Testing Groups of Genes

Manuela Hummel, LMU München
Adrian Alexa, MPI Saarbrücken
Ulrich Mansmann, LMU München

NGFN-Courses in Practical DNA Microarray Analysis
Regensburg, May 10, 2007

LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

IBE
INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG
BIOMETRIE UND EPIDEMIOLOGIE
DIREKTOR: PROF. DR. RER. NAT. ULRICH MANSMANN

NGFN
National Genome
Research Network

Overview

Analysis of groups of genes

- Motivation
- How to define gene groups
- Assess relevance of gene groups

Group testing methods

- Gene set enrichment: Fisher-test, GSEA
- Holistic approaches: Category, globaltest, GlobalAncova, restandardization approach

Motivation

So far: Gene-wise analysis

- Genes are treated independently
- Correction for multiple testing is crucial
- Resulting lists of interesting genes are rather ’instable’
- Biological interpretation of such gene lists is hard

Now: Analysis of gene sets

- Predefined gene groups provide more biological knowledge
- More meaningful interpretation in biological context
- Number of gene sets to be investigated is smaller than number of individual genes
- Useful for validation of published gene groups
  Example: Does a gene signature have predictive value?

How to Define Gene Groups

Exploratory research, literature search or Bioinformatic algorithms can be used to define

- Pathways
  Networks of interacting genes (KEGG, cMAP, BioCarta)
- Gene Ontology categories
  Biological Process, Molecular Function, Cellular Component
- Regions in the genome
- Signatures for classification
- Gene sets of published results
- . . .

Assess Relevance of Gene Groups

- Outstanding gene expression in a specific group compared to other genes
  Example: Do the cyclin D1 target genes show an extraordinary expression pattern in human tumours?

- Differential gene expression not of single genes but over a specific group of genes
  Example: Does the cell cycle pathway contain (many) differentially expressed genes between cancer types A and B?

- Two basic strategies for analysis:
  Gene set enrichment and holistic approaches

Group Testing

Gene set enrichment

- Idea: Provide biological meaning to a list of interesting genes by means of an over-representation analysis
- Step 1: Gene-wise analysis (e.g. of differential expression)
  Step 2: Score gene groups for enrichment
  (always in comparison with the set of all genes)
- Goal: Find gene groups that contain many interesting genes

Holistic approaches

- Idea: Look directly at gene sets and ask whether they are biologically relevant with respect to differential expression
- Global analysis of differential expression for gene groups (without taking the set of all genes as a reference)
- Goal: Find gene groups that contain at least one interesting gene or many genes with moderate differentiality

Hypergeometric Test

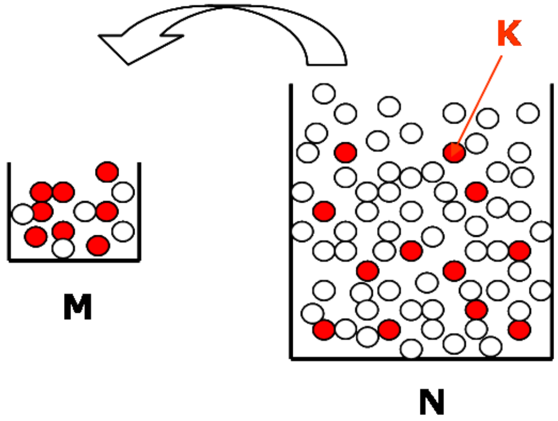
Step 1

- Compute a gene-wise measure (for differential expression, e.g. t-statistic p-values)
- Adjust for multiple testing and choose a cutoff to define a list of interesting genes

Step 2

- Given N genes on the microarray and M genes in a gene group, what is the probability of having x from K interesting genes in this group?

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$



- A p-value for the gene group corresponds to $P(X \geq x|N, M, K)$

Fisher's Exact Test

- The hypergeometric test is equivalent to Fisher's exact test

	$\in$ gene group	$\notin$ gene group	
$\in$ DE genes	x	$K - x$	K
$\notin$ DE genes	$M - x$	$(N - M) - (K - x)$	$N - K$
	M	$N - M$	N

- Fisher-test and similar tests based on gene counts are very often used in Gene Ontology analysis
  (binomial test, χ^2 test, test based on normal z scores)
  Khatri and Draghici (2005)

- All these tests have the hypergeometric as null distribution
  Rivals et al. (2006)

Fisher's Exact Test

Example: $N = 20000$ genes on the microarray, $M = 100$ genes in a gene group of interest, $K = 300$ differentially expressed genes

