Ontologic Analysis: Challenges for Global Testing

Manuela Hummel, LMU München
Ulrich Mansmann, LMU München
Reinhard Meister, TFH Berlin

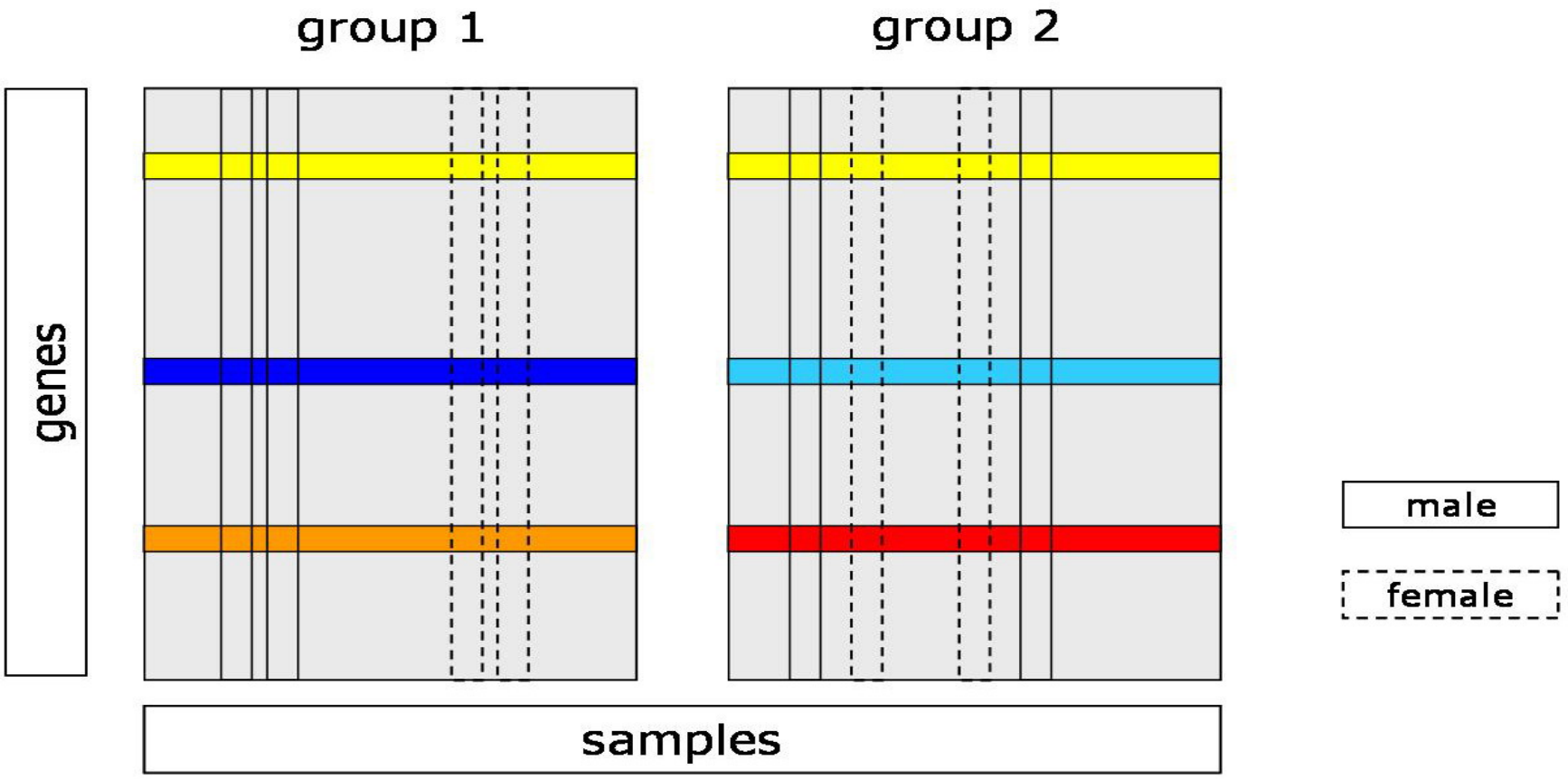
GMDS 2006, Leipzig

IBE
Institut für medizinische Informationsverarbeitung,
Biometrie und Epidemiologie

Overview

- Global tests for groups of genes
- Example of flexible linear modeling with `GlobalAncova`
- Gene Ontology analysis
- How to find significant regions in the GO?

