Group Testing

– Enhancing the analyses of gene expression data –

Adrian Alexa

alexa@mpi-inf.mpg.de

Computational Biology and Applied Algorithmics

Max Planck Institute for Informatics

D-66123 Saarbrücken

Group Seminar, Saarbruecken, 11th of January, 2007

Overview

- Motivation
- Testing gene sets
  - Tests based on counts: **Fisher's exact test** [Khatri and Draghici, 2005]
  - Comparing two-sample distributions: **KS test,** t**-test** [Subramanian, A., *et al.*, 2005] and [Efron and Tibshirani, 2006]
  - Category analysis: **GlobalTest, Category** [Goeman, J. J., *et al.*, 2004] and [Jiang and Gentleman, 2007]
- Gene Ontology issues
  - Accounting for groups dependencies
  - Assessing the performance of different tests
  - Simulation scenario and results
- GO, time series and dimension reduction (preview)
  - Can GO be used for dimension reduction?
  - Time Series data

➢ **Question 1: Find genes correlated with a given phenotype.**

- Standard approach is to treat genes independently (gene-wise differential expression)
- High score genes are further investigated for underlying biology.
- **Problem:** Noisy expression data can induce large number of candidate genes (differentially expressed genes), out of which only few are biologically interesting.

➢ Question 1: Find genes correlated with a given phenotype.

- Standard approach is to treat genes independently (gene-wise differential expression)
- High score genes are further investigated for underlying biology.
- **Problem:** Noisy expression data can induce large number of candidate genes (differentially expressed genes), out of which only few are biologically interesting.

➢ Question 2: Find groups of genes correlated with a given phenotype.

- More biological knowledge is added using predefined groups of genes: GO, KEGG, Transpath, etc.
- Gene set enrichment: gene-wise analysis followed by enrichment analysis of the gene sets.
- Holistic approach: differentially expressed gene sets.

➢ **Main idea:**

- If you look for candidate genes correlated with a given phenotype it is better to look for interesting gene groups first.
- Grouping the genes into biological predefined clusters can be seen as a filtering: genes from the same group share the same biology.

➢ **Analysis steps:**

1. Derive score for genes (p-value, t-statistic, even gene expression value itself).
2. Map genes to biological groups and compute significance of these groups using a suitable test statistic.
3. Screen the significant biological groups for candidate genes.

➢ **Main idea:**

- If you look for candidate genes correlated with a given phenotype it is better to look for interesting gene groups first.
- Grouping the genes into biological predefined clusters can be seen as a filtering: genes from the same group share the same biology.

➢ **Analysis steps:**

1. Derive score for genes (p-value, t-statistic, even gene expression value itself).
2. Map genes to biological groups and compute significance of these groups using a suitable test statistic.
3. Screen the significant biological groups for candidate genes.

➢ **Advantages:**

- Easier to find biologically related genes sharing the same pattern.
- Fewer groups to be investigated for differential expression than individual genes.
- Easier to find genes with sensible small change in expression.

Example

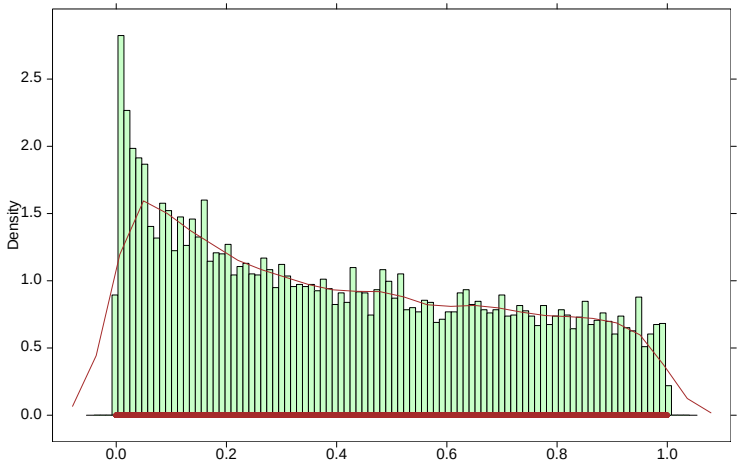
- Analysis of synergistic effect between hypomethylation and changes on chromosome 8 for prostate cancer patients (23 patients).
- **Statistical model:**

$$\log(geneExpr) = \alpha_0 + \alpha_1 I_{hypo} + \alpha_2 I_{chorm8} + \alpha_3 I_{hypo} I_{chrom8} + \epsilon$$

- **Test for interaction effect:**

$$H_0 : \alpha_3 = 0 \quad \text{vs} \quad H_1 : \alpha_3 \neq 0$$

- **Distribution of p-values**

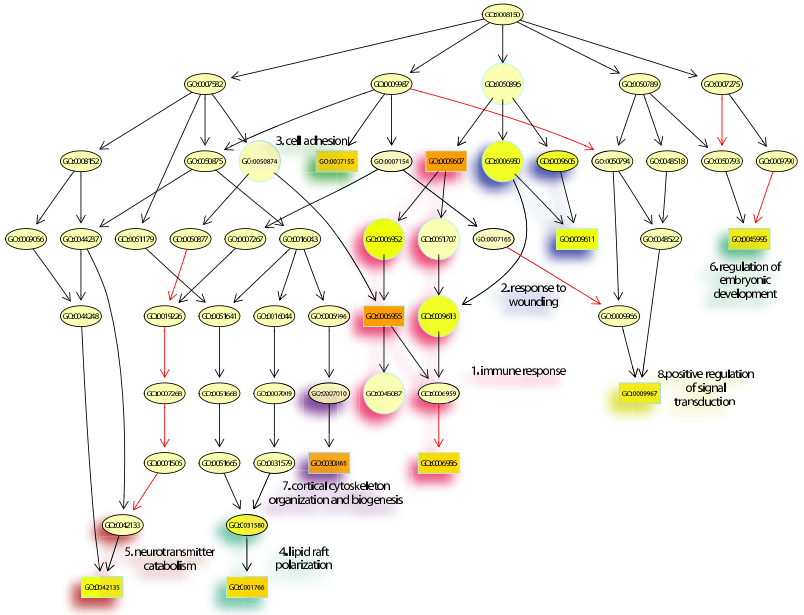


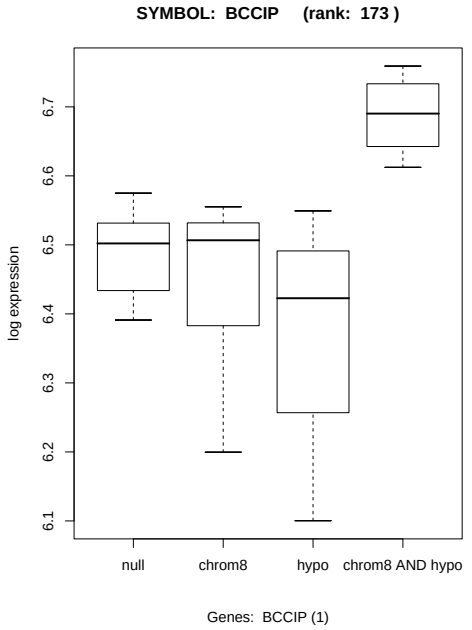
Example

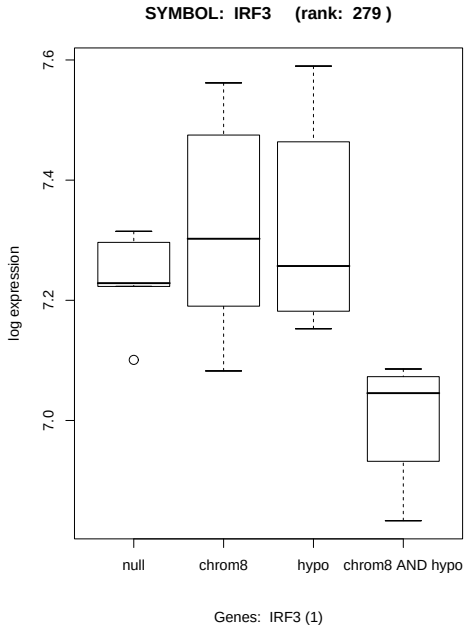
- Interpretation of high ranking GO groups.
- New candidate genes found by screening the high rank GO groups. These genes are hard to find by analysing the list of differentially expressed genes!
- Wolfang A. Schulz, Adrian Alexa, Volker Jung, Christiane Hader, Michele J. Hoffmann, Masanori Yamanaka, Sandy Fritzsche, Agnes Wlazlinski, Mirko Müller, Thomas Lengauer, Rainer Engers, Andrea R. Florl, Bernd Wullich, Jörg Rahnenführer:

  Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer,

  Molecular Cancer, *to be accepted*.