	$\in$ group	$\notin$ group	
$\in$ DE	3	297	300
$\notin$ DE	97	19603	19700
	100	19900	20000

could be random

p-value $=$ 0.19

	$\in$ group	$\notin$ group	
$\in$ DE	6	294	300
$\notin$ DE	94	19606	19700
	100	19900	20000

not likely random

p-value $=$ 0.004

Fisher's Exact Test

Advantages

- Not restricted to analysis of differential expression
- If you just get a list of somehow interesting genes and want to assess biological background, tests based on gene counts are the only way to go

Problems

- Loss of information because of two separated steps
- Small but consistent differential expression is not detected
- Dividing genes into differentially and non-differentially expressed genes is artificial
- No clear way of defining K: p-value correction and choice of a cutoff are crucial

Gene Set Enrichment Analysis

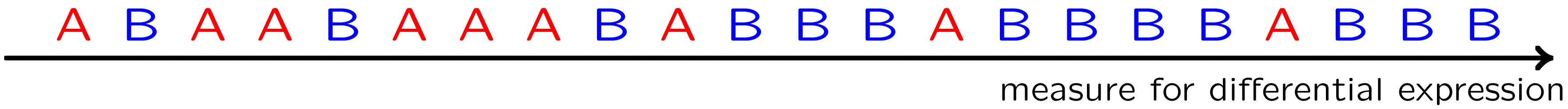
Subramanian et al. (2005)

Step 1

- Compute a gene-wise measure (for differential expression, e.g. absolute t-statistics)
- Rank genes according to this measure

Step 2

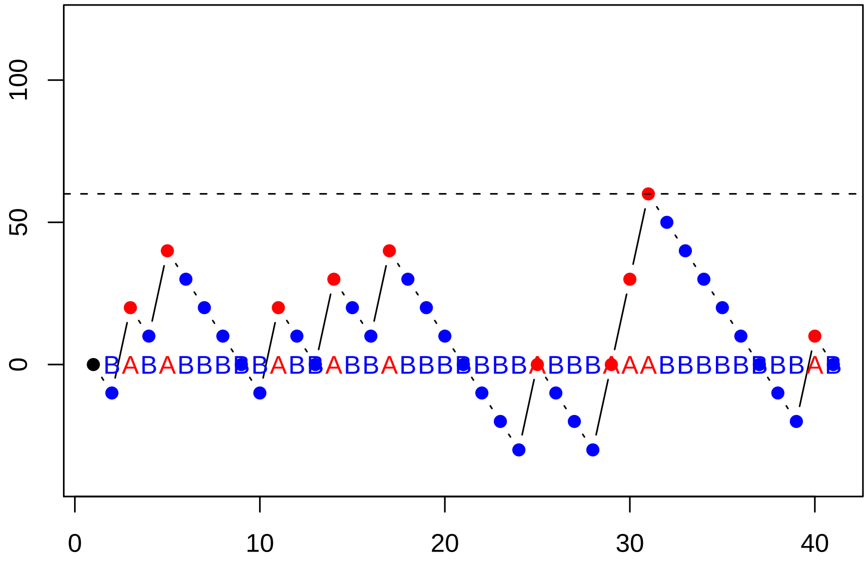
- Assign labels A to genes belonging to a gene group of interest and B to all the other genes
- If group A is enriched with interesting genes, many of it's genes will have high ranks and we will observe a separation in the ordered list

