

From a gene list to biological function

- testing functional groups of genes -

Adrian Alexa

`alex@mpi-sb.mpg.de`

Computational Biology and Applied Algorithmics

Max Planck Institute for Informatics

D-66123 Saarbrcken

Courses in Practical DNA Microarray Analysis, München, May 12, 2005

- The Microarray experiments provide a **long list of genes**.
- Typical studies analyze genes **one by one**:
 1. samples are divided into two groups: **disease vs. healthy** and the genes are **ranked** according to **differential expression**.
 2. genes are sorted according to **correlation** of the expression values with a **phenotype** measurement.

These studies result in an **ordered list** of genes.

- **More important is the group enrichment:**
 - given a **set of genes** with some **biological function**, analyze the positions of these genes in the **ordered list**.
 - the biological function is **relevant**, if all genes are among the **top genes** in the **ordered list**.

- Gene sets:
 - Gene Ontology (GO) terms
 - Metabolic pathways
 - MIPS classes
 - Chromosomes
 - Classes defined via transcription factors
 - Gene sets obtained from other previous experiments

➤ Remark 1:

The score and the gene set must be chosen independently!

➤ Remark 2:

The dependence between gene sets usually make the statistical interpretation of the result harder!

Main idea: Sort genes according to some score and analyze **positions** of members of the investigated gene group in this list.

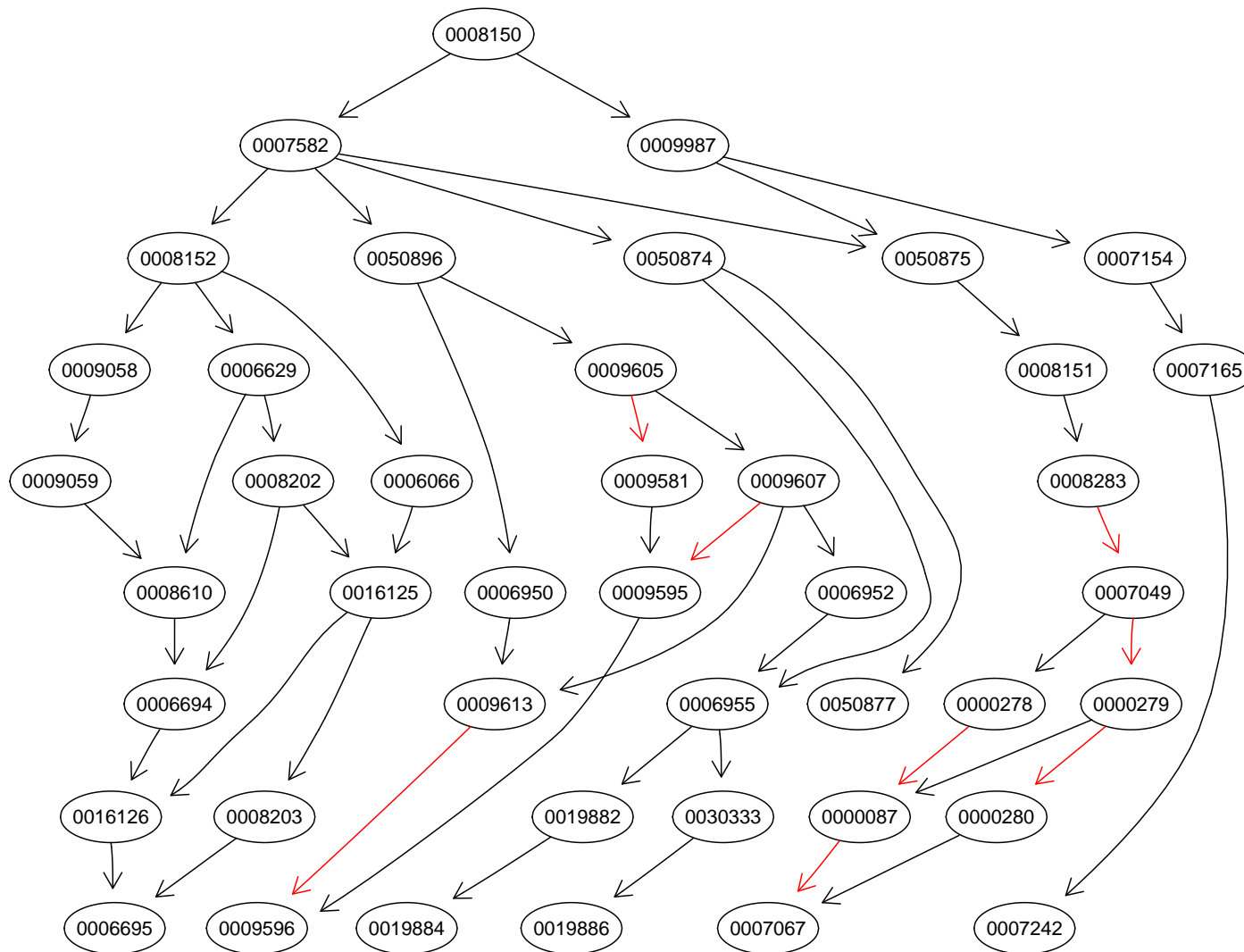
- We want to know if the members of group **a** have significantly **small ranks** (higher in the list). If this is the case, then the group is **enriched**.
- There are two approaches:
 1. Define cutoff and count members of group **a** below and above cutoff.
 2. Analyze distribution of all ranks of members of group **a**.

Gene	Score	Group
$\text{gene}_{\sigma(1)}$	score 1	a
$\text{gene}_{\sigma(2)}$	score 2	b
$\text{gene}_{\sigma(3)}$	score 3	a
$\text{gene}_{\sigma(4)}$	score 4	a
.....
$\text{gene}_{\sigma(100)}$	score 100	b
$\text{gene}_{\sigma(101)}$	score 101	a
.....
$\text{gene}_{\sigma(9905)}$	score 9905	b

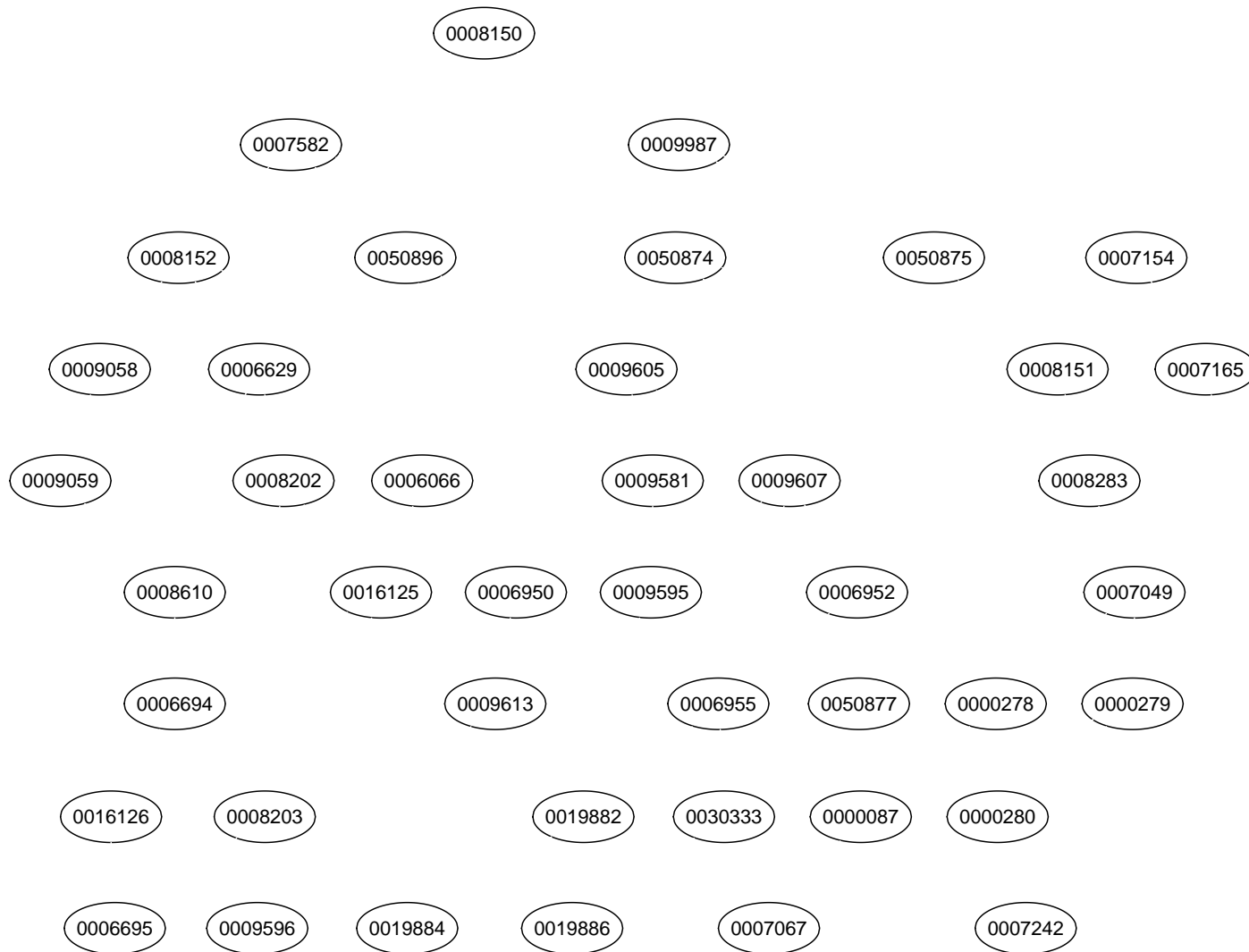
- Using a predefined cutoff
- Optimizing the cutoff
- Global test

- **Using a predefined cutoff**
- Optimizing the cutoff
- Global test

- Obtain the **Gene Expression Data** from the microarrays experiments (this is the normalized and cleaned data: [Long list of genes](#))
- Select a **set of significant genes** (use some test statistic: *t-test*, *permutation-test*)
- Map all the genes to the corresponding **GO terms**
- Analyze the GO terms for significance ([pretty tricky](#))
 - Remark:** the GO terms are considered to be independent and the significance is computed for each one separately.
- Current implementation of these ideas: [Onto-Expres](#), [GOstat](#), [GoMiner](#), [FunSpec](#), [FatiGO](#), [GO::TermFinder](#)
- All the above methods try to account for the [dependence](#) between the GO terms by applying **Multiple Hypothesis Correction** (Bonferroni, FDR)



Note: The labels of the nodes are the GO IDs: 0008150 \cong GO:0008150



Note: The labels of the nodes are the GO IDs: $0008150 \cong \text{GO:}0008150$

Small example: suppose that we have a GO term for which we expect ~ 10 genes to be significant.

genes expected	genes in data	
10	10	random
10	12	still random
10	20	better than random
10	40	significant

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)

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

(also called Fisher's exact test)

The score for a GO term is the **degree of independence** between the two characteristics:

$\mathcal{A} = \{\text{gene is in the list of significant genes}\}$ and $\mathcal{B} = \{\text{gene is found in the GO term}\}$.

	Significant genes	Not significant genes	Sum
Genes in G	$ \text{sigGenes} \cap \text{funcGenes} $	$ \overline{\text{sigGenes}} \cap \text{funcGenes} $	$ \text{funcGenes} $
Genes in \overline{G}	$ \text{sigGenes} \cap \overline{\text{funcGenes}} $	$ \overline{\text{sigGenes}} \cap \overline{\text{funcGenes}} $	$ \overline{\text{funcGenes}} $
Sum	$ \text{sigGenes} $	$ \overline{\text{sigGenes}} $	$ \text{allGenes} $

Testing the independence of two groups in the above contingency table corresponds to **Fisher's exact test**.