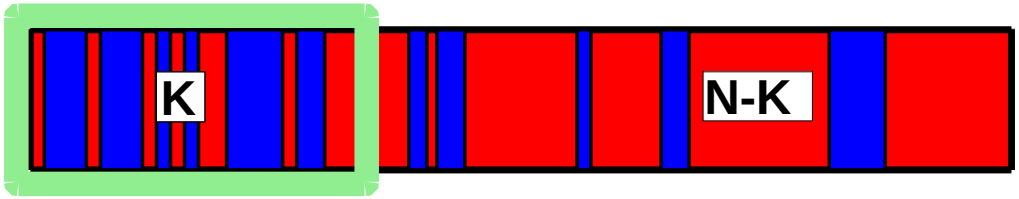




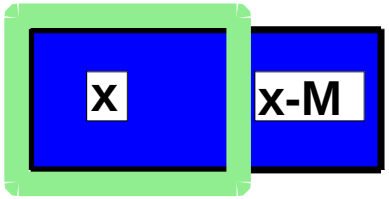
- Motivation
- **Testing gene sets**
  - Tests based on counts: **Fisher's exact test** [Khatri and Draghici, 2005]
  - Comparing two-sample distributions: **KS test,** t**-test** [Subramanian, A., *et al.*, 2005] and [Efron and Tibshirani, 2006]
  - Category analysis: **GlobalTest, Category** [Goeman, J. J., *et al.*, 2004] and [Jiang and Gentleman, 2007]
- Gene Ontology issues
- GO, time series and dimension reduction (preview)

- **Group enrichment:** given a **gene group** with some biological function, analyse the positions of these genes in the **ordered list**. The **gene group** is relevant, if all genes are among the top genes in the **ordered list**.
- **Idea:** Sort genes according to some score (diff. expression) and investigate the ranks of the members of group **A** (the biological function) in this list.
- Define cutoff and count members of group **A** below and above cutoff. Basically, one wants to compare the following ratios:

$$\frac{\mathbf{K}}{\mathbf{N}} \leq \frac{\mathbf{x}}{\mathbf{M}}.$$



N (gene on the microarray)



M (genes in group)

For computing the significance of a gene set, we can use a *hypergeometric test*:

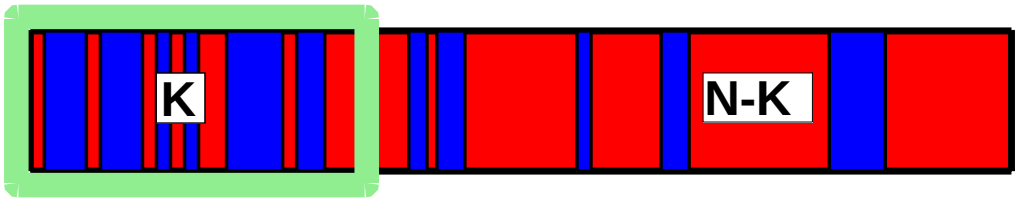
- N genes are on microarray
- Bio is a GO term
  - M genes $\in Bio$
  - $N - M$ genes $\notin Bio$
- Let K be the no. of significant genes
- What is the probability of having exactly x genes from K of type Bio ?

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

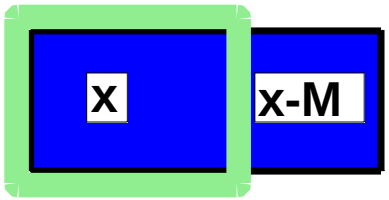
- This is the probability of getting exactly x by chance (not what we want)

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}.$$

(also called Fisher's exact test)



N (gene on the microarray)



M (genes in group)

Fisher's exact test

For computing the significance of a gene set, we can use a *hypergeometric test*:

- N genes are on microarray
- Bio is a GO term
  - M genes $\in Bio$
  - $N - M$ genes $\notin Bio$
- Let K be the no. of significant genes
- What is the probability of having exactly x genes from K of type Bio ?

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

- This is the probability of getting exactly x by chance (not what we want)

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}.$$

(also called Fisher's exact test)

➢ Depends on p-value adjustment procedure. No clear way to define K.

➢ Since genes are divided into two disjoints sets (differentially and non-differentially expressed genes) small but consistent differential expression is not accounted for.

Category analysis

- Generalize the concept of differentially expressed gene to differentially expressed groups.
- Enrichment analysis using count statistics can not capture gene expression patterns for a gene group.
- Gentleman's category and Goeman's global test aggregates per gene statistics/gene expression within a gene group.
- The idea behind is that small but coordinate changes in gene expression are relevant for phenotypic differences.

- Generalize the concept of differentially expressed gene to differentially expressed groups.
- Enrichment analysis using count statistics can not capture gene expression patterns for a gene group.
- Gentleman's category and Goeman's global test aggregates per gene statistics/gene expression within a gene group.
- The idea behind is that small but coordinate changes in gene expression are relevant for phenotypic differences.

- The association between genes and the gene groups can be seen as an incidence matrix **A**.
- The numbers of genes in each category is given by the row sums.
- The number of groups a gene belongs to is given by the column sums.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1B} \\ \vdots & & & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KB} \end{pmatrix}$$

$$a_{ij} = \begin{cases} 1, & \text{if } g_j \in GO_i \\ 0, & \text{if } g_j \notin GO_i. \end{cases}$$

- The correlation between the phenotype and the genes is summarised in a vector $\mathbf{Z}$:

$$\mathbf{Z} = (z_1, z_2, \ldots, z_b).$$

z_i is the gene-wise statistic for gene i, for example t-statistic between two groups.

- The gene set statistics are defined by:

$$\mathbf{X} = \frac{\mathbf{A} \cdot \mathbf{Z}}{\sqrt{row\_sum(\mathbf{A})}}, \quad \text{or more generally:} \quad \mathbf{X} = f(\mathbf{A}, \mathbf{Z}).$$

f can be defined as: *mean, median, sign of the test*, etc.

➢ The correlation between the phenotype and the genes is summarised in a vector $\mathbf{Z}$:

$$\mathbf{Z} = (z_1, z_2, \ldots, z_b).$$

z_i is the gene-wise statistic for gene i, for example t-statistic between two groups.

➢ The gene set statistics are defined by:

$$\mathbf{X} = \frac{\mathbf{A} \cdot \mathbf{Z}}{\sqrt{row\_sum(\mathbf{A})}}, \quad \text{or more generally:} \quad \mathbf{X} = f(\mathbf{A}, \mathbf{Z}).$$

f can be defined as: *mean, median, sign of the test*, etc.

➢ When t-statistic is used for $\mathbf{Z}$, then $\mathbf{X} \sim N(0, 1)$ (standardised sum of normals). This usually holds when the number of samples is large and the summands are independent!

➢ Otherwise, permutation test can be used for assessing the significance of the statistic.

- Permuting genes: is the test statistic for given group unusual?
- Permuting samples: is the group statistic unusual w.r.t. the entire expression?

➢ **The main idea:** Compare correlation structure of members of investigated group with correlation structure of phenotype values:

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

➢ **The main idea:** Compare correlation structure of members of investigated group with correlation structure of phenotype values:

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

➢ Test Statistic:

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

- Y is the phenotype vector.
- $R \sim XX^T$ is the covariance matrix of the gene expression data of members of G.
- The first sum is taken over genes, the second over samples

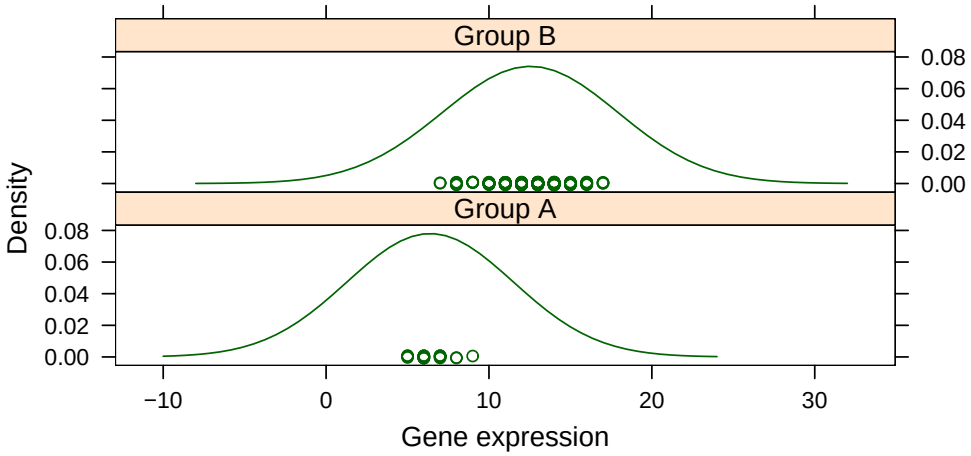
➢ Two interpretations of test statistic Q:

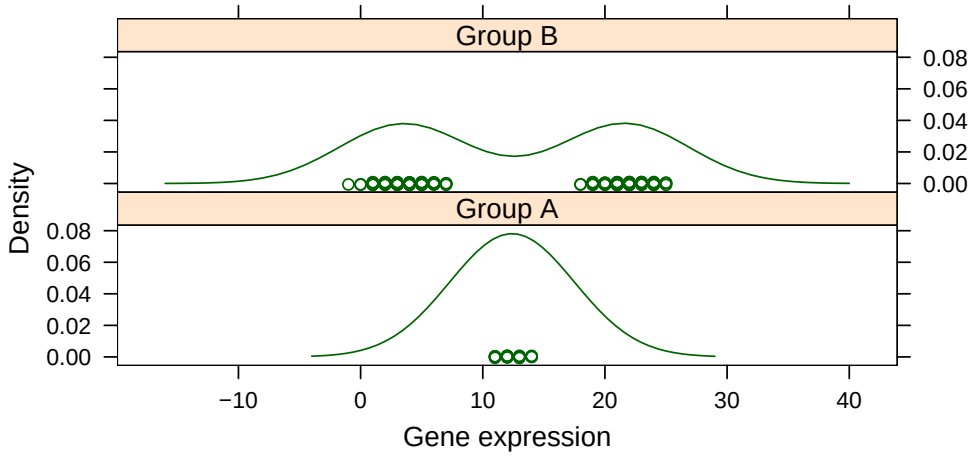
- Average covariance of expression vector of members of G and phenotype values.
- Quantification of how much covariance structure between expression data resembles covariance structure between phenotype values.

Testing equality of two distributions

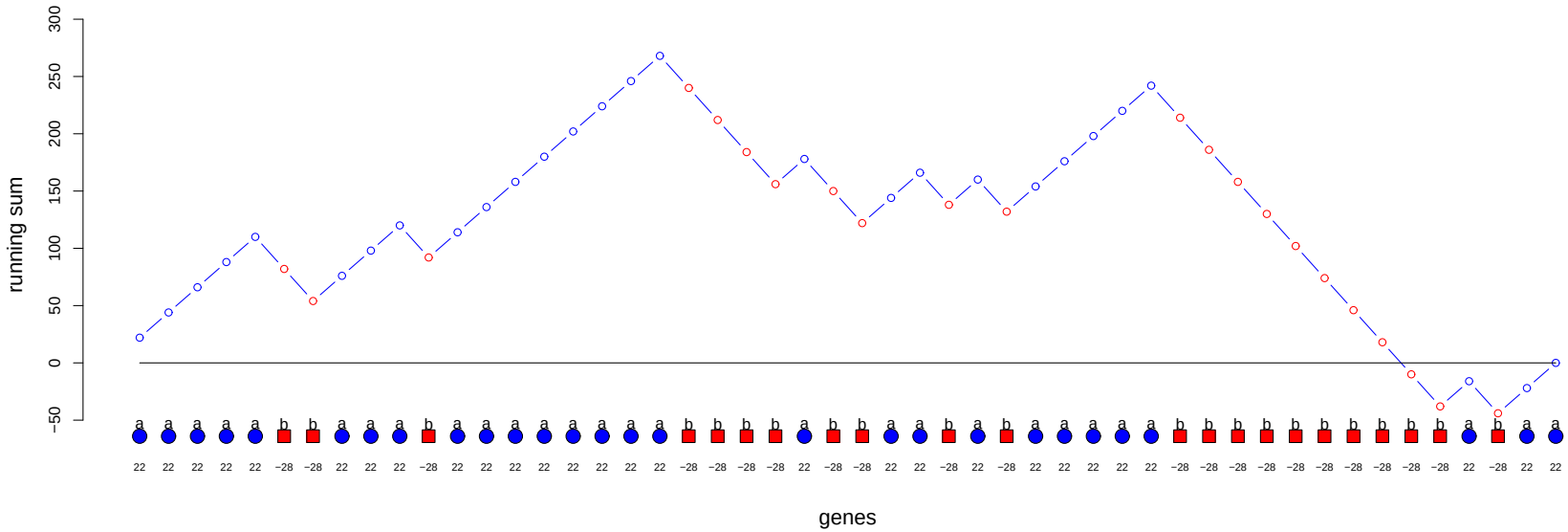
- Fisher's exact test is not optimal due to the loss of information. Genes are partitioned into two sets and the information embedded in the genes below the cutoff is not used. Also the position of the genes is not considered with Fisher's exact test.
- Category analysis accounts only for the gene expression pattern inside the group. Larger groups inherently contain a larger amount of differential expression.

- Fisher's exact test is not optimal due to the loss of information. Genes are partitioned into two sets and the information embedded in the genes below the cutoff is not used. Also the position of the genes is not considered with Fisher's exact test.
- Category analysis accounts only for the gene expression pattern inside the group. Larger groups inherently contain a larger amount of differential expression.
- What we want is to compare the distribution of a gene set, group **A**, with the distribution of all other genes present on the array, group **B**.





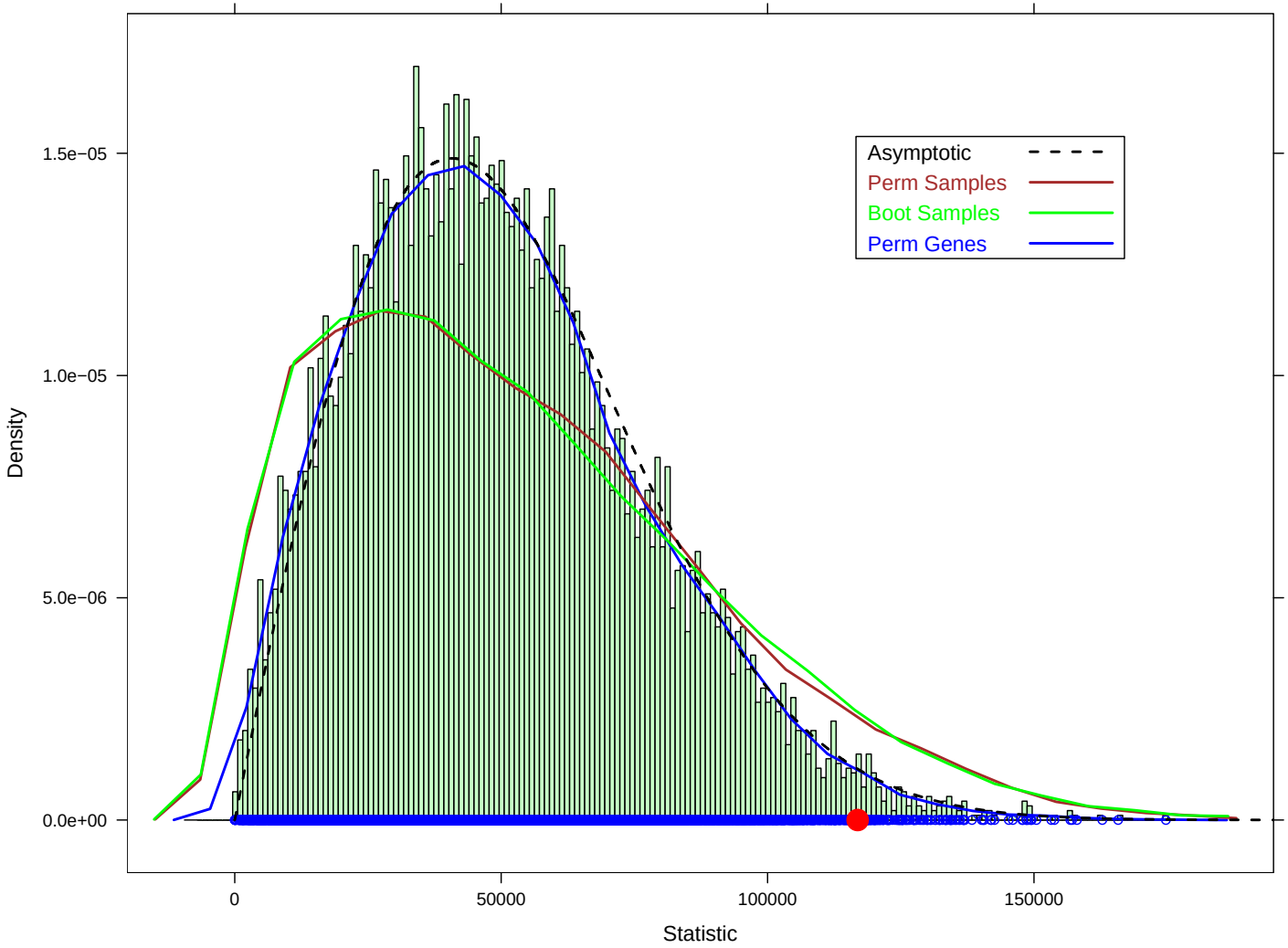
Testing equality of two distributions



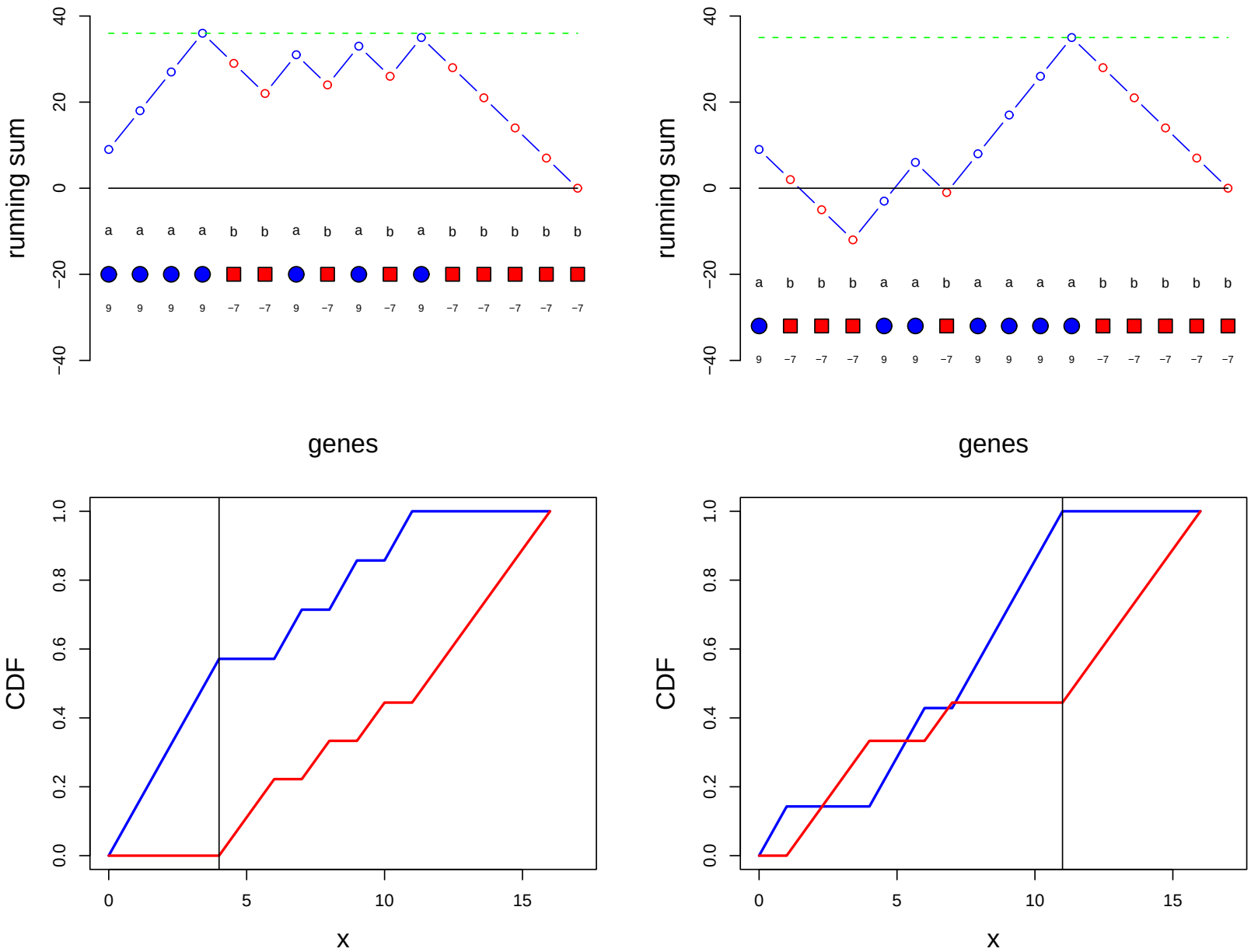
- Genes are ordered with respect to a measure that quantifies the expression differences in the phenotype.
- A running-sum statistic is computed: If the next gene belongs to group **a**, add n_b to the current sum. If not, subtract n_a from the sum. The total sum is always 0.
- Group **a** is found significant if a high value of the maximal deviation from 0 is obtained. This is a two sided test.
- The significance of running-sum statistic is computed by randomly permuting genes (under the null hypothesis that the genes are uniformly mixed between groups).

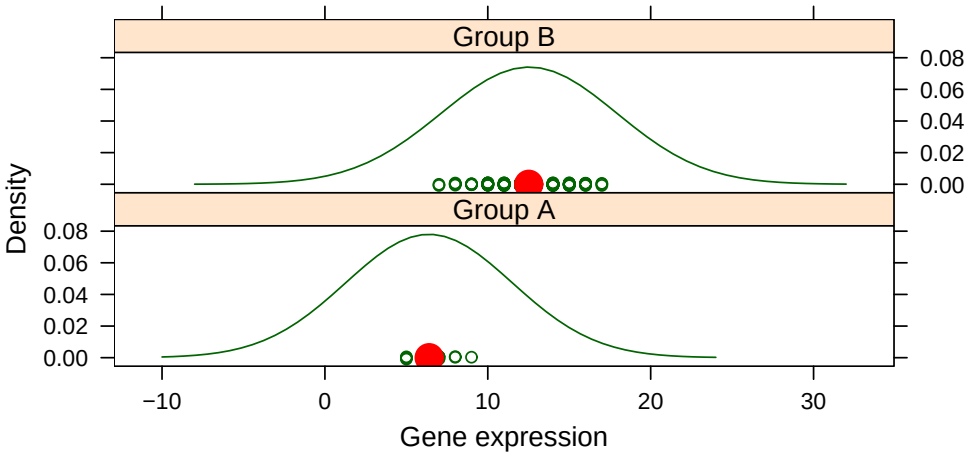
The null hypothesis

- The asymptotic distribution agrees with the one of permuting genes (KS test).
- Permuting samples gives a different null distribution than permuting genes.
- Permutation of samples can be done when sufficient arrays are available (in this case 128).
- Bootstrapping can be used as an alternative to permuting samples.



Shift in distribution





➢ If we are interested in a distribution shift, then a simple t-test can be used between groups $\mathbf{A}$ and $\mathbf{B}$:

$$H_0 : \mu_{\mathbf{A}} = \mu_{\mathbf{B}} \quad \text{versus} \quad H_1 : \mu_{\mathbf{A}} \neq \mu_{\mathbf{B}},$$

where the test statistic is defined by:

$$\mathbf{T} = \frac{\mu_{\mathbf{A}} - \mu_{\mathbf{B}}}{\hat{\sigma}}.$$

➢ $\hat{\sigma}$ is an estimate of the variance. Usually a t-test with equal variance for the two samples gives better results.

➢ **Problems:** not proportional sample sizes: group $\mathbf{A} \sim 10$ genes vs. group $\mathbf{B} \sim 10000$!

- Each test answers a different question!
- Category based tests analyse only the distribution inside the group. This can be an artifact of the whole experiment (each group is differentially expressed).
- **Generalisation:** Group testing can be seen as comparing two multi-dimensional densities.
  - Genes are k-dimensional vectors (k is the number of samples).
  - Group **A** contains n_A genes belonging to a biological group (GO term).
  - Group **B** contains all the other genes from the array.
  - Test statistic to compare the multivariate distributions of group **A**, respectively group **B**:

$$H_0 : F_A = F_B \quad \text{vs.} \quad H_1 : F_A \neq F_B.$$

- Considering genes as k-dim. vectors can avoid the problems with the two stage approach. First compute a gene-wise statistic (t-statistic) an then perform a group test.

- Motivation
- Testing gene sets
- **Gene Ontology issues**
  - Accounting for groups dependencies
  - Assessing the performance of different tests
  - Simulation scenario and results
- GO, time series and dimension reduction (preview)

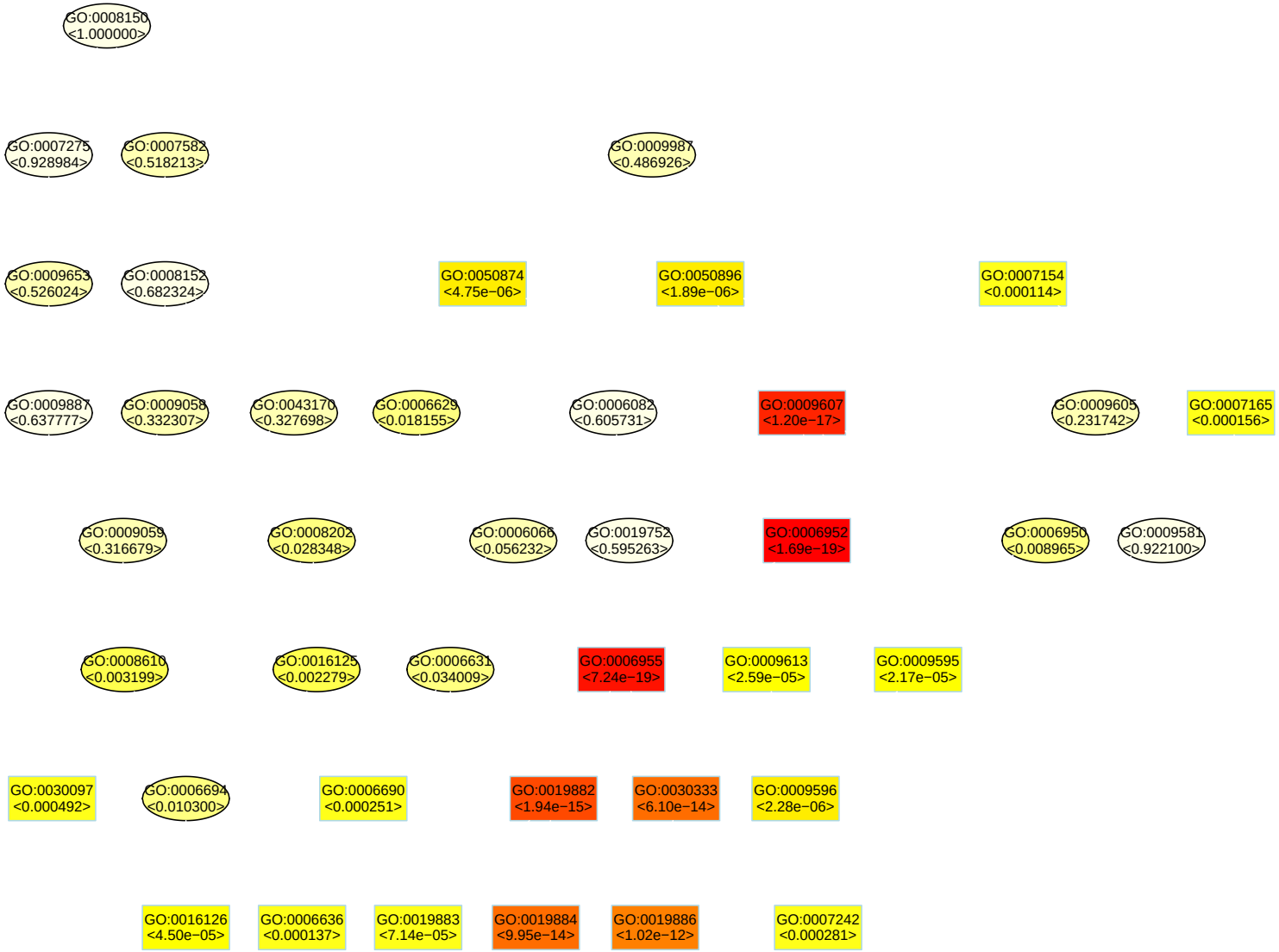
max planck institut informatik

Given:

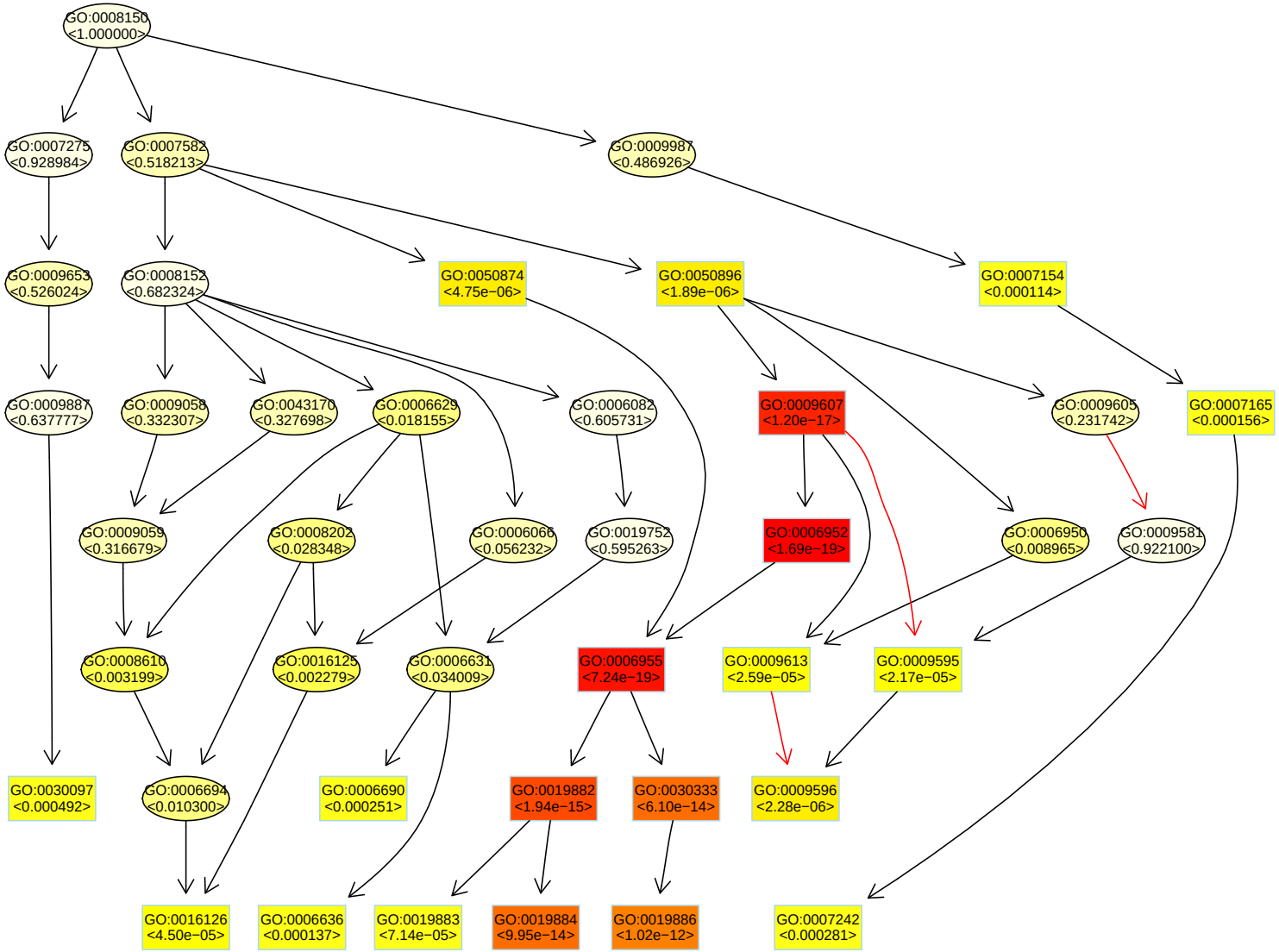
- a directed acyclic graph (GO graph) and a set of items (genes) s.t.:
  - each node in the graph contains some genes
  - the parent of a node contains all the genes of its child
  - a node can contain genes that are not found in the children
- a subset of genes that we call significant genes (differentially expressed genes)

Goal:

- find the nodes from the graph (biological functions) that best represent the significant genes w.r.t some scoring function (some test statistic)

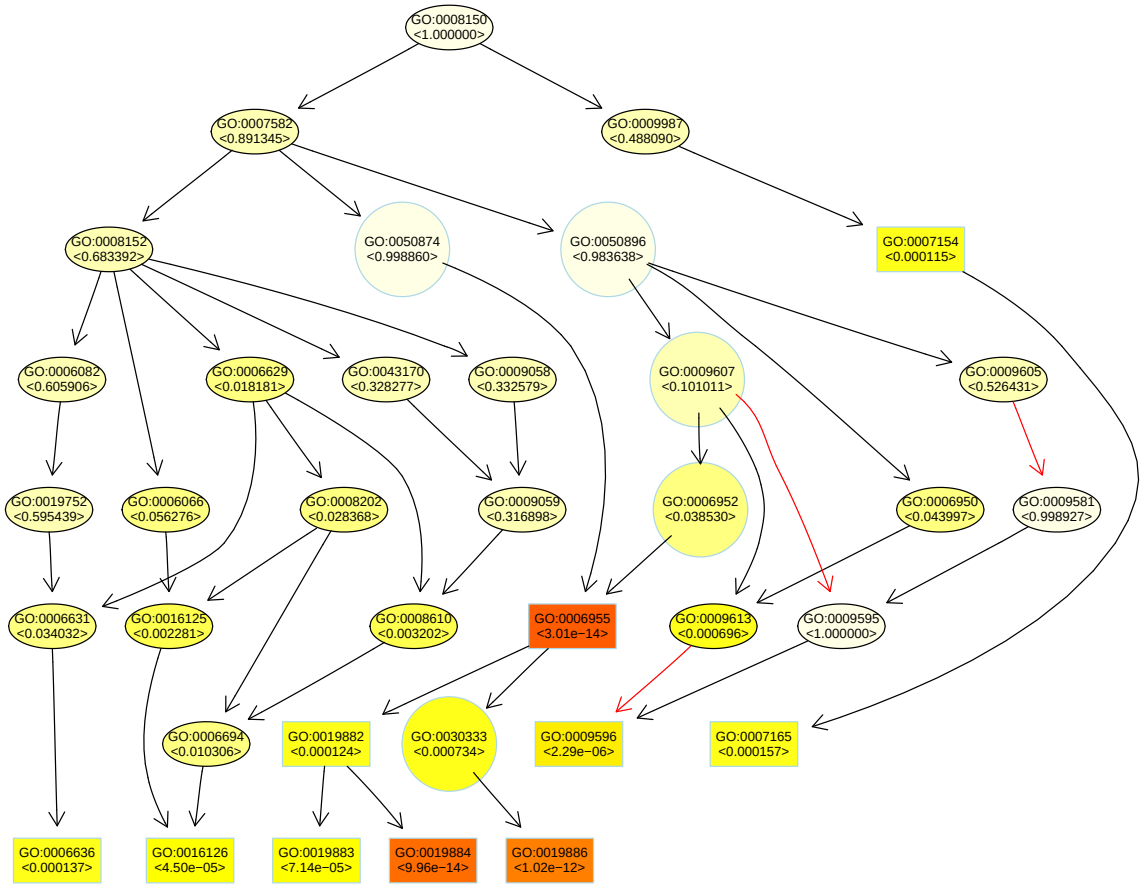


Note: The coloring of the nodes represent the *relative* significance of the GO terms: dark red is the most significant, light yellow is the least significant from the graph



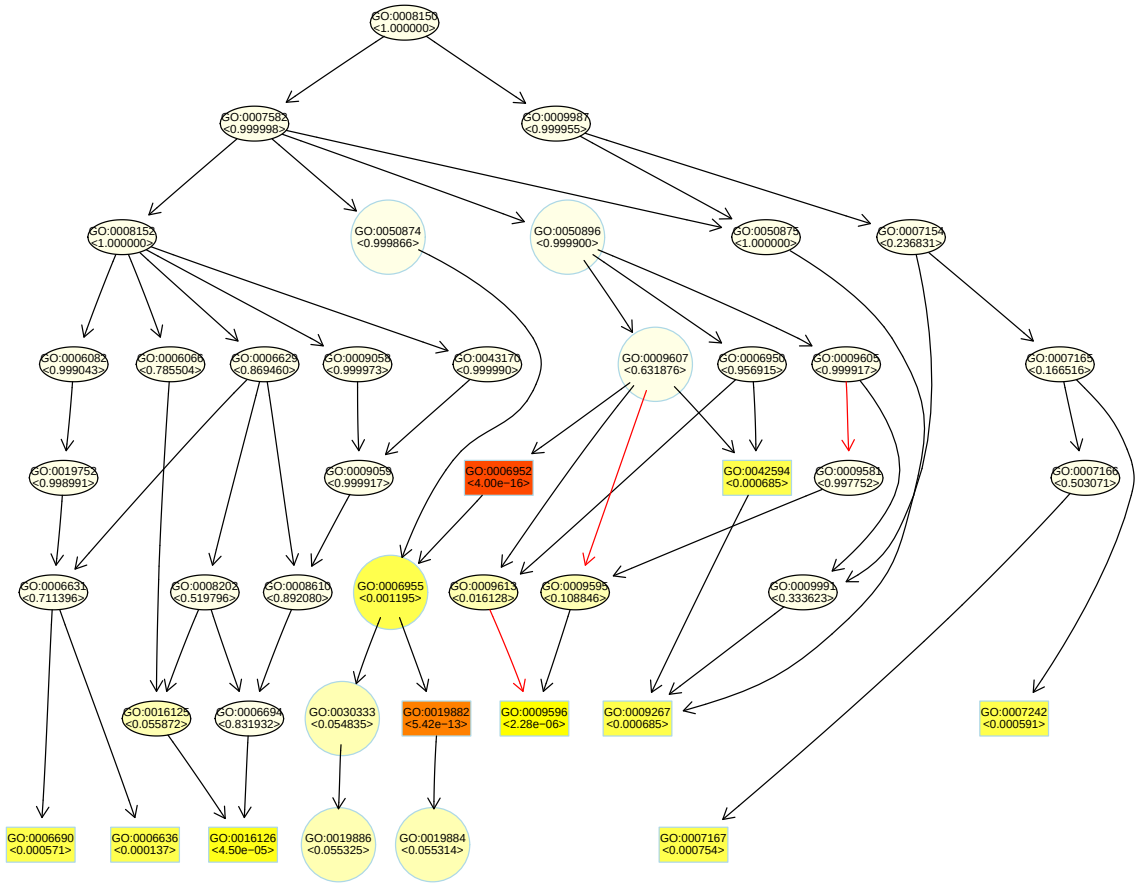
Note: The coloring of the nodes represent the *relative* significance of the GO terms: dark red is the most significant, light yellow is the least significant from the graph

- Nodes are processed bottom-up in the GO graph.
- It iteratively removes the genes annotated to significant GO terms from more general GO terms.
- Intuitive and simple to interpret.



max planck institut informatik

- The genes obtain weights that denote the gene relevance in the significant nodes.
- To decide if a GO term u better represents the interesting genes, the enrichment score of node u is compared with the scores of its children.
- Children with a better score than u better represent the interesting genes; their significance is increased
- Children with a lower score than u have their significance reduced.



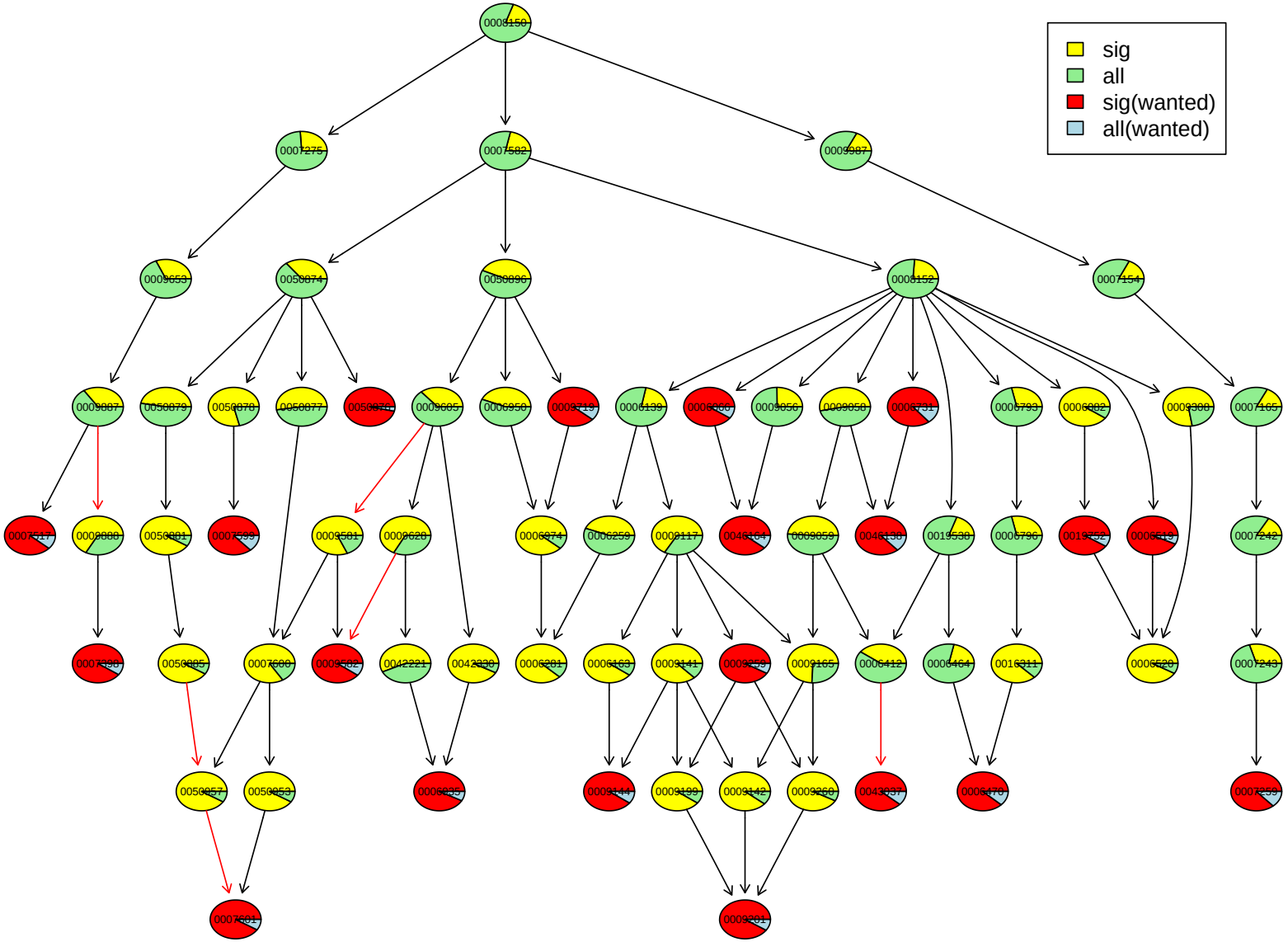
Evaluation on simulated data

- We use the GO graph structure (~ 3000 nodes), and all the genes from HGU95aV2 Affymetrix chip (~ 10000 mapped to the GO graph)
- Select only the nodes that have the number of mapped genes in some range ($10 \ldots 100$)
- Choose randomly a number of nodes (50 in our case) from the selected nodes. These nodes represent the enriched nodes (interesting nodes).
- Set as significant genes all the genes from the enriched nodes.
- Some noise is introduced:
  - Pick 10% from all significant genes
  - Remove them from the significant list
  - Replace the genes that we removed with other genes

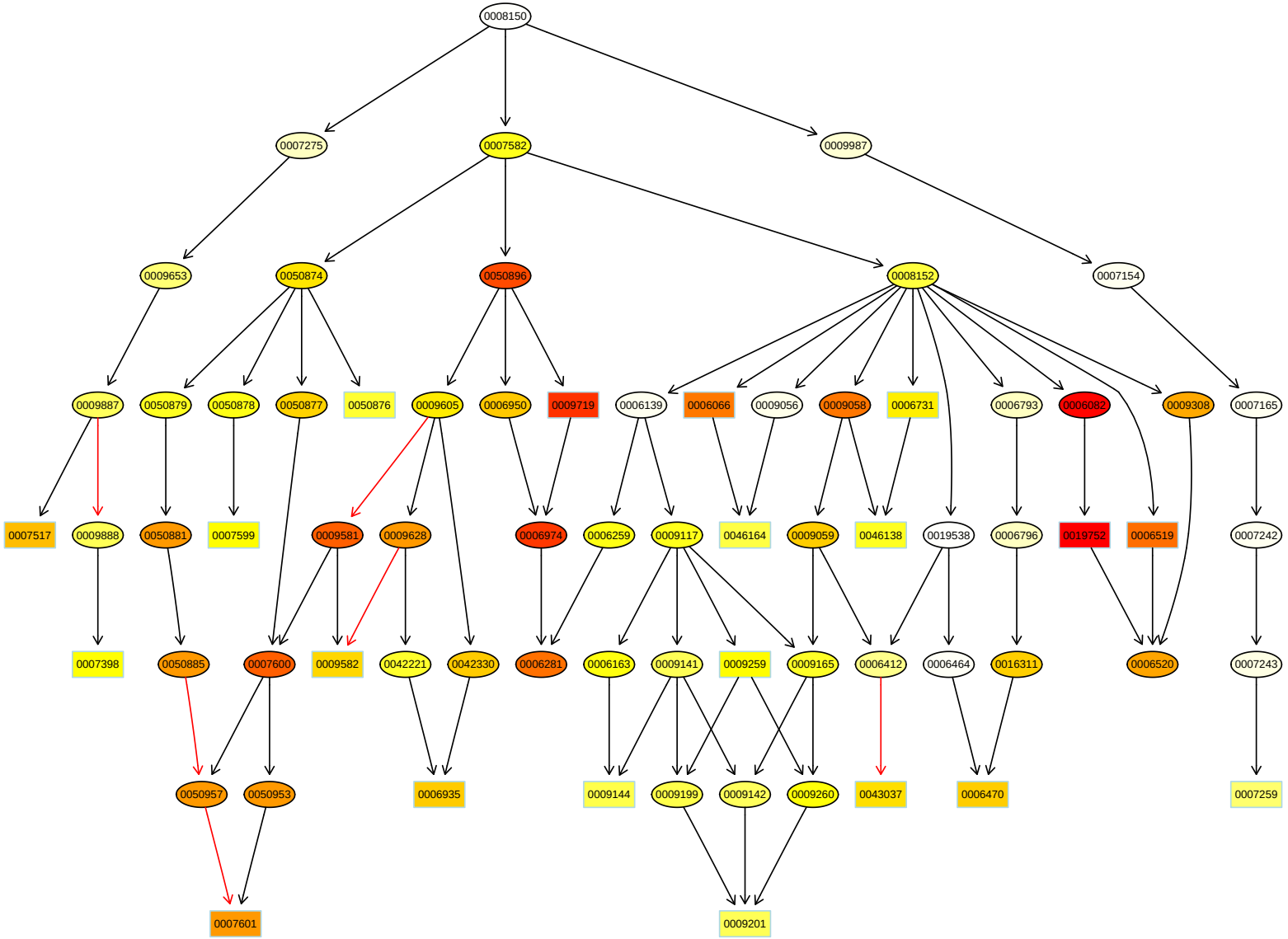
Evaluation on simulated data

- We use the GO graph structure (~ 3000 nodes), and all the genes from HGU95aV2 Affymetrix chip (~ 10000 mapped to the GO graph)
- Select only the nodes that have the number of mapped genes in some range $(10 \ldots 100)$
- Choose randomly a number of nodes (50 in our case) from the selected nodes. These nodes represent the enriched nodes (interesting nodes).
- Set as significant genes all the genes from the enriched nodes.
- Some noise is introduced:
  - Pick 10% from all significant genes
  - Remove them from the significant list
  - Replace the genes that we removed with other genes
- The goal is to recover as best as possible the enriched nodes.

Simulated dataset



Simulated dataset



➢ **Scenario 2:** use of real data sets.

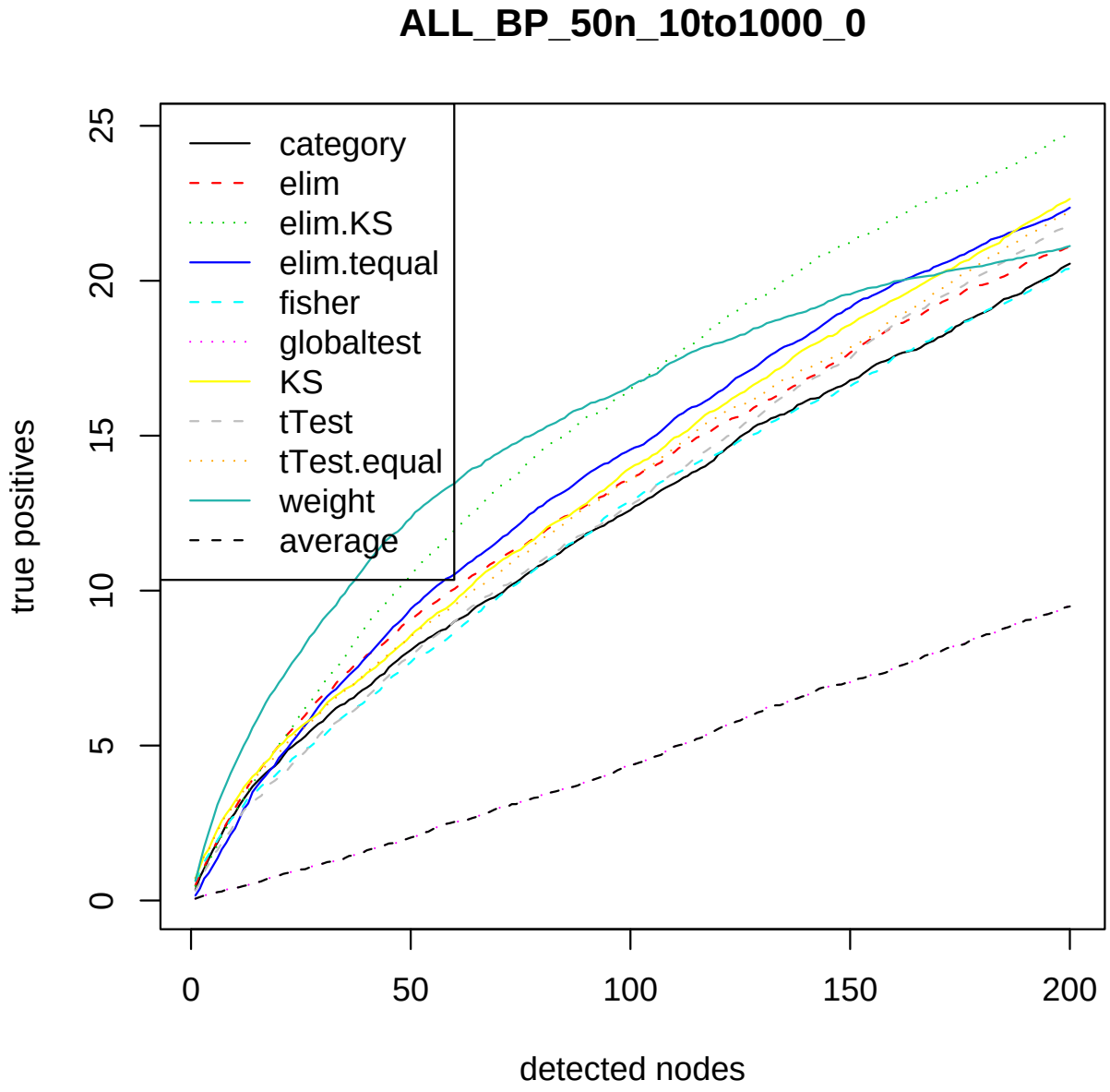
- Issue with first scenario: no measure for differential expression is assigned to genes. Thus, only test based on counts can be applied.
- Define gene group **A** as the group containing all genes annotated to the enriched nodes.
- Sort all genes from the array w.r.t. the correlation with a given phenotype.
- Permute the gene labels such that group A contains the top differentially expressed genes.

➢ **Scenario 2:** use of real data sets.

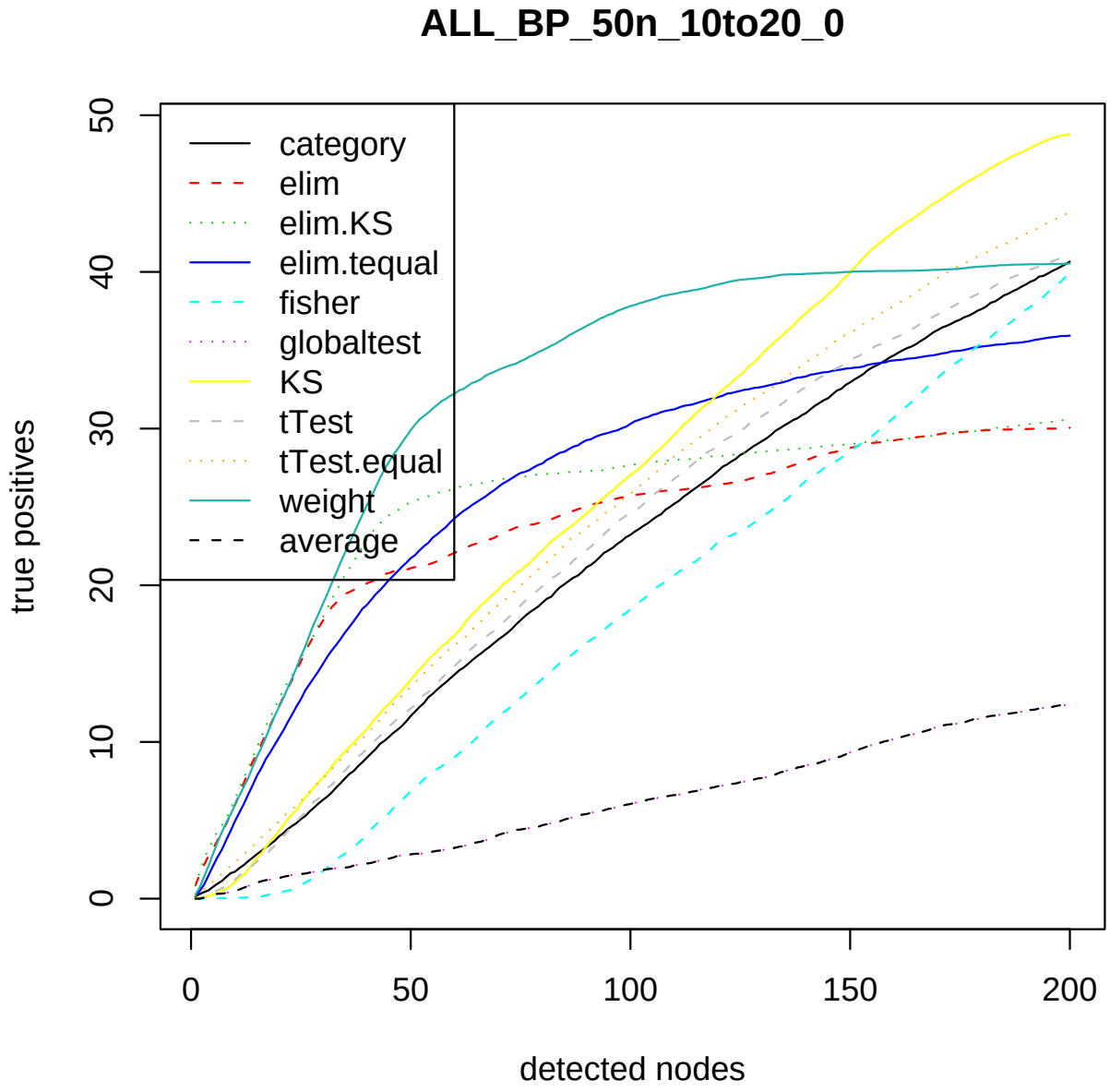
- Issue with first scenario: no measure for differential expression is assigned to genes. Thus, only test based on counts can be applied.
- Define gene group **A** as the group containing all genes annotated to the enriched nodes.
- Sort all genes from the array w.r.t. the correlation with a given phenotype.
- Permute the gene labels such that group A contains the top differentially expressed genes.

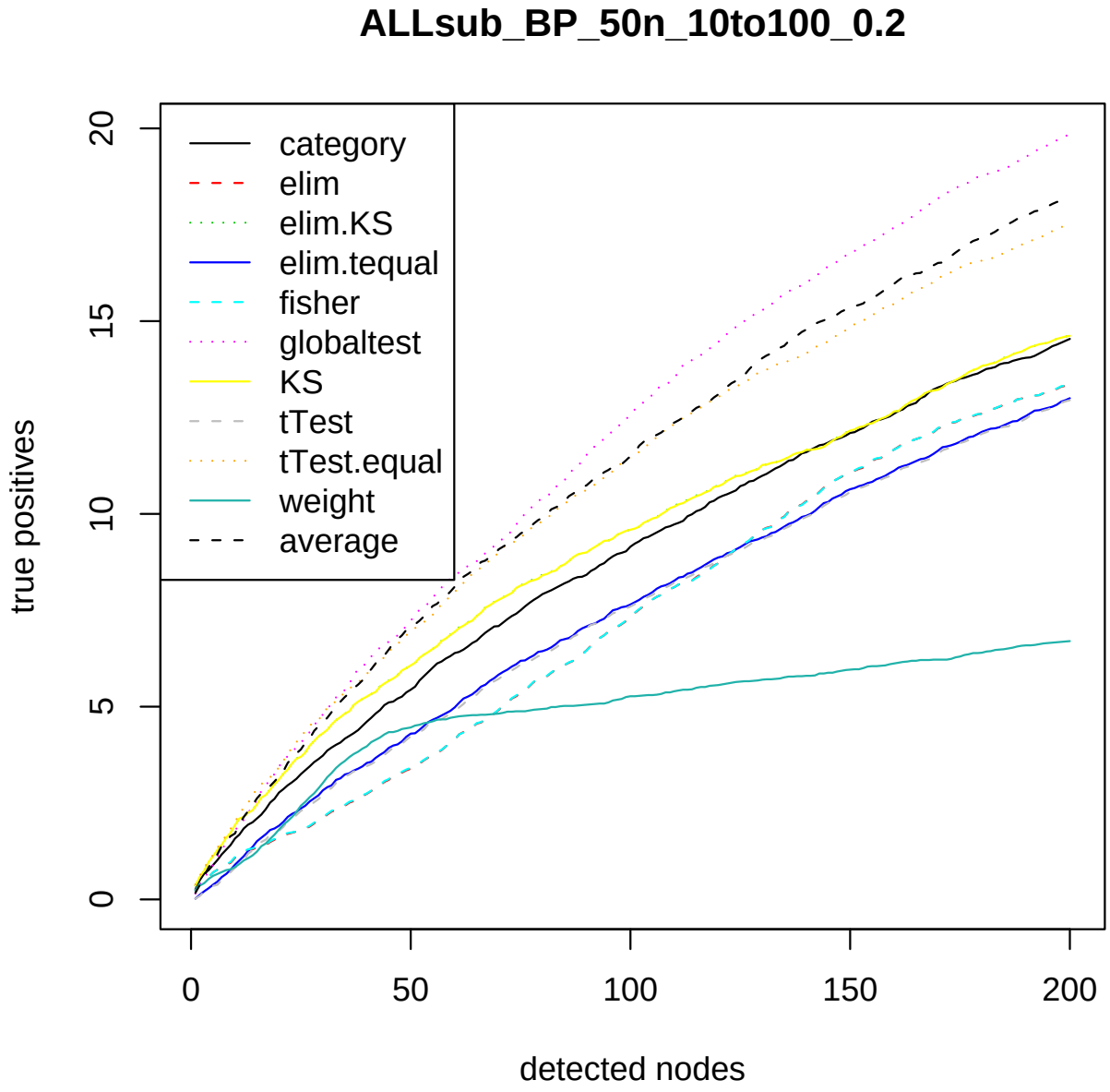
➢ **Scenario 3:** use of real data sets and conserving the correlation structure between the genes.

