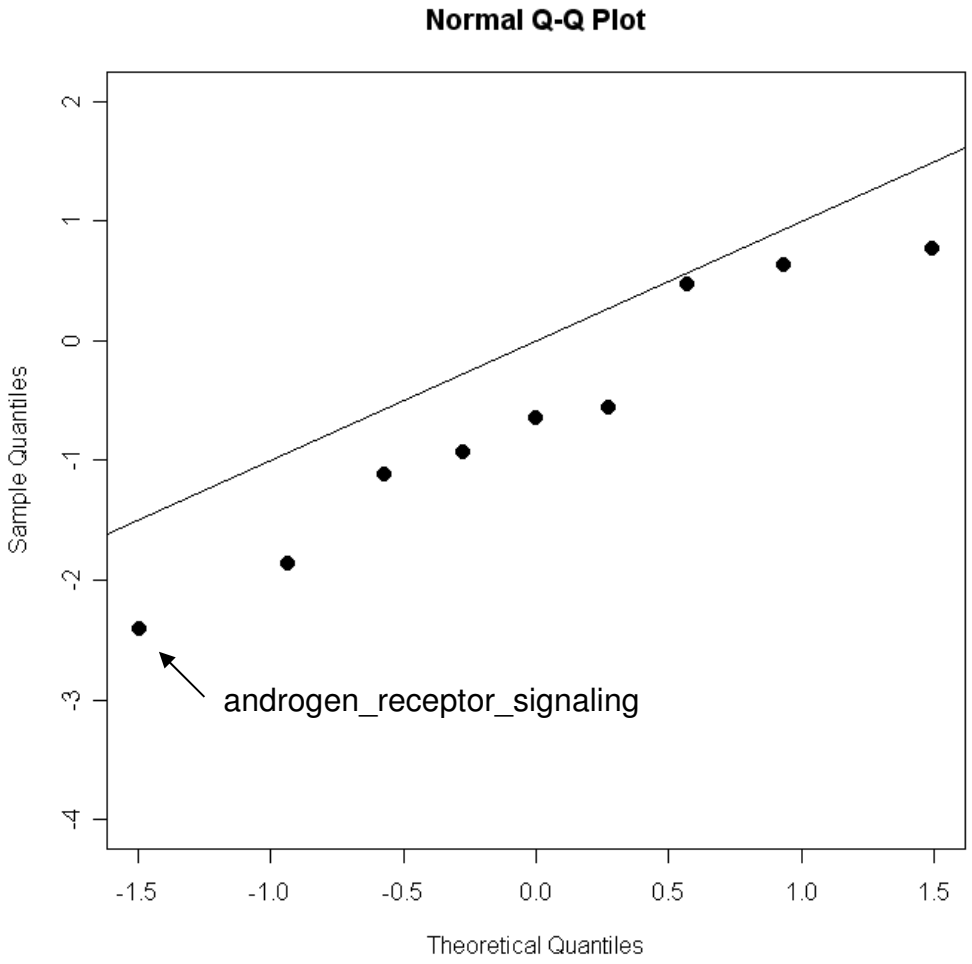
- The approach is well suited to fit the reasoning in a proper statistical framework

Gentleman's categories – Colon Cancer

Results for the colon data:

	z.statistic	p.Lower	p.Upper
androgen_receptor_signaling	-2.4124695	0.082	0.918
apoptosis	0.6270588	0.598	0.402
cell_cycle_control	0.7682091	0.690	0.310
notch_delta_signalling	-0.6442985	0.393	0.607
p53_signalling	-0.9325874	0.261	0.739
ras_signalling	0.4736045	0.586	0.414
tgf_beta_signaling	-1.1235767	0.376	0.624
tight_junction_signaling	-0.5652049	0.462	0.538
wnt_signaling	-1.8580599	0.288	0.712

