Multiple testing with gene expression array data

Anja von Heydebreck

Max–Planck–Institute for Molecular Genetics,
Dept. Computational Molecular Biology, Berlin, Germany

heydebre@molgen.mpg.de

Slides partly adapted from S. Dudoit, Bioconductor short course 2002

Multiple hypothesis testing

❍ Suppose we want to find genes that are differentially expressed between different conditions/phenotypes, e.g. two different tumor types.

❍ On the basis of independent replications for each condition, we conduct a statistical test for each gene $g = 1, \ldots, m$.

❍ This yields test statistics T_g, p-values p_g.

❍ p_g is the probability under the null hypothesis that the test statistic is at least as extreme as T_g. Under the null hypothesis, $Pr(p_g < \alpha) = \alpha$.

Statistical tests: Examples

- t-test: assumes normally distributed data in each class

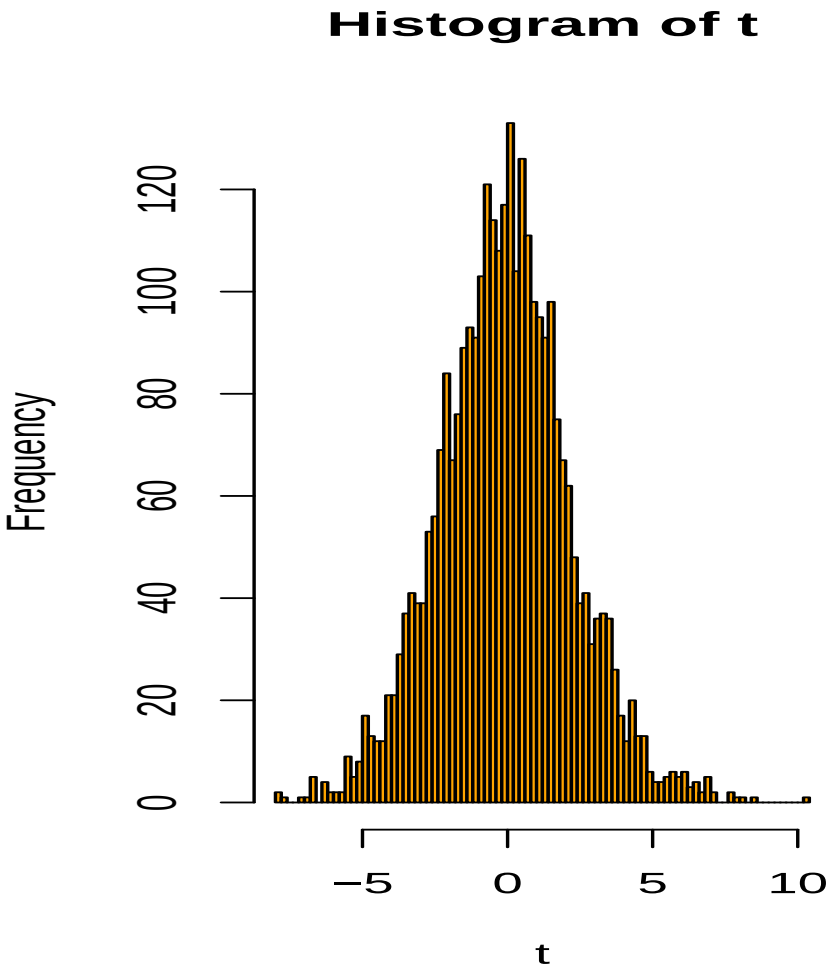
- Wilcoxon test: non–parametric, rank–based