Differential Gene Expression



Differential Gene Expression

Question A:

Which genes differ in expression between biological entities?

→ Single tests for each gene

Question B:

Do functional groups of genes (e.g. pathways, areas in the genome, Gene Ontology terms) contain genes showing differential expression?

→ Global tests for groups of genes

Global Testing

Y : clinical outcome, $X : p \times n$ gene expression matrix
(p genes, n samples)

Approach A: $H_0 : P(Y = 1|X) = P(Y = 2|X)$

Random Coefficient Generalized Linear Model; Score test

Goeman et al. (2004)

R package `globaltest`

Approach B: $H_0 : P(X|Y = 1) = P(X|Y = 2)$

ANCOVA: Comparison of adjusted means; F test

Mansmann, Meister (2005)

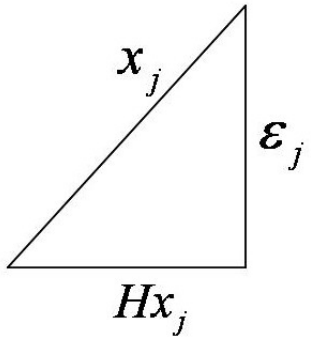
R package `GlobalAncova`

Hypotheses are equivalent by Bayes theorem

GlobalAncova

- Question of interest:
  How is gene expression X influenced by phenotype Y?
- The expectation for gene j follows a linear model
  $E(x_j) = D\beta_j = Hx_j$, with $H = D(D'D)^{-1}D'$
- The design matrix D, e.g. in the two group case and with an additional covariate z, e.g. sex, may look like this

$$\begin{array}{c} \\ \text{sample 1} \\ \text{sample 2} \\ \text{sample 3} \\ \text{sample 4} \\ \dots \end{array} \begin{array}{c} \begin{array}{ccc} \text{Int} & Y & z \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ & \dots & \end{pmatrix} \end{array}$$



- Residual sum of squares for gene j
  $\varepsilon_j{}'\varepsilon_j$, with $\varepsilon_j = (I - H)x_j$

GlobalAncova

- Question of interest:
  Do we need the variable Y to explain the data X?
  $\rightarrow$ Extra sum of squares principle

- Design matrices:
  $D_{full} = (\mathbf{1}, Y, z)$, $D_{reduced} = (\mathbf{1}, z)$

- Extra residual sum of squares:
  $SSR_{extra} = SSR_{reduced} - SSR_{full}$, with $SSR = \sum_{j=1}^{p} \varepsilon_j{}' \varepsilon_j$

- F statistic:
  $F = MSR_{extra}/MSR_{full}$

GlobalAncova p-values

Theoretical F distribution p-values

Those are not valid in the case of correlations between genes or non-normality

Permutation p-values

Sample labels are permuted and a p-value is estimated as the fraction of corresponding permutation F statistics that are greater than the observed F statistic

Asymptotic distribution of the test statistic

The test statistic has an asymptotic scaled F distribution $\sim b \cdot F(h_1, h_2)$ where b, h_1 and h_2 depend on eigenvalues of the $p \times p$ gene expression covariance matrix and adequate differences of $n \times n$ model hat matrices

Linear Models

The global ANCOVA approach can easily be extended to a general linear model framework with various modeling capabilities

design	full model	reduced model
Various groups	~ `group + cov`	~ `cov`
Continuous variable	~ `dose + cov`	~ `cov`
Time trends in groups	~ `group * time + cov`	~ `group + time + cov`
Gene-gene interaction	~ `gene + cov`	~ `cov`
Co-expression	~ `group + gene + cov`	~ `group + cov`
Differential co-expression	~ `group * gene + cov`	~ `group + gene + cov`
...	...	...