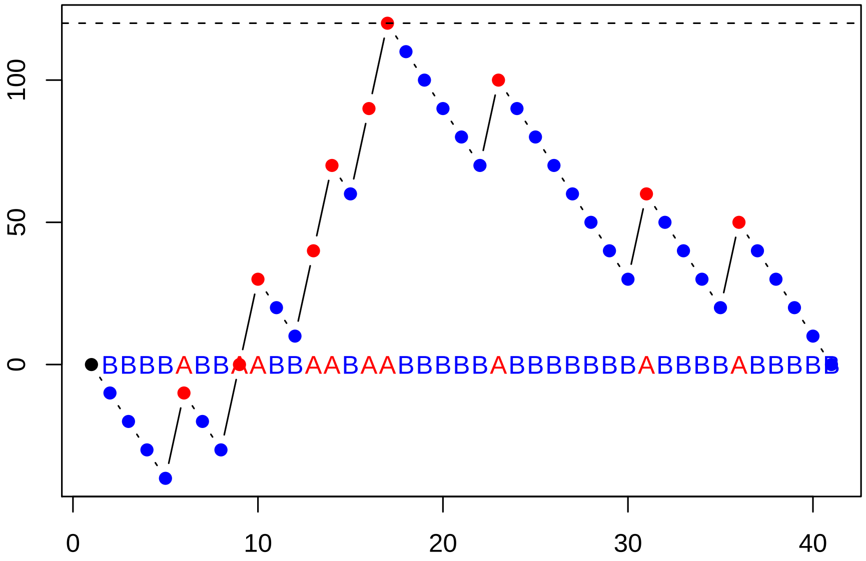

A B A A B A A A B A B B B A B B B B A B B B

measure for differential expression

Gene Set Enrichment Analysis

- Assign score n_B to all genes A and $-n_A$ to all genes B
- Draw the cumulative sum of these scores
- Is the maximum M of the cumulative sum unusually high? (Kolmogorov-Smirnov test)





GSEA Permutation Test

Permute genes

- Permute labels A and B in the ordered list P times
- Calculate the maximum M^* of the cumulative sum for each permutation
- Empirical p-value: $p = \#(M^* \geq M)/P$
- Hypothesis: group is extreme w.r.t. random mixing

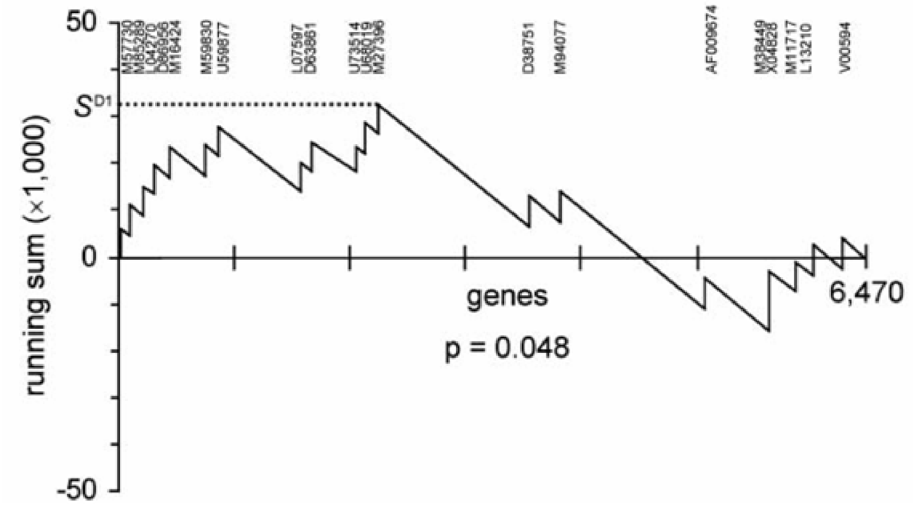
Permute subjects

- Permute phenotype labels in the expression matrix
- Compute the gene-wise measure for each permutation
- For each resulting gene ranking calculate M^* and then a p-value as above
- Hypothesis: group is extreme w.r.t. overall expression

GSEA Example

- *Lamb et al. (2003)* investigate activity of cyclin D1 in human tumours: Does the cyclin D1 target gene set play a prominent role in different tumour entities? Being present as highly expressed genes

- Group A: cyclin D1 target gene set
  Group B: all other genes



Gene Set Enrichment Analysis

Advantages

- Not restricted to analysis of differential expression
- Ranking of genes is considered
- No cutoff has to be chosen

Problems

- Loss of information because of two separated steps
- Small but consistent differential expression is not detected

Category

Gentleman (2006)

- Goal is to find gene categories whose genes show small but consistent expression changes in the same direction

- Calculate vector $\mathbf{x}$ of genewise statistics indicating differential expression, e.g. t-test statistics or more general $\mathbf{x} = f_1(\mathbf{X})$

- Get an incidence matrix $\mathbf{A}$ representing the mappings between predefined categories and genes

$$\mathbf{A} = \underset{\begin{array}{c}\uparrow\\ \text{genes}\end{array}}{\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 1 & 0 & \ldots \\ 1 & 1 & 0 & 1 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}} \longleftarrow \text{categories}$$

- Row sums: numbers of genes in each category
  Column sums: numbers of categories each gene belongs to

Category

- Define a statistic $\mathbf{z}$ that reflects which categories are extreme:

$$\mathbf{z} = \frac{\mathbf{Ax}}{\sqrt{rowsums(\mathbf{A})}} \quad \text{or more general} \quad \mathbf{z} = f_2(\mathbf{A}, \mathbf{x})$$

- When $\mathbf{x}$ is a vector of t-statistics and $\mathbf{z}$ as shown, then $\mathbf{z} \sim N(0, 1)$ (unfortunately only when genes are independent)

- Comparisons are possible
  Within categories: For a given category, is the observed test statistic unusual?
  Between categories: Are any of the observed category statistics unusually w.r.t. the entire reference distribution?