Gentleman's categories – Colon Cancer



Goeman's global test

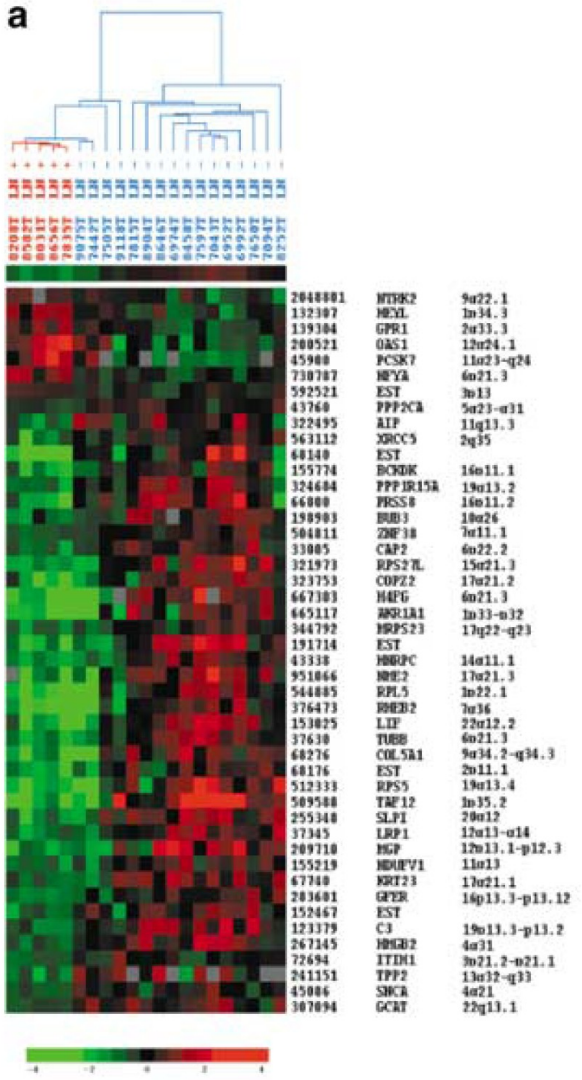
- Tests if global expression pattern of a group of genes is significantly related to some outcome of interest (groups, continuous phenotype)
- If this relationship exists, then the knowledge of gene expression helps to improve the prediction of the phenotype of interest.
  If the prediction can not be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

$$Q \sim (Y-\mu)'R\,(Y-\mu)$$
$$\sim \Sigma\,[X_i'(Y-\mu)]^2 \qquad \text{sum over genes of the pathway}$$
$$\sim \Sigma\,\Sigma\, R_{ij}(Y_i-\mu)\,(Y_j-\mu) \qquad \text{sum over subjects}$$

μ: 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



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

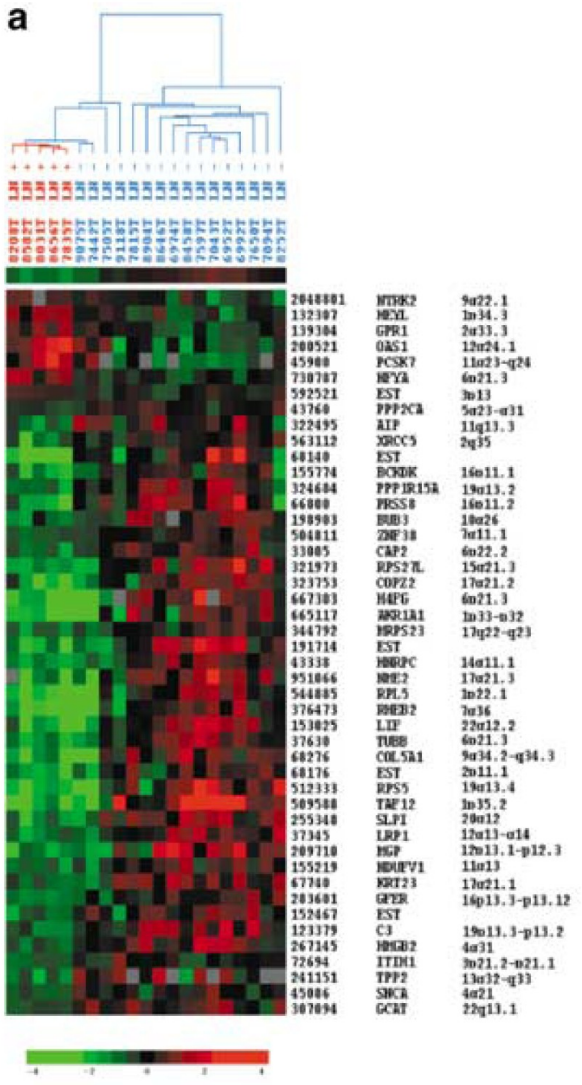
Bertucci et al. present a gene signature consisting of 46 genes which is claimed to be able to discriminate between LN- and LN+ colorectal cancer.