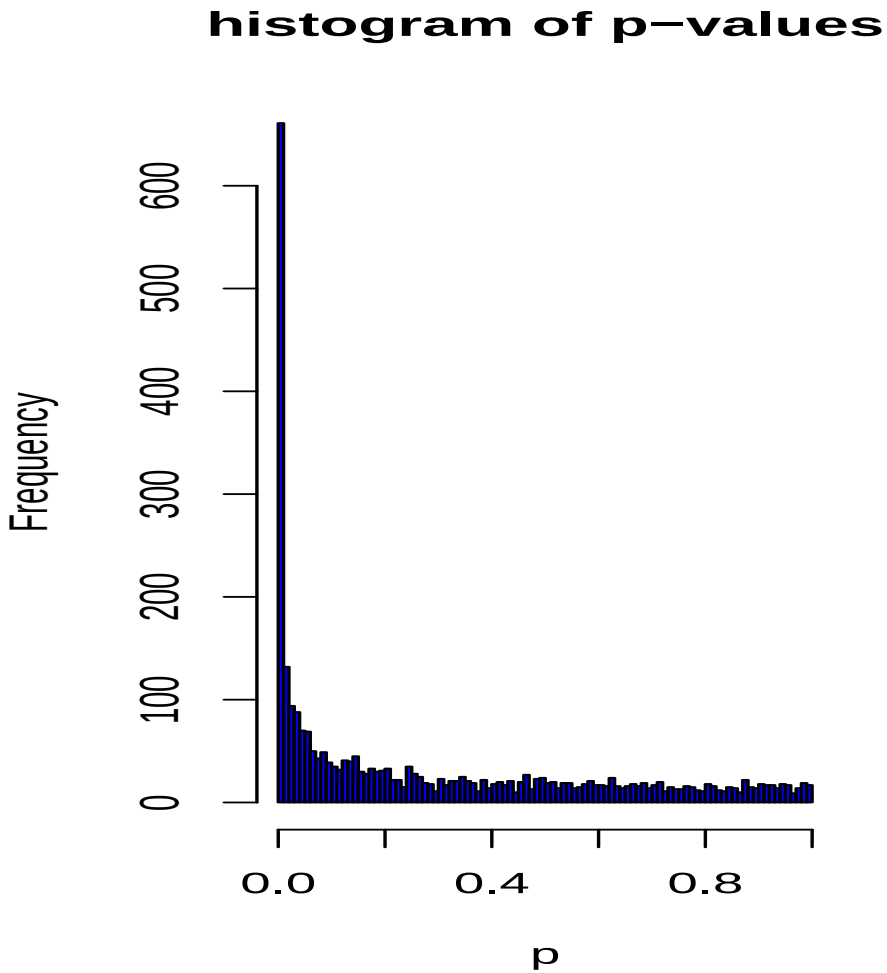
- permutation test: estimate the distribution of the test statistic (e.g., the t-statistic) under the null hypothesis by permutations of the sample labels:
The p–value p_g is given as the fraction of permutations yielding a test statistic that is at least as extreme as the observed one.

Perform statistical tests on normalized data; often a $\log$– or arsinh–transformation is advisable.

Example

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.





t-test: 1045 genes with $p < 0.05$.