Example

- *Van t'Veer et al. (2002)* present a gene signature of 70 genes to predict recurrence of breast cancer
- We derived 9 cancer related pathways from a literature research
- Questions:
  Is it possible to relate the signature genes to the pathways?
  Are signature genes co-expressed with pathways?
- Explored clinical outcome: development of distant metastases within 5 years (yes/no)
- For demonstration we pick the cell cycle pathway and signature gene „AL137718“

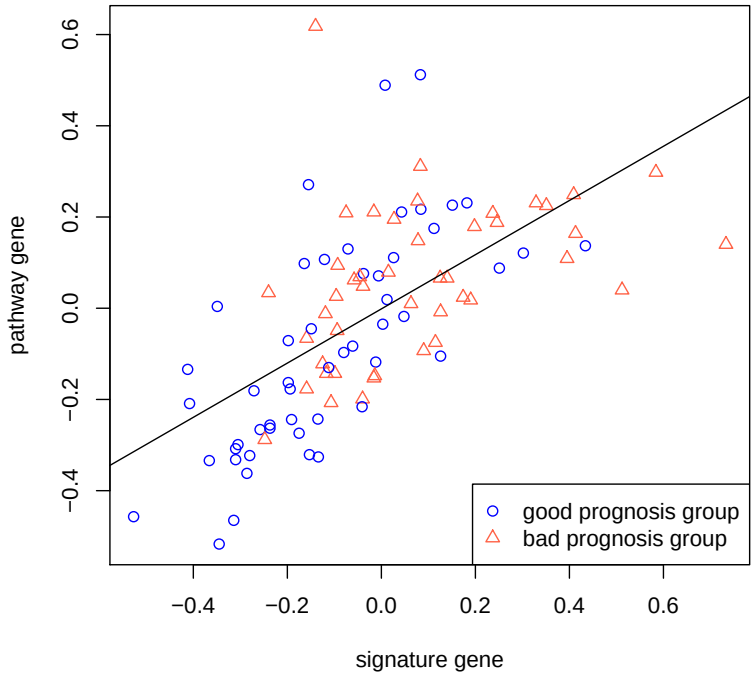
Example

Is there a correlation between the expression of the signature gene and the pathway genes?

Full model: $\sim$ `signature.gene`

Reduced model: $\sim$ `1`

```
$ANOVA
              SSQ    DF         MS
Effect  27.24455    31 0.8788564
Error  117.75240 2914 0.0404092

$test.result

F.value 21.74892
p.value  0.00000
p.perm   0.00000

$terms
[1] "(Intercept)" "signature.gene"
```



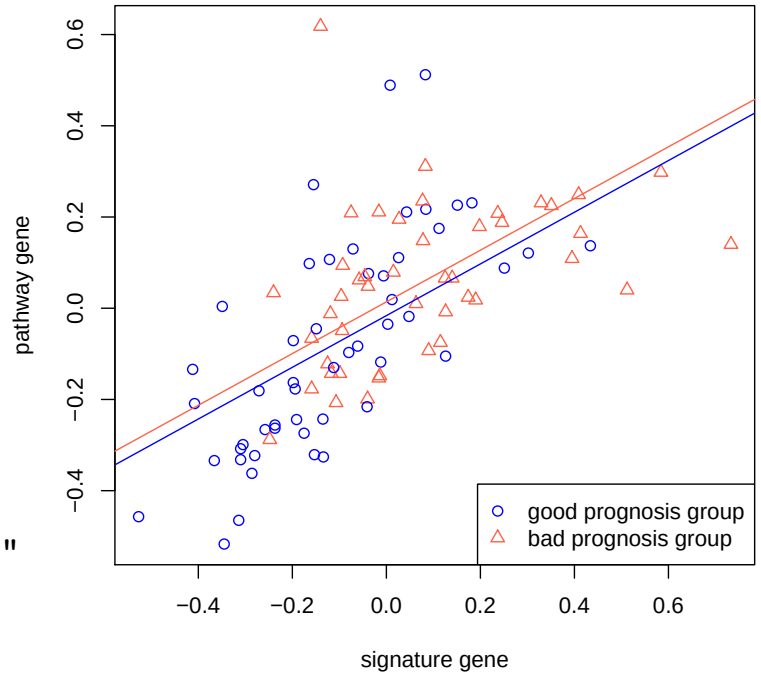
Example

Is there co-expression between signature gene and pathway regarding the clinical outcome?

Full model: $\sim$ `metastases + signature.gene`

Reduced model: $\sim$ `metastases`

```
$ANOVA
              SSQ    DF          MS
Effect  21.27201    31 0.68619383
Error  116.90453 2883 0.04054961

$test.result

F.value 16.92233
p.value  0.00000
p.perm   0.00000

$terms
[1] "(Intercept)" "metastases"  "signature.gene"
```



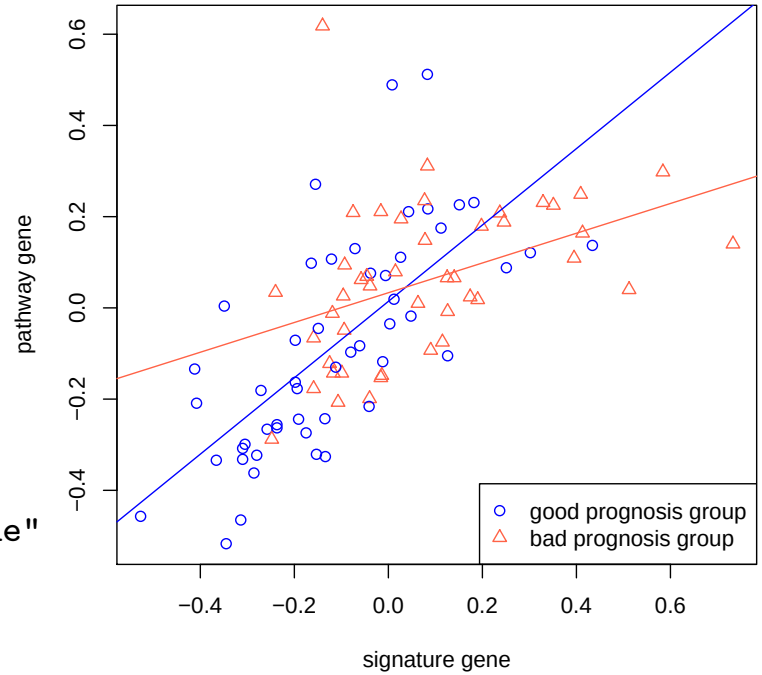
Example

Is there differential co-expression between a signature gene and a pathway regarding the clinical outcome?

Full model: $\sim$ `metastases * signature.gene`

Reduced model: $\sim$ `metastases + signature.gene`

```
$ANOVA
                SSQ   DF          MS
Effect    3.694421   31 0.11917486
Error   113.210105 2852 0.03969499

$test.result

F.value 3.002265e+00
p.value 5.741837e-08
p.perm  5.000000e-04

$terms
[1] "(Intercept)"  "metastases"  "signature.gene"
[4] "metastases:signature.gene"
```