Category Permutation Test

Permute genes

- Permute columns of $\mathbf{A}$ P times
- Calculate category statistic $\mathbf{z}^*$ for each permutation
- Empirical p-value: $p = \#(\mathbf{z}^* \geq \mathbf{z})/P$
- Hypothesis: group is extreme w.r.t. random mixing

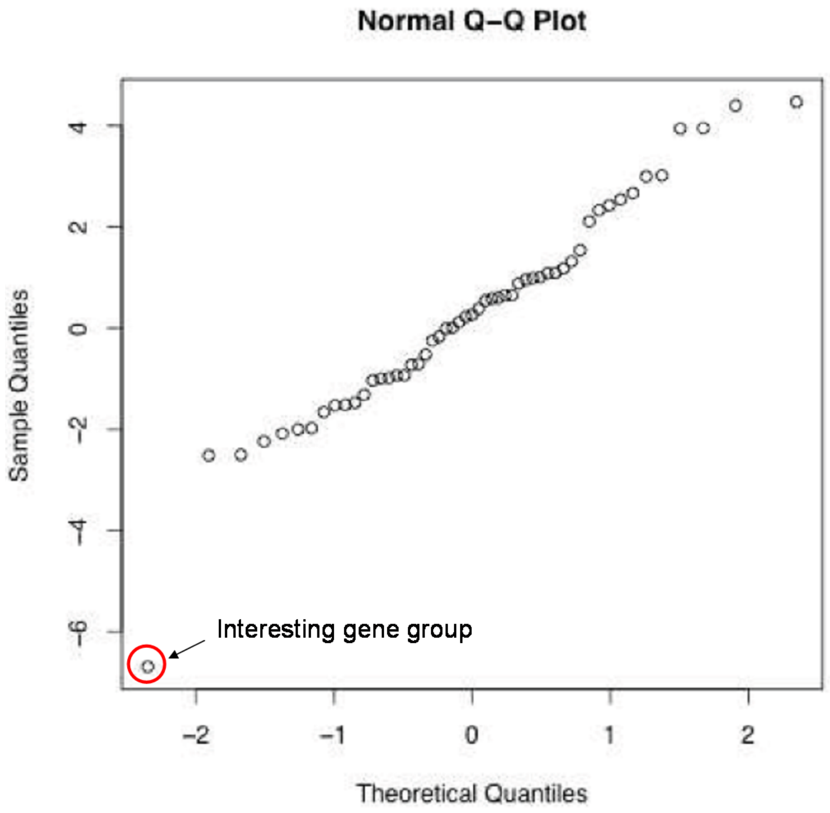
Permute subjects

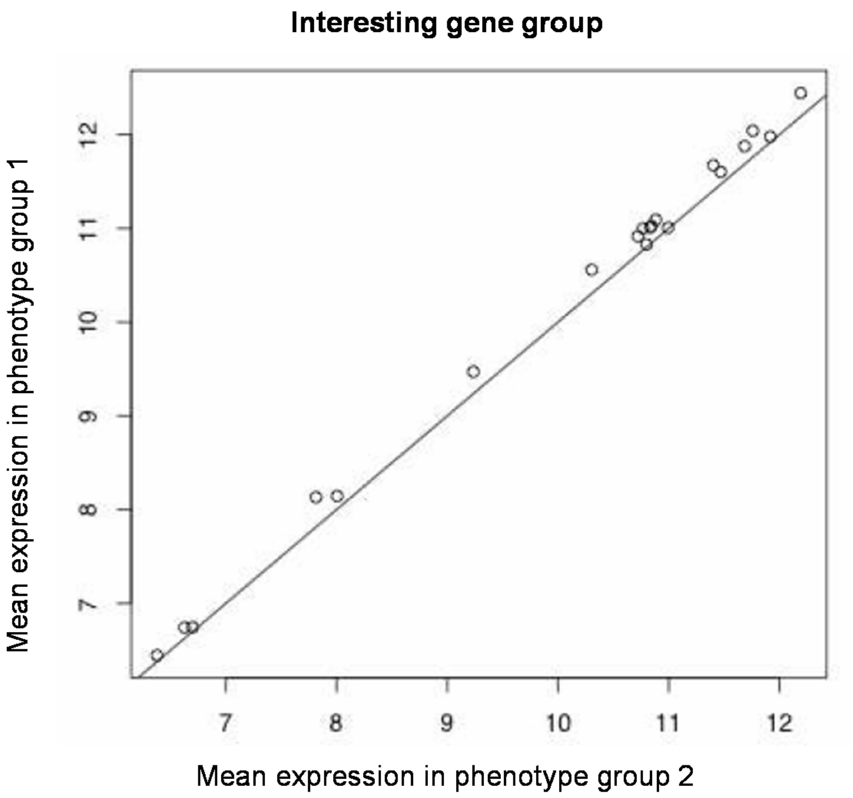
- Permute phenotype labels in the expression matrix $\mathbf{X}$
- Compute the gene-wise measure $\mathbf{x}^*$ for each permutation
- Calculate category statistic $\mathbf{z}^*$ with $\mathbf{A}$ and each $\mathbf{x}^*$ and then a p-value as above
- Hypothesis: group is extreme w.r.t. overall expression

Category

- qq-plots of the category statistics can help to reveal interesting gene groups

- These groups can further be explored by plotting expression means in the two clinical entities against each other





Category

Advantages

- Proper statistical framework
- Very flexible through choice of functions f_1 and f_2
- Ability to find groups with interesting expression patterns missed by gene set enrichment approaches

Problems

- Categories with both up- *and* down-regulated genes will eventually not be found because their t-statistics will cancel out in the overall sum