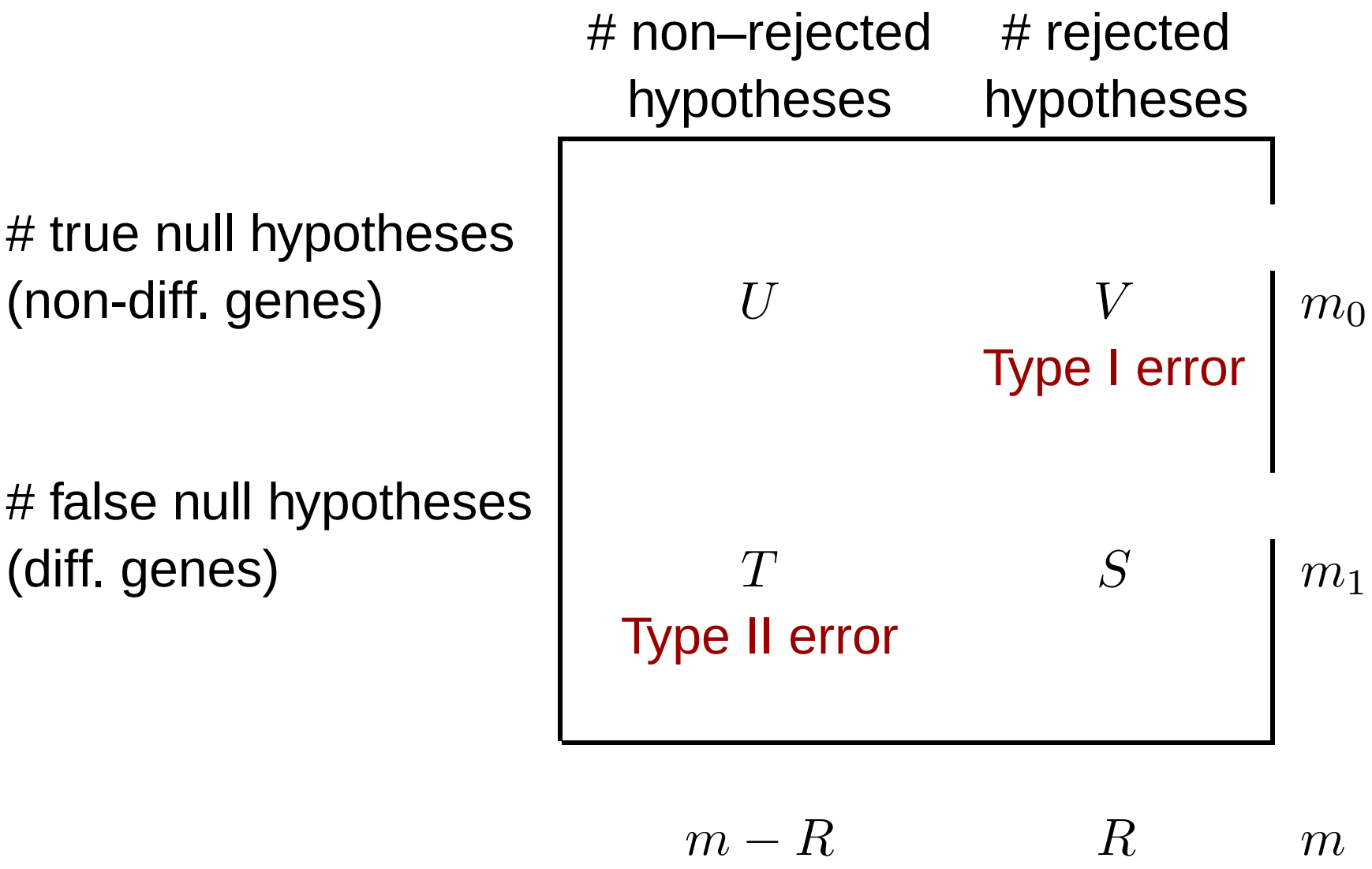
Multiple testing: the problem

Multiplicity problem: thousands of hypotheses are tested simultaneously.

- Increased chance of false positives.
- E.g. suppose you have 10,000 genes on a chip and not a single one is differentially expressed. You would expect $10000 * 0.01 = 100$ of them to have a p-value < 0.01.
- Individual p–values of e.g. 0.01 no longer correspond to significant findings.

Need to **adjust for multiple testing** when assessing the statistical significance of findings.

Multiple hypothesis testing



From Benjamini & Hochberg (1995).

Type I error rates

1. **Family–wise error rate (FWER)**. The FWER is defined as the probability of at least one Type I error (false positive):

$$FWER = Pr(V > 0).$$

2. **False discovery rate (FDR)**. The FDR (Benjamini & Hochberg 1995) is the expected proportion of Type I errors among the rejected hypotheses:

$$FDR = E(Q),$$

with

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

Multiple testing: Controlling a type I error rate

❍ Aim: For a given type I error rate α, use a procedure to select a set of "significant" genes that guarantees a type I error rate $\leq \alpha$.

FWER: The Bonferroni correction

Suppose we conduct a hypothesis test for each gene $g = 1, \ldots, m$, producing

- an observed test statistic: T_g
- an unadjusted p–value: p_g.

Bonferroni adjusted p–values:

$$\tilde{p}_g = \min\bigl(mp_g, 1\bigr).$$

FWER: The Bonferroni correction

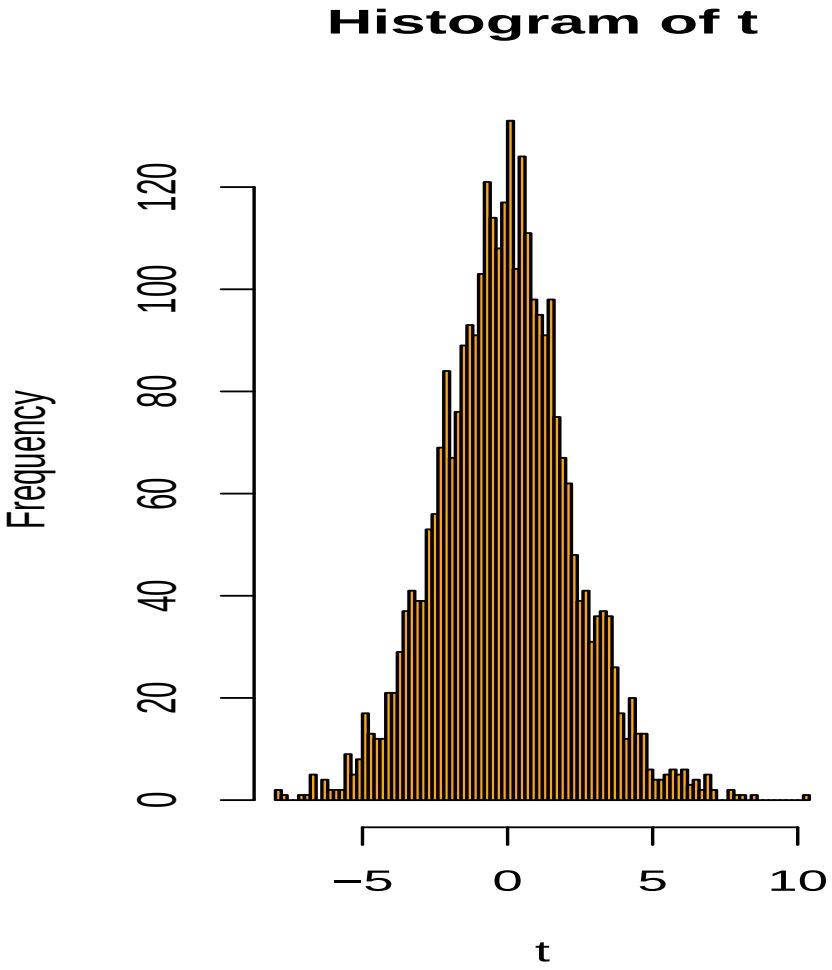
Choosing all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α. Under the complete null hypothesis H_0 that no gene is differentially expressed, we have:

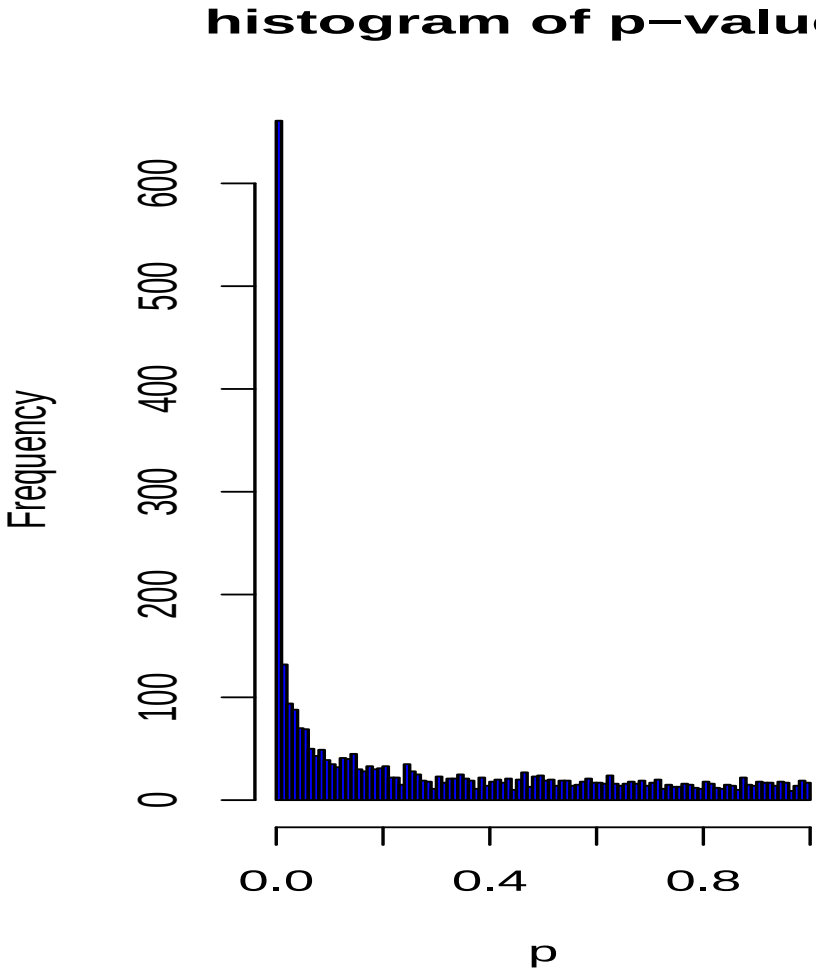
$$
\begin{aligned}
FWER = Pr(V > 0|H_0) = & \quad Pr(\text{at least one } \tilde{p}_g \leq \alpha|H_0) \\
= & \quad Pr(\text{at least one } p_g \leq \alpha/m|H_0) \\
\leq & \quad \sum_{g=1}^{m} Pr(p_g \leq \alpha/m|H_0) \\
= & \quad m * \alpha/m = \alpha
\end{aligned}
$$

(analogously for other configurations of hypotheses).