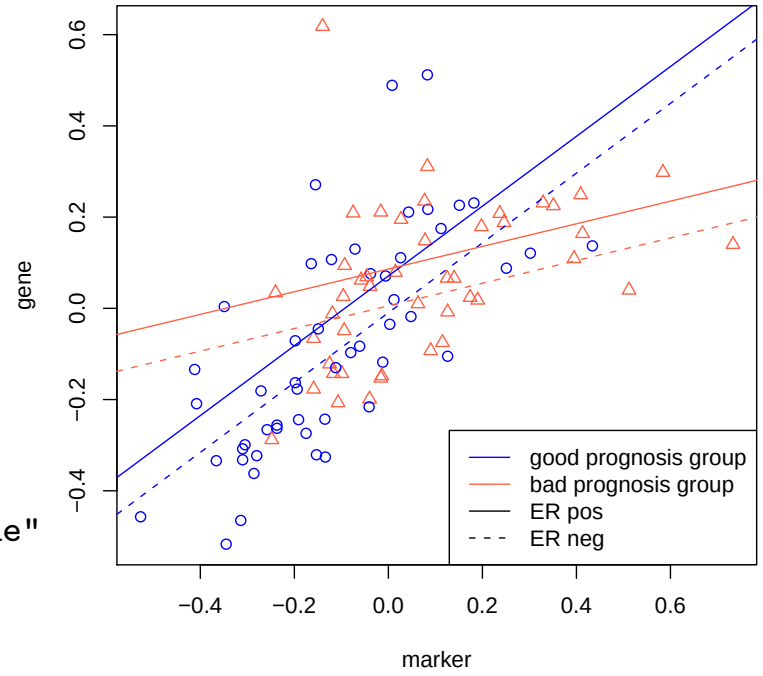
Example

With covariate adjustment

Full model: ∼ `metastases * signature.gene + ER`

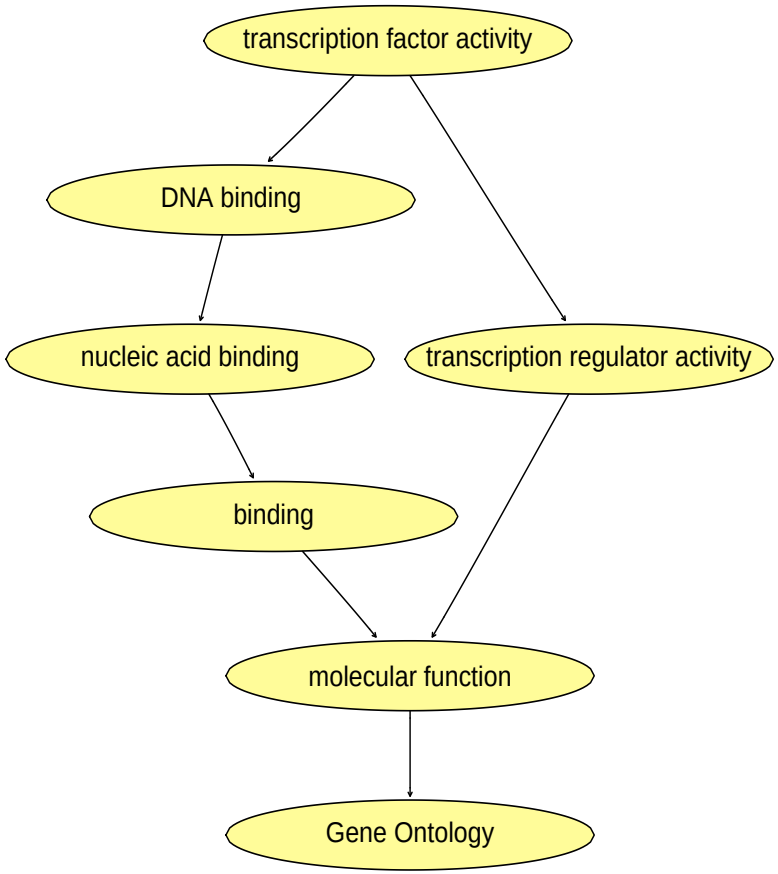
Reduced model: ∼ `metastases + signature.gene + ER`

```
$ANOVA
                SSQ DF         MS
Effect    3.665263 31 0.1182343
Error   107.548645 31 0.0381243

$test.result

F.value 3.101284e+00
p.value 2.031171e-08
p.perm  5.000000e-04

$terms
[1] "(Intercept)"  "metastases"  "signature.gene"
[4] "ER"  "metastases:signature.gene"
```



Gene Ontology

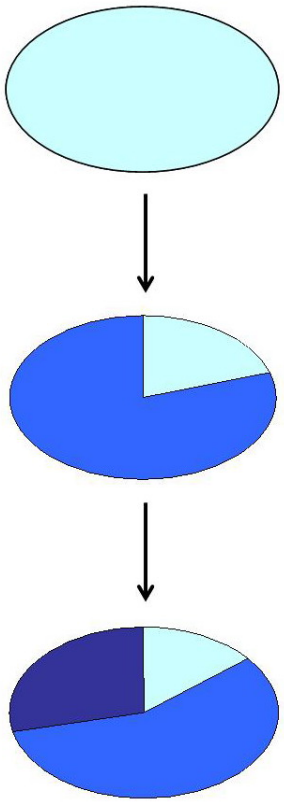
- The Gene Ontology (GO) is a controlled vocabulary to describe gene and gene product attributes (*http://www.geneontology.org/*)
- Three Ontologies: Molecular Function, Biological Process, Cellular Component
- Relations between GO terms are displayed in directed acyclic graphs – direction from specific to general terms



Gene Ontology

- Genes known to be associated with some attributes are mapped to corresponding GO terms

- Inheritance: each gene associated with some term is also mapped to all its ancestors



Biological Questions

- Provide biological meaning to a list of genes found differentially expressed by means of an over-representation analysis
  → Gene set enrichment approaches

- Find biological coherences regarding differential gene expression
  → Holistic approaches

- Find essentially enriched terms given the relationship structure of the GO
  → GO inheritance approaches

Some Methods

Gene set enrichment approaches

- Define a list of differentially expressed genes and score GO terms using the hypergeometric distribution
- Define a Kolmogorov-Smirnov like running sum test statistic for ranked genes *(Subramanian et al. (2005))*

Holistic approaches

- Score GO terms directly using `GlobalAncova` *(Mansmann and Meister (2005))* or `globaltest` *(Goeman et al. (2005))*
- Category approach *(Gentleman (2005))*

GO inheritance approaches

- Decorrelating the GO (*Alexa et al. (2006)*)
- Parent-child approach (*Grossmann et al. (2006)*)

Drawbacks of Gene Set Enrichment

- 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
- p-value correction is crucial (correlations between genes, power of detecting genes, ...)