- Permutation of genes destroys correlations between genes, permutation of subjects ignores overall distribution of group statistics – what to do?

Global Tests

Is the global expression pattern of a group of genes significantly related to some clinical variable of interest?

globaltest: Does knowledge of gene expression X help to improve prediction of the variable Y?
$H_0 : P(Y = 1|X) = P(Y = 0|X)$
Goeman et al. (2004)

GlobalAncova: How is gene expression X influenced by the structure of the variable Y?
$H_0 : P(X|Y = 1) = P(X|Y = 0)$
Mansmann and Meister (2005)

Tests are equivalent under the null hypothesis of no relationship between Y and X

Globaltest

- Does knowledge of gene expression X help to improve prediction of the variable Y?

- Test statistic

$$\begin{aligned} Q &\sim (Y-\mu)^T R (Y-\mu) & \\ &\sim \sum_g [X_g(Y-\mu)]^2 & \text{sum over genes} \\ &\sim \sum_i \sum_j R_{ij}(Y_i-\mu)(Y_j-\mu) & \text{sum over subjects} \end{aligned}$$

  $R = X^T X$ matrix of correlations between gene expression of subjects

- Test to see whether subjects with similar expression also have similar outcomes

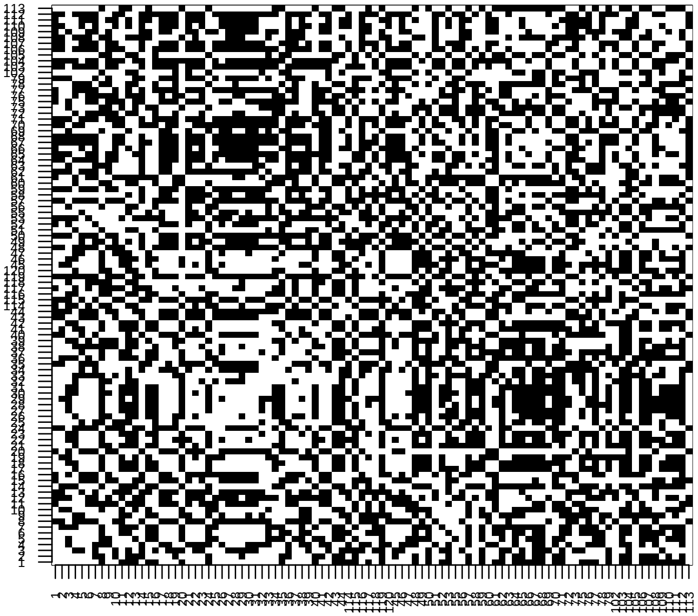
- Permutation based and asymptotic p-values are available