Is it possible to prove with new data that the signature has discriminative value? Can we reject the Null hypothesis

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

Y : LN+/LN-
X: expression pattern of 46 genes

Goeman's global test - Lymph node metastases



Global 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 IV: Colon Cancer

Study:
18 patients with UICC II colon cancer, 18 with UICC III colon cancer,
snap-frozen material, laser microdissection,
HG-U133A-arrays, 22.283 probesets representing ~18.000 genes

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

Analysis:
Goeman's global test

Goeman's global test – Colon Cancer

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

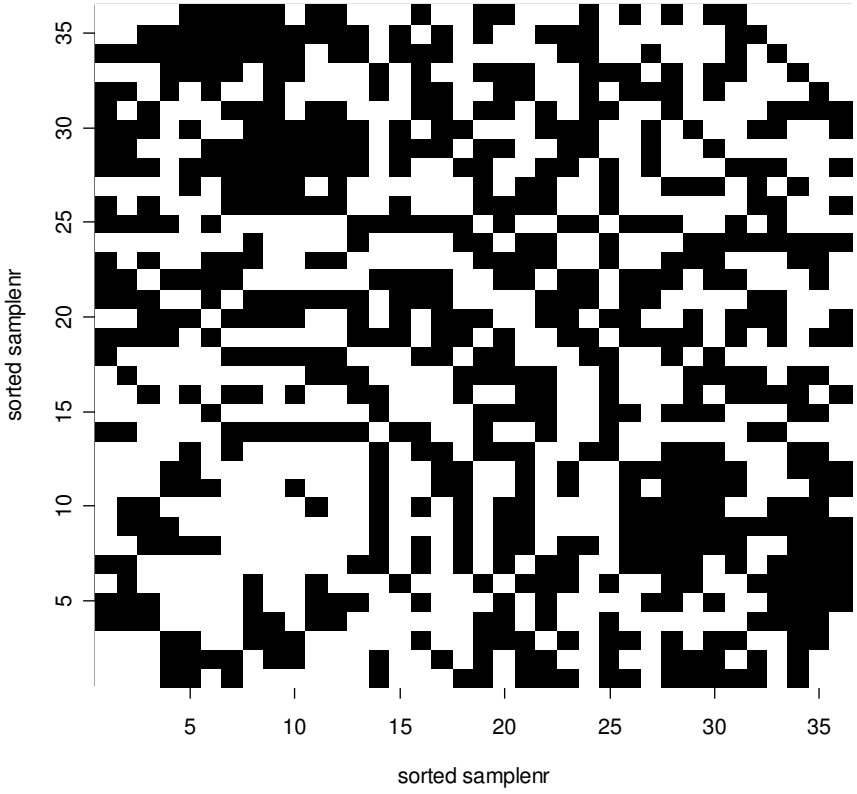
```
> gt <- globaltest(expressions,UICC.stage,genesets=pathways["p53_signalling"])
> gt

Global Test result:

Data: 36 samples with 1407 genes; 1 pathway tested
Model: logistic
Method: Asymptotic distribution

               Genes Tested Statistic Q Expected Q sd of Q  P-value
p53_signalling    45     45      20.446     9.1621  4.3335 0.021357
```

- 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 – Colon Cancer

```
> checkerboard(gt)
```