How to Find Significant Regions in the GO?

- Since many tests are performed some correction for multiple testing is required
- There are already various adjustment methods but it would be desirable to incorporate the structure of the GO
- GO inheritance approaches make use of parent – child relationships in order to find truly enriched nodes
- Those are modifications of classical gene set enrichment
- Alternative: Find significant regions in the graph based on the family of global null hypotheses

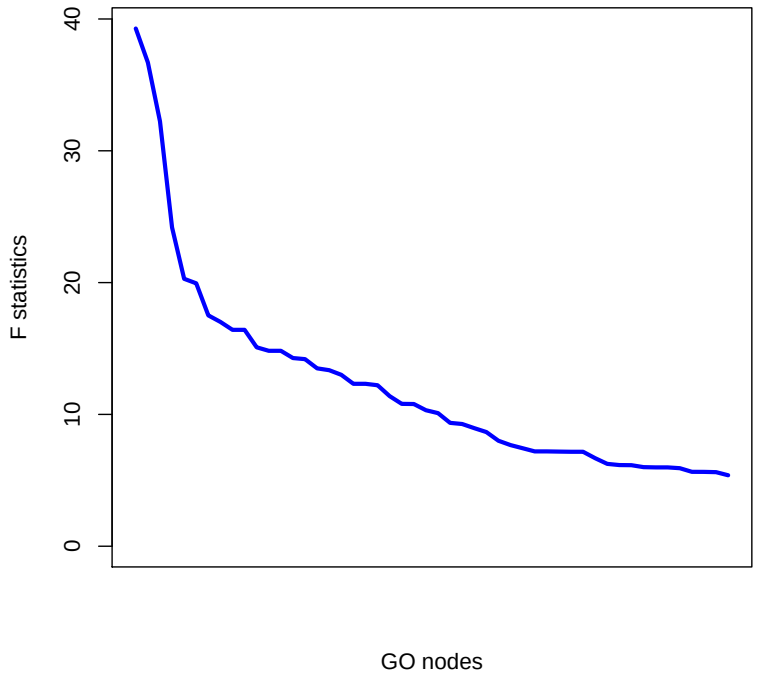
How to Find Significant Regions in the GO?

- For each of the N nodes of the GO graph consider the null hypothesis of no differential expression and corresponding global test statistic F_n
- P_{H_0} denotes the distribution of vector $(F_n)_{n=1,\ldots,N}$ under the family of null hypotheses
- Applying global tests to all nodes yields observed values of the test statistics F_n^{obs}
- Goal: Find a set of nodes $W \subset \{1, \ldots, N\}$ for which

$$P_{H_0}\left\{F_w > F_w^{obs}, \text{ for all } w \in W\right\} < \alpha$$

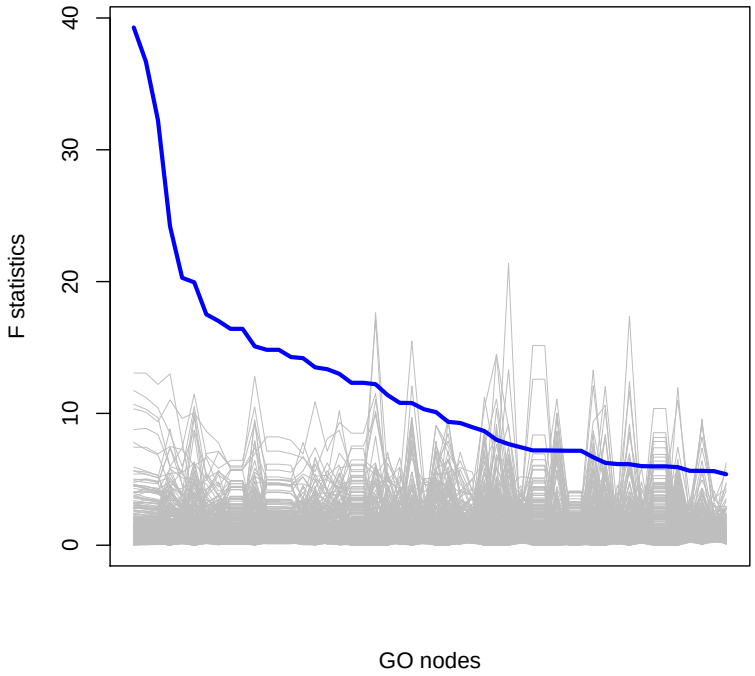
How to Find Significant Regions in the GO?