- Select of some GO nodes as enriched nodes.
- Define gene group **A** as the group containing all genes annotated to the enriched nodes.
- Set a group test procedure as a reference test, for example KS-test or Global test.
- Permute the phenotype and for each permutation apply the test statistic to group A.
- The final dataset is the one with the most extreme test.


ALL_BP_50n_10to1000_0
category
elim
elim.KS
elim.tequal
fisher
globaltest
KS
tTest
tTest.equal
weight
average
true positives
0
5
10
15
20
25
0
50
100
150
200
detected nodes

Quality of GO scoring methods




ALLsub_BP_50n_10to100_0.2
category
elim
elim.KS
elim.tequal
fisher
globaltest
KS
tTest
tTest.equal
weight
average
true positives
0
5
10
15
20
0
50
100
150
200
detected nodes

- Motivation
- Testing gene sets
- Gene Ontology issues
- **GO, time series and dimension reduction (preview)**
  - Can GO be used for dimension reduction?
  - Time Series data

- Since there are so many group testing procedures available and some are based on gene scores one can think of using GO for dimension reduction.
- **Main idea**: For each array first map genes to GO groups and perform enrichment analysis on the obtained GO groups. If the raw expression value is used, then high scoring GOs can be thought as groups accumulating high gene expression.
- This can be seen as a dimension reduction: from ~ 40.000 genes to ~ 2.000 GO groups.
- The issue is which measure is best (Category can be better than KS or t-test).

KS test (row)

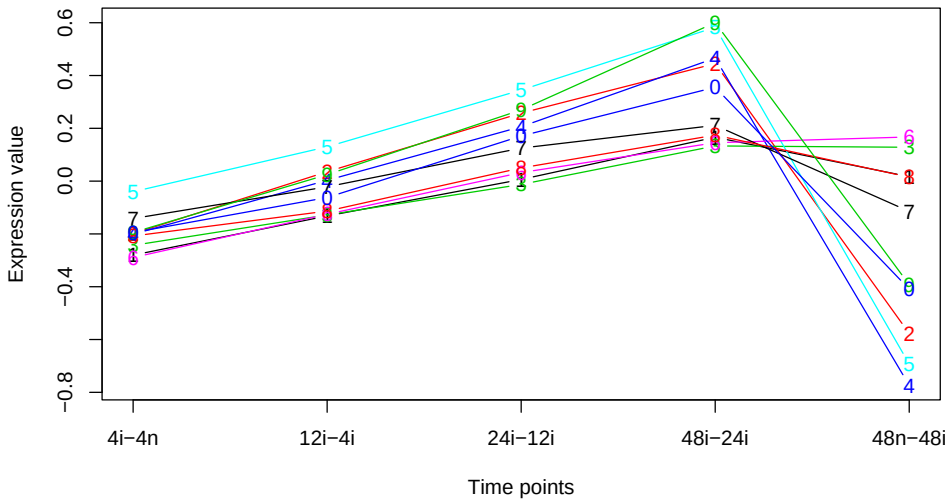
GO ID	4h_C	4h	12h	24h	48h	48h_C
GO:0050874	1	1	1	1	1	1
GO:0006952	2	2	2	2	2	2
GO:0006955	3	3	4	3	3	3
GO:0007186	4	5	3	5	5	5
GO:0009607	5	4	5	4	4	4
GO:0007154	6	6	6	7	7	6
GO:0007275	7	8	7	9	9	7
GO:0007267	8	7	9	6	6	8
GO:0007166	9	9	8	8	8	9
GO:0050877	10	11	11	10	10	10

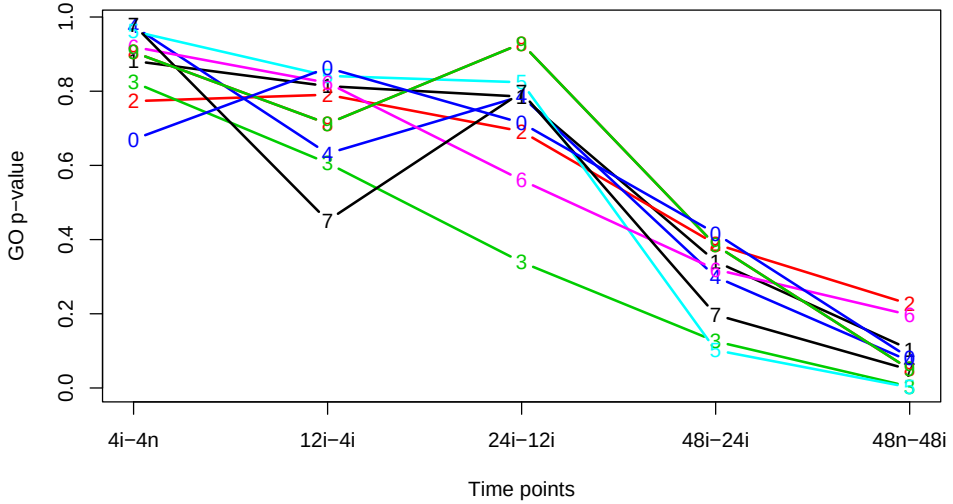
t-test (row)

GO ID	4h_C	4h	12h	24h	48h	48h_C
GO:0050874	1	3	1	7	2	2
GO:0007154	2	4	5	8	4	3
GO:0007186	3	5	7	9	3	4
GO:0007166	4	6	6	11	5	5
GO:0006952	5	7	9	10	6	6
GO:0007275	6	9	10	15	9	7
GO:0007165	7	8	8	14	11	9
GO:0006955	8	10	12	12	7	8
GO:0009607	9	11	11	13	8	10
GO:0007267	10	12	13	16	10	11

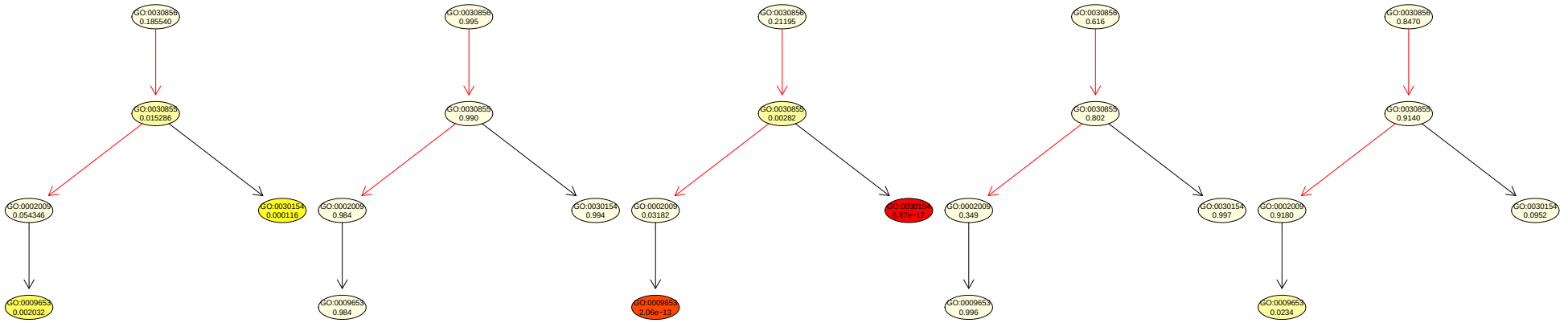
KS-test ($t_{i+1} - t_i$)

GO ID	4h-4h_C	12h-4h	24h-12h	48h-24h	48h_C-48h
GO:0050874	1	3269	1	1873	1869
GO:0006952	2	3309	2	2413	1421
GO:0009607	3	3306	4	2393	1699
GO:0006955	4	3322	3	2568	1334
GO:0007154	5	3390	7	3254	85
GO:0007166	6	3395	9	2513	433
GO:0007186	7	3396	5	1938	1317
GO:0007165	8	3391	12	3332	65
GO:0007275	9	3281	6	2992	193
GO:0050896	10	2266	10	2459	2438





4h-4h_C	12h-4h	24h-12h	48h-24h	48h_C-48h



References

[Efron and Tibshirani, 2006] Efron, B. and Tibshirani, R. (2006). On testing the significance of sets of genes. Technical report, Bepts. of Statistics, Stanford University.

[Goeman, J. J., *et al.*, 2004] Goeman, J. J., *et al.* (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

[Jiang and Gentleman, 2007] Jiang, Z. and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics*, Advance Access.

[Khatri and Draghici, 2005] Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595.

[Subramanian, A., *et al.*, 2005] Subramanian, A., *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550.

Acknowledgements

- MPI Informatics, Saarbrücken

  Jörg Rahnenführer

  Prof. Thomas Lengauer

  Joachim Büch

- Department of Urology, Heinrich-Heine-University, Düsseldorf

  Prof. Wolfgang A. Schulz

- IBE, LMU, Munchen

  Manuela Hummel

  Prof. Ulrich Mansmann