Simultaneous correlation of phenotype and expression

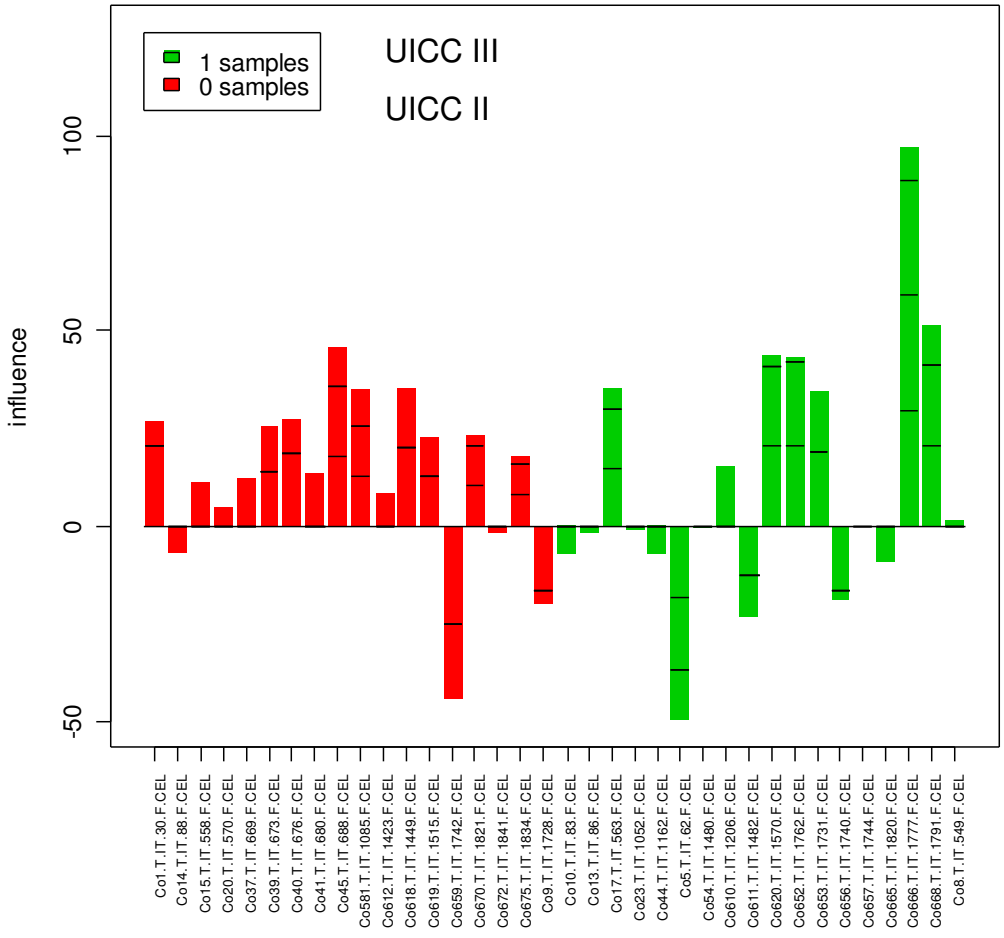


$$\Sigma \Sigma R_{ij}(Y_i-\mu)(Y_j-\mu)$$

Values dichotomized around median

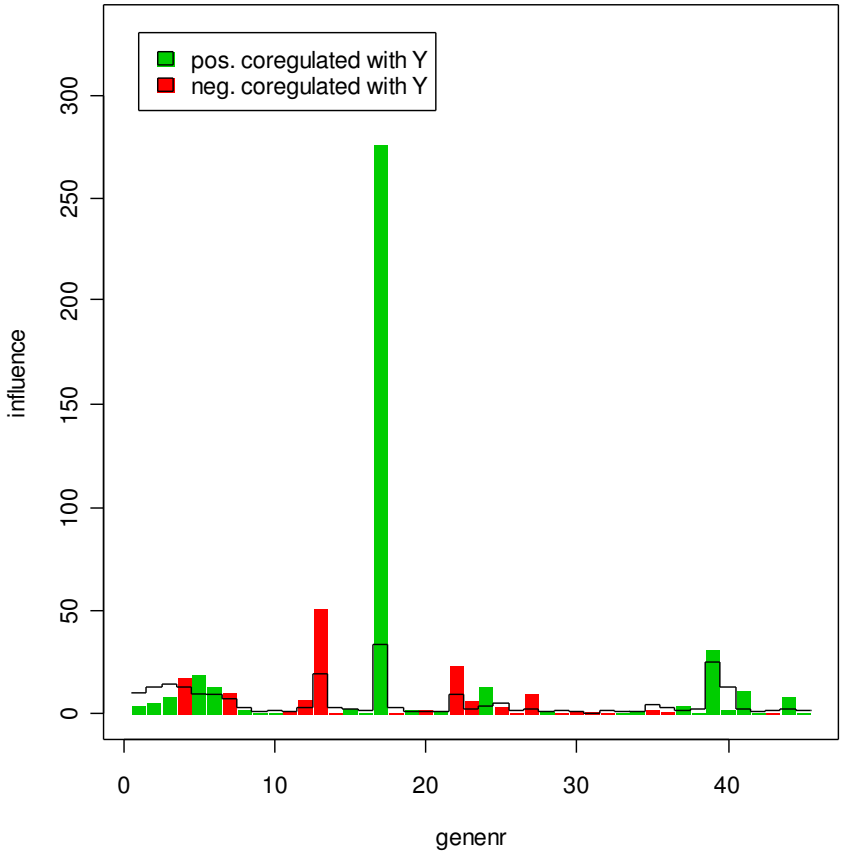
Goeman's Global Test – Colon Cancer

```
> sampleplot(gt)
```