- First idea to define subset W: Sort nodes by corresponding observed statistics and find suitable cutoff
- The approach is carried out permutation based



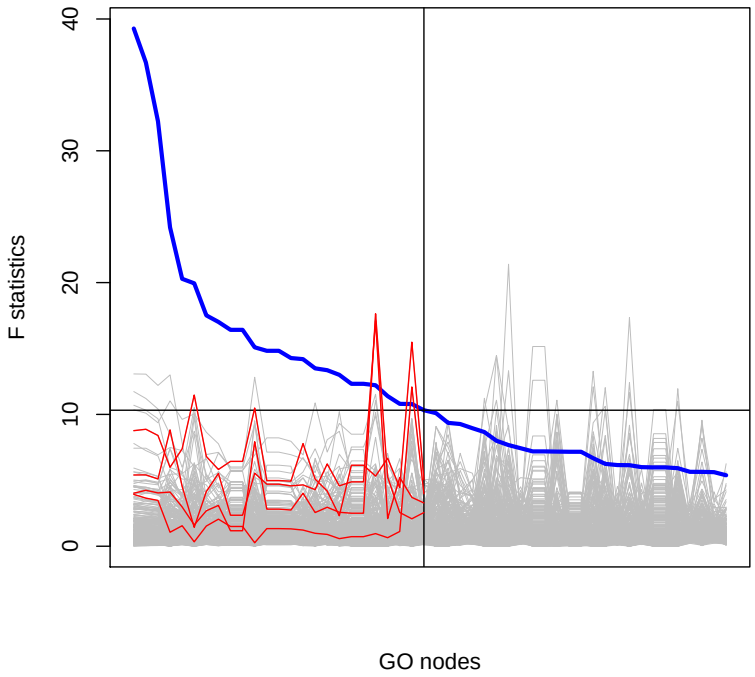
How to Find Significant Regions in the GO?

- First idea to define subset W: Sort nodes by corresponding observed statistics and find suitable cutoff
- The approach is carried out permutation based



How to Find Significant Regions in the GO?

- First idea to define subset W: Sort nodes by corresponding observed statistics and find suitable cutoff
- The approach is carried out permutation based



Outlook

- Define subset W of interesting terms by ordering nodes according to the graph structure
- Use full annotation of GO terms or only the 'node-specific' genes (without genes of respective descendant nodes)
- Use complete annotation but shrink expression values of offspring genes before calculating global statistics
- MANOVA approach with additional variable indicating whether a gene is specific at a node

Literature

1. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006.
2. Box, GEP. Some theorems on quadratic forms applied in the study of ANOVA problems (I). Annals of Mathematical Statistics 1954; 25.
3. Dudoit, S, Keles S, van der Laan MJ. Multiple tests of association with biological annotation metadata. U.C. Berkeley Division of Biostatistics Working Paper Series 2006. Working Paper 202.
4. Gentleman R. Using GO for statistical analyses. 2004.
5. Gentleman R. Using Categories to Model Genomic Data. 2005.
6. 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.
7. Grossmann S, Bauer S, Robinson PN, Vingron M. An Improved Statistic for Detecting Over-representated Gene Ontology Annotations in Gene Sets. Research in Computational Molecular Biology: 10th Annual International Conference, RECOMB 2006, Venice, Italy, April 2-5, 2006. Proceedings: Lecture Notes in Computer Science 3909, Mar 2006, 85-98.
8. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005.
9. Mansmann U, Meister R. Testing differential gene expression in functional groups. Methods Inf Med 2005; 44(3).
10. 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.
11. van t'Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536.
12. Westfall PH, Young SS. Resampling-based multiple testing – examples and methods for p-value adjustment. New York: Wiley; 1993.