Example

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.





98 genes with Bonferroni-adjusted $\tilde{p}_g < 0.05 \Leftrightarrow p_g < 0.000016$ (t-test)

More is not always better

❍ Suppose you produce a small array with 500 genes you are particularly interested in.

❍ If a gene on this array has an unadjusted p-value of 0.0001, the Bonferroni-adjusted p-value is still 0.05.

❍ If instead you use a genome-wide array with, say, 50,000 genes, this gene would be much harder to detect, because roughly 5 genes can be expected to have such a low p-value by chance.

FWER: Improvements to Bonferroni (Westfall/Young)

- The minP adjusted p-values (Westfall and Young):
- $\tilde{p}_g = Pr(\min_{k=1,\ldots,m} P_k \leq p_g | H_0)$.
- Choosing all genes with $\tilde{p}_g \leq \alpha \Leftrightarrow p_g \leq c_\alpha$ controls the FWER at level α.
- But how to obtain the probabilities $\tilde{p}_g$?

Estimation of minP-adjusted p-values through resampling

- For $b = 1, \ldots, B$, (randomly) permute the sample labels.
- For each gene, compute the unadjusted p-values p_{gb} based on the permuted sample labels.
- Estimate $\tilde{p}_g = Pr(\min_{k=1,\ldots,m} P_k \leq p_g | H_0)$ by

$$\#\{b : \min_g p_{gb} \leq p_g\}/B.$$

Example

❍ Suppose $p_{\min} = 0.0003$ (the minimal unadjusted p-value).

❍ Among the randomized data sets (permuted sample labels), count how often the minimal p-value is smaller than 0.0003. If this appears e.g. in 4% of all cases, $\tilde{p}_{min} = 0.04$.

Westfall/Young FWER control

❍ Advantage of Westfall/Young: The method takes the dependence structure between genes into account, which gives in many cases (positive dependence between genes) higher power.

❍ Step-down procedure (Holm): Enhancement for Bonferroni and Westfall/Young: same adjustment for the smallest p-value, successively smaller adjustment for larger ones.

Westfall/Young FWER control

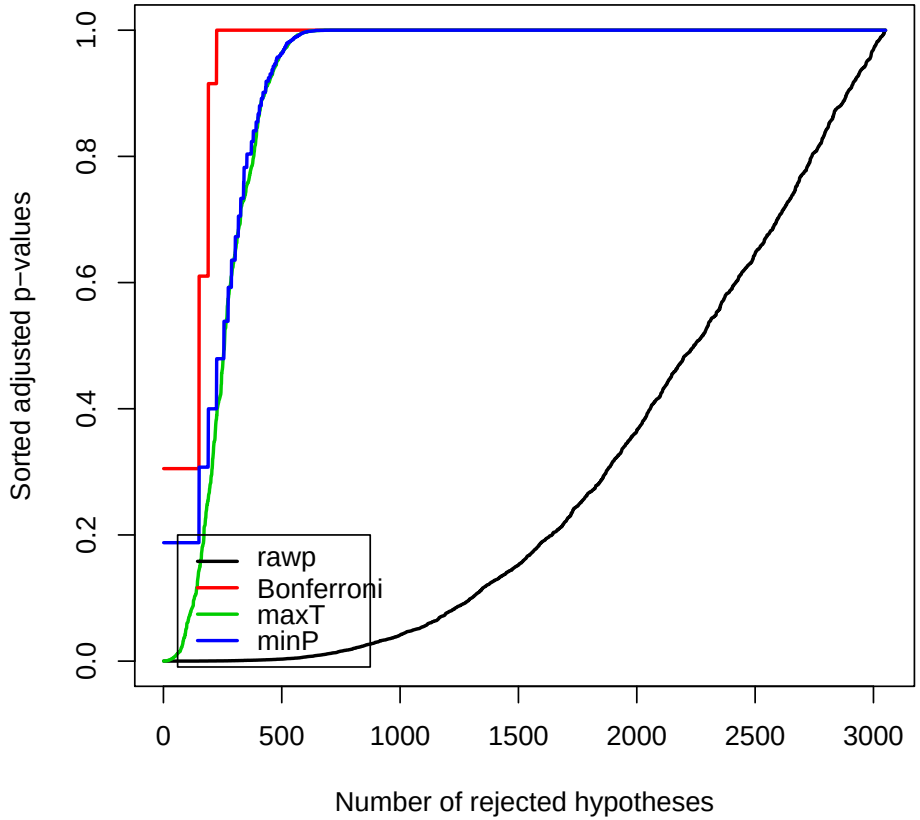
❍ Computationally intensive if the unadjusted p-values arise from permutation tests.