Goeman's Global Test – Colon Cancer

```
> geneplot(gt)
```



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

GlobalANCOVA (I)

- P[X|Y] how is gene expression influenced by structure of Y?
- General framework: p genes, n probes, p>>n

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix} = M + E \qquad \text{E[E]=0} \qquad M = \begin{pmatrix} 11 & \cdots & 1n \\ \vdots & & \vdots \\ p1 & \cdots & pn \end{pmatrix} = \begin{pmatrix} (1) \\ \vdots \\ (p) \end{pmatrix}$$

- The expectation for gene i and n probes follows a linear model $(i)' = D\beta^{(i)}$
- Example for D^t in the two group case with two covariates

Probe	P1	P2	P3	P4	P5	P6	P7	P8
Gene (i) specific mean	1	1	1	1	1	1	1	1
Group	0	0	0	0	1	1	1	1
Sex	M	M	F	F	F	F	M	M
Localization	C	R	C	R	C	R	C	R

GlobalANCOVA (II)

Design	Full model	Reduced model
Groups	~ group + cov	~ cov
Dose – Response	~ dose + cov	~ cov
Group * Dose	~ group*dose + cov	~ group + dose + cov
time trends in groups	~ group*time + cov	~ group + time + cov
gene gene interaction	~ gene + cov	~ cov
differential co-expression	~ group*gene + cov	~ group + gene + cov
etc.		

GlobalANCOVA studies the following question: Is the observed gene expression sufficiently explained by the reduced model or does the full model improve the model fit substantially?

GlobalANCOVA fits gene-wise linear models but summarizes the fit of all single models to a global measure

Summary: Two perspectives on gene groups

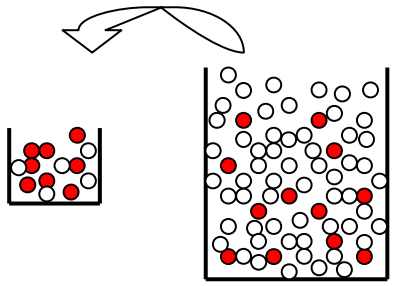
Question 1: Two groups of genes have to be compared with respect to gene expression: Is the gene expression in gene group A different from the expression in gene group B?

Genes of group A | Genes of group B

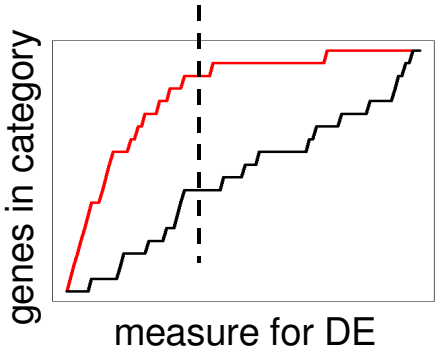
Question 2: Is there differential gene expression between different biological entities not in terms of single genes but with respect to a defined group of genes?

Summary: Perspectives of group testing methods

- **Fisher-test approach**
  Are there more interesting genes in the category the expected by randomly drawing?



- **Gene set enrichment**
  Do the genes in the category have high ranks with respect to differential expression?



- **Global test / GlobalANCOVA**
  Can there be found differential expression in the category?