	GO:0006955	GO:0009059
Term name	immune response	macromolecule biosynthesis
Definition	Any process involved in the immunological reaction of an organism to an immunogenic stimulus	The formation from simpler components of macromolecules, large molecules including proteins, nucleic acids and carbohydrates
Ontology	BP	BP
# mapped genes	780	568

- The genes are sorted based on a two sided t -test statistic. There are a total of 9905 genes on the array.
- A **cutoff** of 559 is chosen (the number of genes which are found significant at a level $\alpha = 0.01$ test after a Bonfferoni adjustment procedure is employed).

Contingency table for GO:0006955

	Significant genes	Not significant genes	Sum
Genes in G	107	673	780
Genes in \bar{G}	452	8673	9125
Sum	559	9346	9905

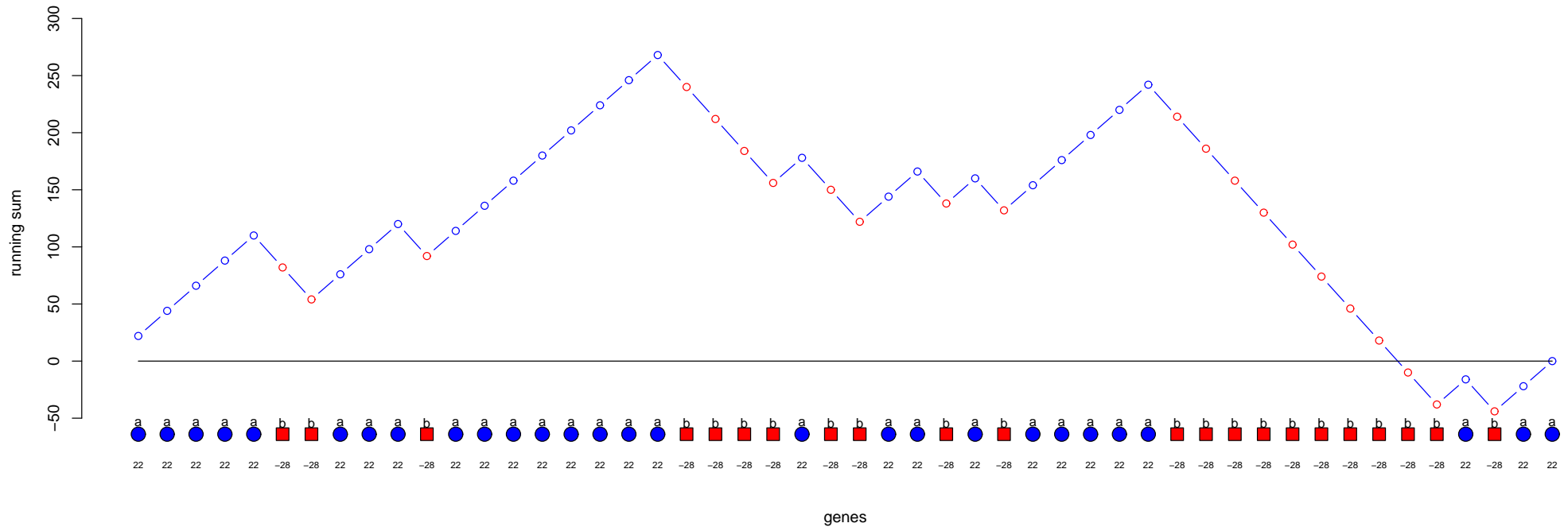
Contingency table for GO:0009059

	Significant genes	Not significant genes	Sum
Genes in G	35	533	568
Genes in \bar{G}	524	8813	9337
Sum	559	9346	9905

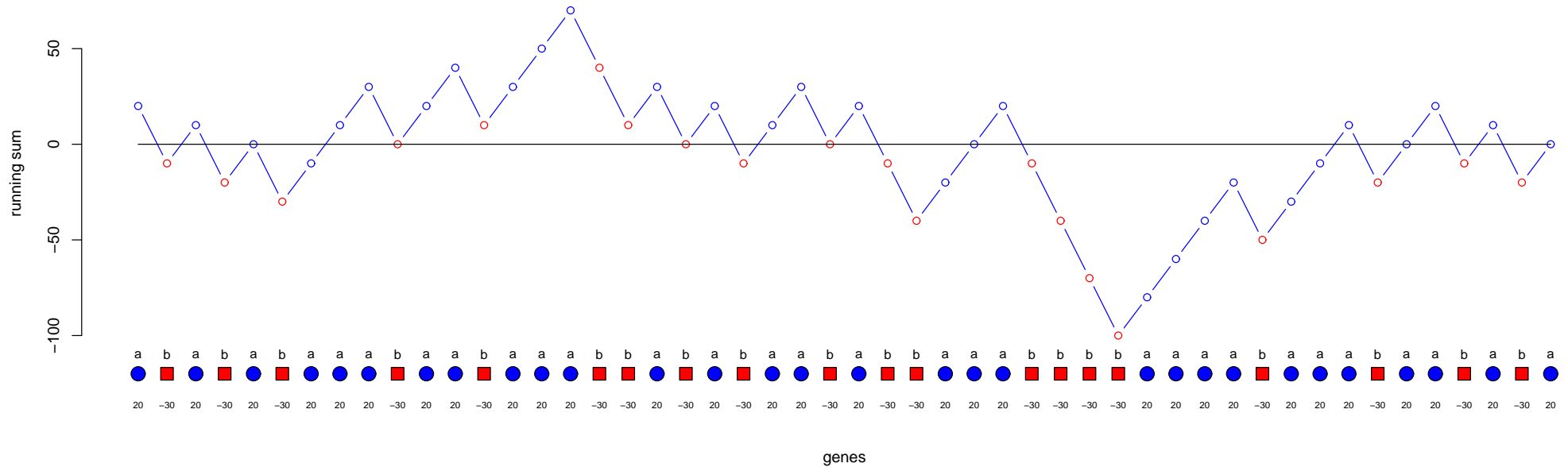
	GO:0006955	GO:0009059
Observed	107	33
Expected	44.020	32.055
Standard deviation	6.186	5.339
raw p -value (Fisher)	7.3e-19	0.3166
adj p -value (Fisher)	7.3e-15	1
raw p -value (Z score)	1.2e-24	0.291

- Using a predefined cutoff
- **Optimizing the cutoff**
- Global test

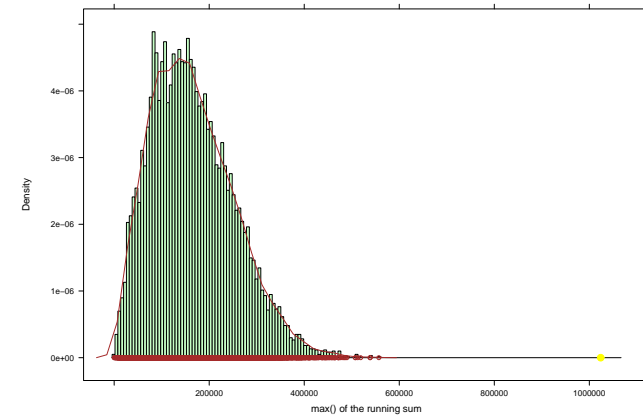
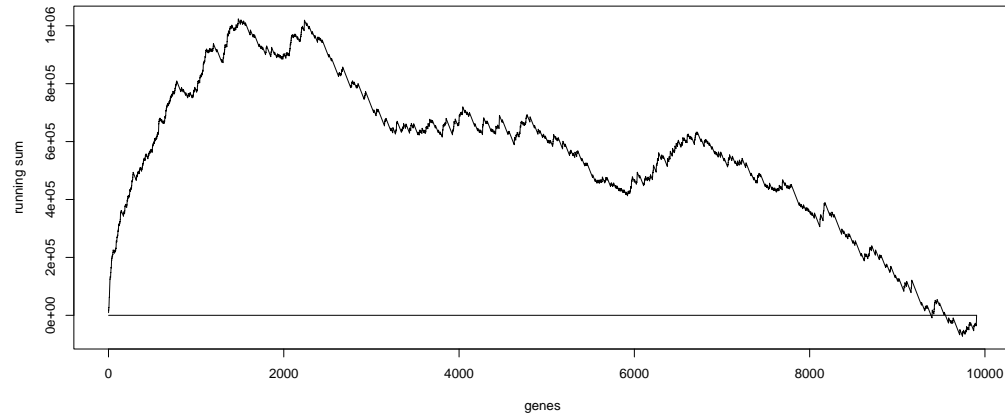
- Fixing a **cutoff** and looking only at the top genes can be sometimes misleading. Also the position of the genes is not considered in the previous approach.
- The information embedded in the genes **below the cutoff** is not used.
- We want to analyze **the distribution of all ranks** of members of group **a**.
- **Main idea:** Use a Kolmogorov-Smirnov like test.
 - 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 this statistic is computed by randomly permuting genes (the null hypothesis is that the genes are uniformly mixed between groups).



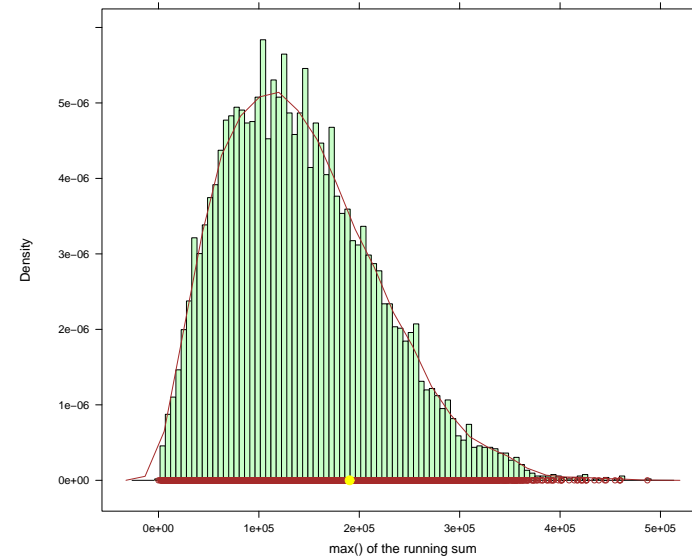
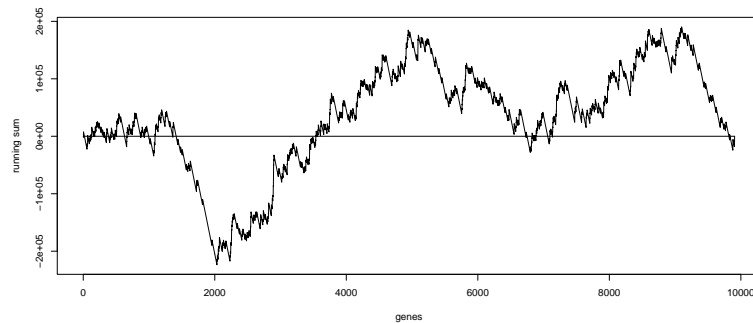
- The maximum deviance from 0 is ~ 270 .
- Since this is positive, we have that group **a** is more enriched than group **b**.



➤ Typical case in which genes are uniformly distributed across the list.



The p -value for GO:0006955 is 0



The p -value for GO:0009059 0.2492

- Using a predefined cutoff
- Optimizing the cutoff
- **Global test**

- Generalize the concept of **differentially expressed gene** to group of genes.
- **The main idea:** Compare **correlation structure** of members of investigated group with correlation structure of phenotype values.

- 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.

- Tim Beissparth and Terry Speed, **GOstat: Find statistically overrepresented Gene Ontologies within a group of genes**, Bioinformatics, Vol. 1 no 1 2004
- Sorin Draghici et al., **Global functional profiling of gene expression**, Genomics 81, 2003
- Jelle J. Goeman et al., **A global test for groups of genes: testing association with a clinical outcome.**, Bioinformatics 20(1):93-99 (2004).
- Berrar Daniel, Dubitzky Werner, Granzow Martin, **A Practical Approach to Microarray Data Analysis**, Kluwer Academic, 2003
- Michael Ashburner, **Gene Ontology: tool for the unification of biology**, Nature genetics, Vol. 25, 2000