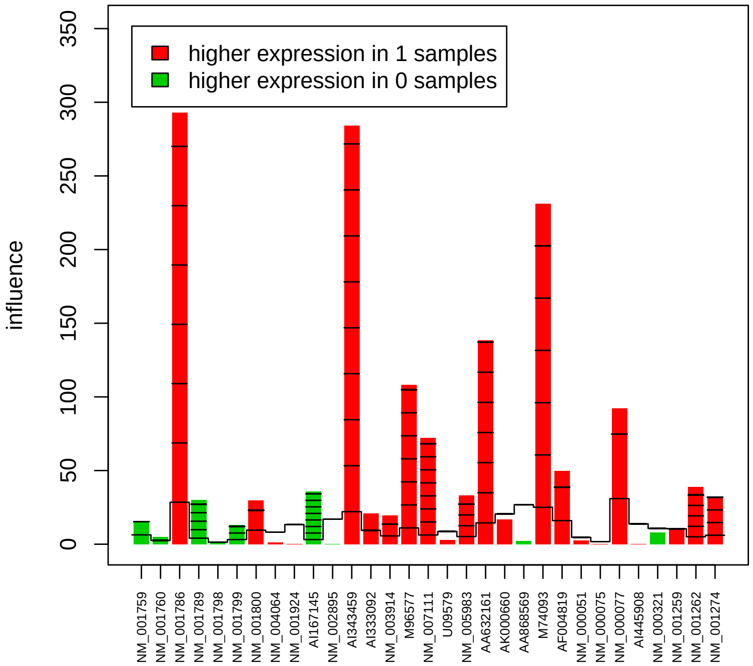
- Also multicategorical, continuous or survival variables can be considered and adjustment for covariates is possible

Globaltest

- Checkerboard plots help to illustrate whether subjects of the same clinical group also have similar expression patterns

- Gene plots show the influence of single genes in the gene sets on the global test statistic





GlobalAncova

- How is gene expression X influenced by the structure of the variable Y?

- The expectation for gene j follows a linear model $E(x_j) = D\beta_j$

- The design matrix D, e.g. in the two group case and with an additional covariate z, may look like this

$$\begin{array}{lc} & \begin{array}{ccc}\text{Int} & Y & z\end{array} \\ \begin{array}{l}\text{sample 1}\\ \text{sample 2}\\ \text{sample 3}\\ \text{sample 4}\\ \quad\ldots\end{array} & \begin{pmatrix}1 & 0 & 0\\ 1 & 0 & 1\\ 1 & 1 & 1\\ 1 & 1 & 0\\ & \ldots & \end{pmatrix}\end{array}$$

- The full model containing the clinical parameter of interest is compared to a reduced model without it via the extra sum of squares principle

- Gene-wise linear models are summarized to a global F-test

GlobalAncova

- Permutation p-values:
  Permutation of subjects and calculation of empirical p-values
  Asymptotic p-values:
  Approximation of the test statistic distribution

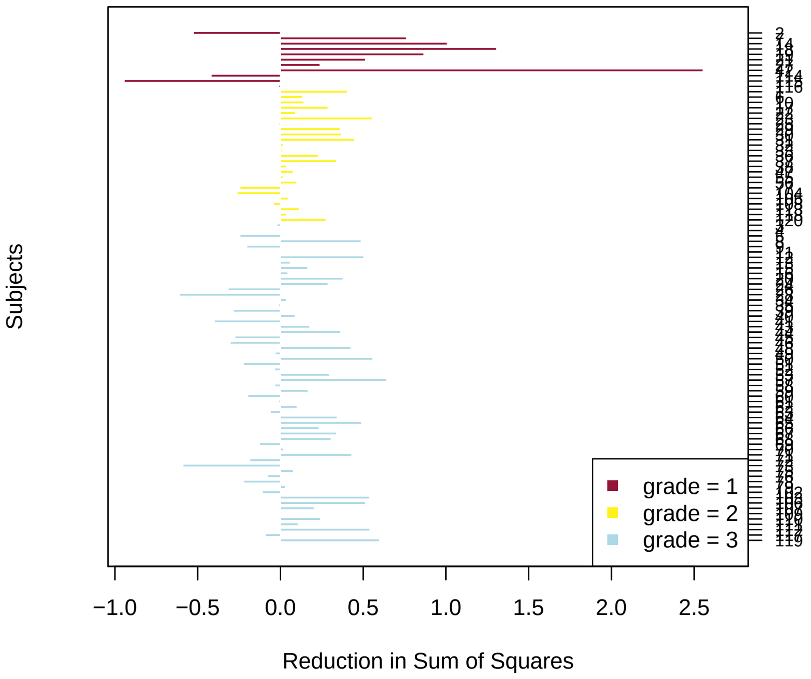
- General linear model framework allows analysis of

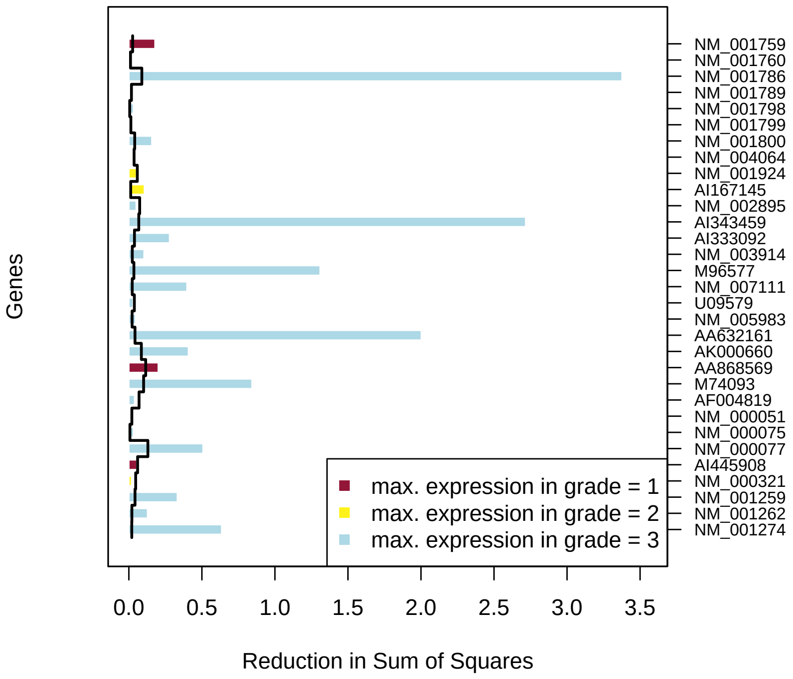
Design	Full model	Reduced model
Various groups	∼ `group + cov`	∼ `cov`
Dose-response	∼ `dose + cov`	∼ `cov`
Group by dose interaction	∼ `group * dose + cov`	∼ `group + dose + cov`
Differential time trends	∼ `group * time + cov`	∼ `group + time + cov`
Gene gene interaction	∼ `gene + cov`	∼ `cov`
Differential co-expression	∼ `group * gene + cov`	∼ `group + gene + cov`
...		

GlobalAncova

- Subject plots help to detect subjects that ’do not fit’ into their clinical groups

- Gene plots show the influence of single genes in the gene sets on the global test statistic





Global Tests

Advantages

- Gene groups with few strongly as well as groups with many moderately differentially expressed genes are detected
- Flexible frameworks suitable for many kinds of applications

Problems

- Only analysis of expression patterns within groups – it is not accounted for the overall distribution of group statistics
- Eventually too sensitive for data with much differential expression

Restandardization

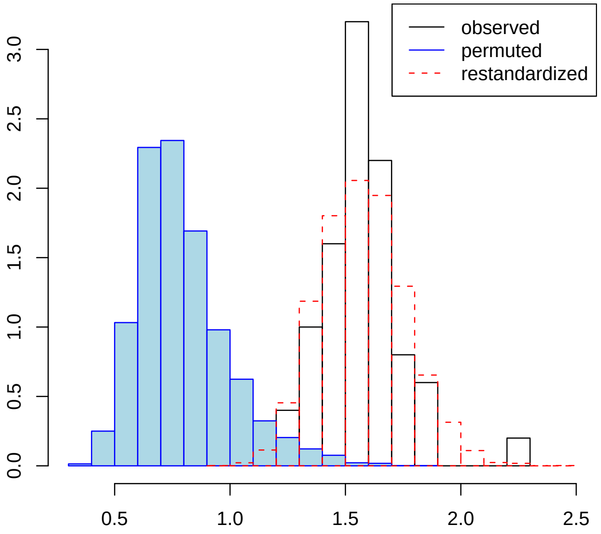
Efron and Tibshirani (2007)