❍ Similar method (maxT) under the assumption that the statistics T_g are equally distributed under the null hypothesis - replace p_g by $|T_g|$ and $\min$ by $\max$. Computationally less intensive.

❍ All methods are implemented in the Bioconductor package multtest, with a fast algorithm for the minP method.

FWER: Comparison of different methods

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



Example taken from the multtest package in Bioconductor.

The FWER is a conservative criterion: many interesting genes may be missed.

Estimation of the FDR
(according to SAM and Storey 2001)

Idea: Depending on the chosen cutoff-value(s) for the test statistic T_g, estimate the expected proportion of false positives in the resulting gene list through a permutation scheme.

1. Estimate the number m_0 of non-diff. genes.

2. Compute the number of significant genes under permutations of the sample labels. The average of these numbers, multiplied with $\hat{m}_0/m$, gives an estimate of the expected number of false positives $E(V)$.

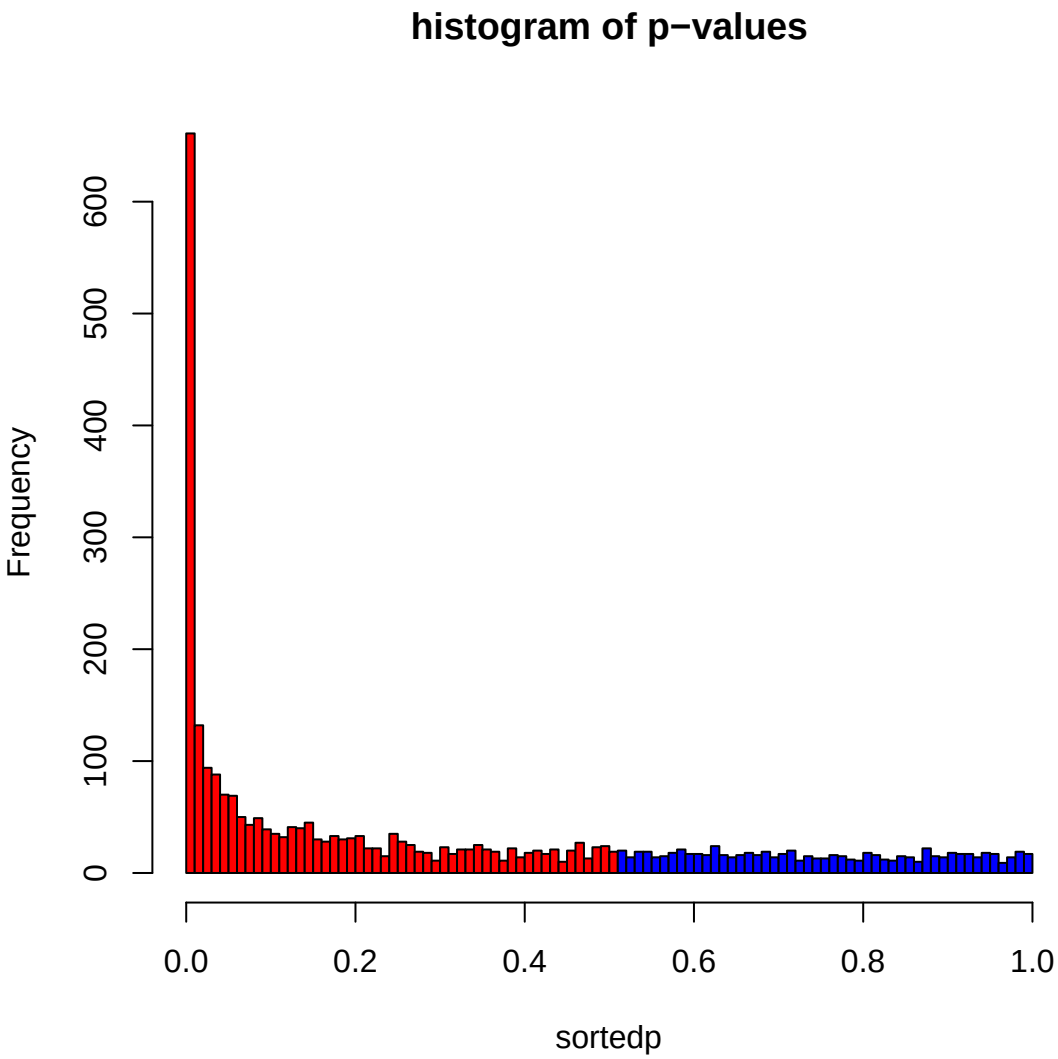
3. Estimate the FDR $E(V/R)$ by $\widehat{E(V)}/R$.

FDR - 1. Estimating the number m_0 of invariant genes

❍ Consider the distribution of p-values: A gene with $p > 0.5$ is likely to be not differentially expressed.

❍ As p-values of non-diff. genes should be uniformly distributed in $[0, 1]$, the number $2 * \#\{g|p_g > 0.5\}$ can be taken as an estimate of m_0.

❍ In the Golub example with 3051 genes, $\hat{m}_0 = 1592$.



2. Estimation of the FDR

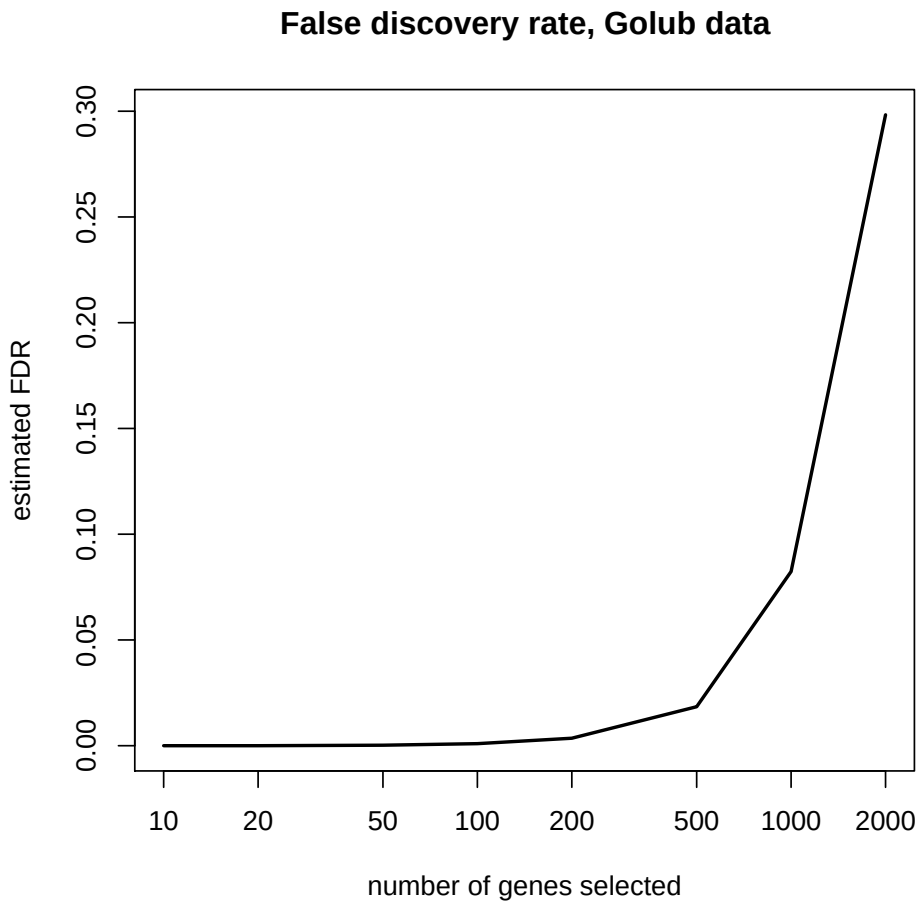
❍ For $b = 1, \ldots, B$, (randomly) permute the sample labels – this corresponds to the complete null hypothesis. Compute test statistics T_{gb} for each gene.