- How can a proper null distribution for some gene set statistic S be simulated?
- Randomization of genes destroys correlations between genes: variability of S will be underestimated
- Permutation of subjects does not account for the overall distribution: If *all* genes are equally differential, all gene groups will look significant though none of them is more extreme than the others
- Restandardized gene set statistic

$$S^{**} = \mu^{+} + \frac{\sigma^{+}}{\sigma^{*}}(S^{*} - \mu^{*})$$

μ^{+}, σ^{+} mean and standard deviation of S^{+} for a randomly selected gene set of same size

μ^{*}, σ^{*} corresponding quantities for S^{*}, which are computed based on sample permutations



Restandardization

Advantages

- Applicable for arbitrary gene set statistics
- Combines ideas of a global group statistic and at the same time comparison with all remaining genes

Problems

- For complex group statistics a nested simulation is required
- Is it really necessary to account for the overall distribution of gene set statistics?
- Gene randomization is problematic

Gene versus Subject Sampling

Goeman and Bühlmann (2007)

Subject sampling model: A new sample corresponds to measurements of the same variables (= genes) for a new subject

Gene sampling model: A new sample would correspond to a sample of new genes for the same subjects
(this is also the underlying model for hypergeometric tests)

- Gene sampling reverses the roles of samples and variables
- Interpretation of p-values is different
- Misleading sample size in gene sampling model, i.e. the number of genes m does not correspond to the biological sample size $n =$ number of subjects
- Assumption of independence between genes in the gene sampling model may lead to anti-conservative tests

Summary: Two Perspectives on Gene Groups

Question 1

Is the gene expression in gene set A different from the expression in gene set B?

Gene set A

Gene set B

Question 2

Is there differential expression between different biological entities, not in terms of single genes but with respect to a defined gene set?

Entity 1

Gene set X

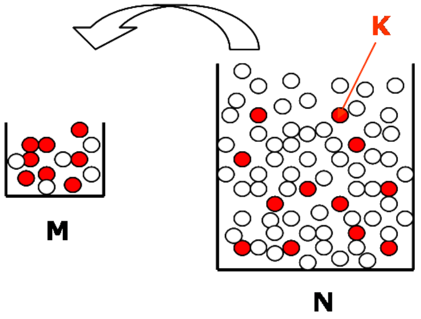
Entity 2

Gene set X

Summary: Perspectives of Group Testing

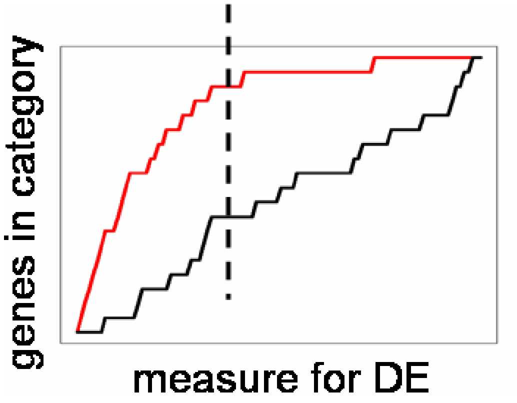
Fisher-test approaches

Are there more interesting genes in the gene set than expected by randomly drawing?



Gene set enrichment analysis

Do the genes in the gene set have high ranks with respect to differential expression?



Globaltest / GlobalAncova / Category

Can there be found differential expression in the gene set?

Outlook

- Gene versus subject sampling: Also tests based on gene counts in a contingency table could be modified to subject sampling procedures

- Annotation: Only genes annotated to the considered gene sets are involved in the analysis, all others are missed

- When testing large collections of gene sets we have to face a multiple testing problem

- Dependencies between gene sets complicate statistical analysis and interpretation
  Special example: Gene Ontology

References

1. Efron B, Tibshirani R. On testing the significance of sets of genes. Annals of Applied Statistics 2007, to appear.

2. Gentleman R with contributions from Falcon S. Category: Category Analysis. R package version 2.0.0.

3. Goeman JJ, de Kort F, van de Geer SA, van Houwelingen JC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004; 20(1): 93-99.

4. Goeman JJ, Bühlmann P. Methodological issues in gene set testing based on microarray data. Submitted.

5. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005.

6. Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, Kittrell FS, Zahnow CA, Patterson N, Golub TR, Ewen ME. A mechanism of Cyclin D1 Action Encoded in the Patterns of Gene Expression in Human Cancer. Cell 2003; 114: 323-334.

7. Mansmann U, Meister R. Testing differential gene expression in functional groups. Methods Inf Med 2005; 44(3).

8. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2006.

9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 2005; 102(43): 15545-15550.