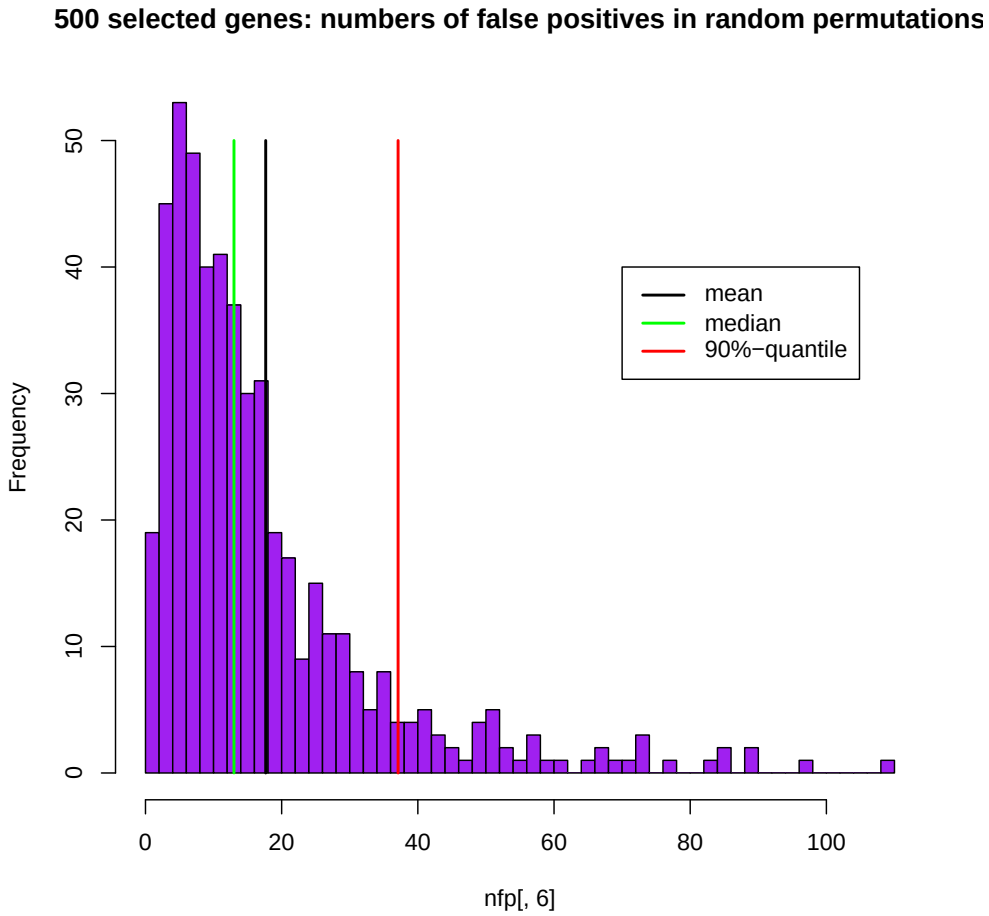
❍ For any threshold t_0 of the test statistic, compute the numbers V_b of genes with $T_{gb} > t_0$ (numbers of false positives).

❍ The estimation of the FDR is based on the mean of the V_b. However, a quantile of the V_b may also be interesting, because the actual proportion of false positives may be much larger than the mean value.

Estimation of the FDR: Example

Golub data





Estimation of the FDR

❍ The procedure takes the dependence structure between genes into account.

❍ The q-value of a gene is defined as the minimal FDR at which it appears significant.

FWER or FDR?

❍ Chose control of the FWER if high confidence in all selected genes is desired. Loss of power due to large number of tests: many differentially expressed genes may not appear as significant.

❍ If a certain proportion of false positives is tolerable: Procedures based on FDR are more flexible; the researcher can decide how many genes to select, based on practical considerations.

Prefiltering

- What about prefiltering genes (according to intensity, variance etc.) to reduce the proportion of false positives - e.g. genes with consistently low intensity may not be considered interesting?

- Can be useful, but:

- The criteria for filtering have to be chosen before the analysis - not dependent on the results of the analysis.

- The criteria have to be independent of the distribution of the test statistic under the null hypothesis - otherwise no control of the type I error.

References

- Y. Benjamini and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, Vol. 57, 289–300.
- Dudoit et al. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, Vol. 12, 111–139.
- J.D. Storey and R. Tibshirani (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Preprint, http://www.stat.berkeley.edu/ storey/
- V.G. Tusher et al. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, Vol. 98, 5116–5121.
- P.H. Westfall and S.S. Young (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley.