Differential gene expression

Anja von Heydebreck

anja.von.heydebreck@merck.de

Slides partly adapted from S. Dudoit, Bioconductor short course

Identifying differentially expressed genes

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

❍ Simple fold–change rules give no assessment of statistical significance (data come with biological and technical variability).

❍ Conduct a statistical test for each gene $g = 1, \ldots, m$, giving test statistics T_g.

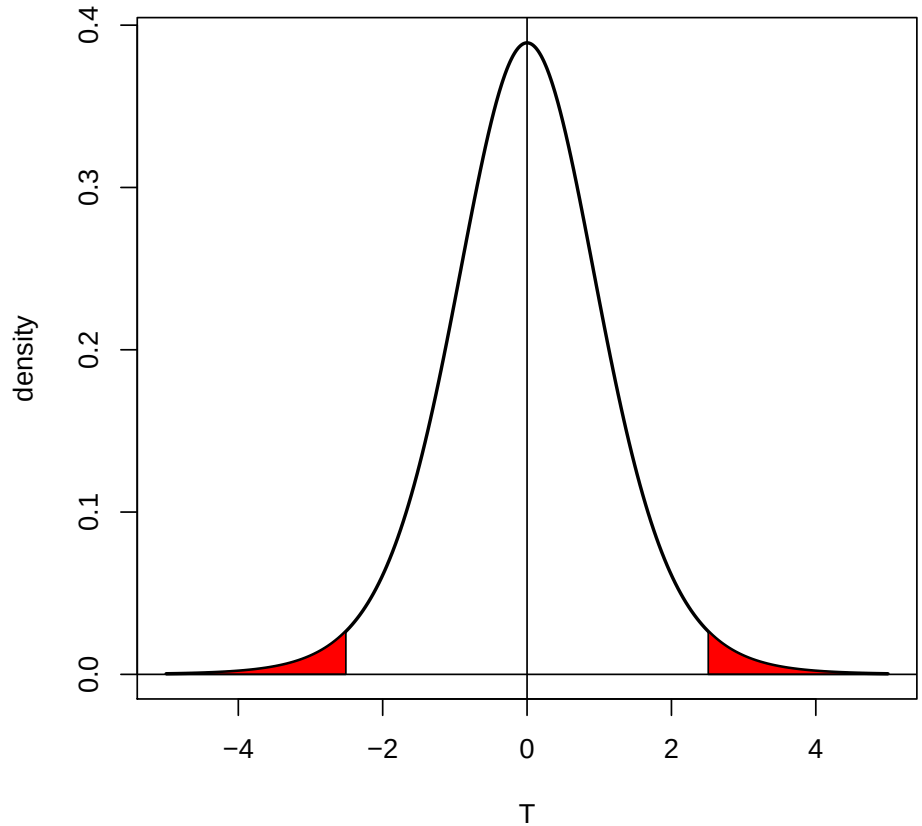
Statistical tests

❍ Example: The two–sample t–statistic

$$T_g = \frac{\bar{X}_{g1} - \bar{X}_{g2}}{s_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is used to test equality of the group means μ_1, μ_2.

❍ The p–value p_g is the probability under the null hypothesis (here: $\mu_1 = \mu_2$) that the test statistic is at least as extreme as the observed value T_g. Under the null hypothesis, $Pr(p_g < \alpha) = \alpha$.



Statistical tests: Examples

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

❍ standard t-test: assumes normally distributed data in each class (almost always questionable), equal variances within classes

❍ Welch t-test: as above, but allows for unequal variances

❍ 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.

Permutation tests

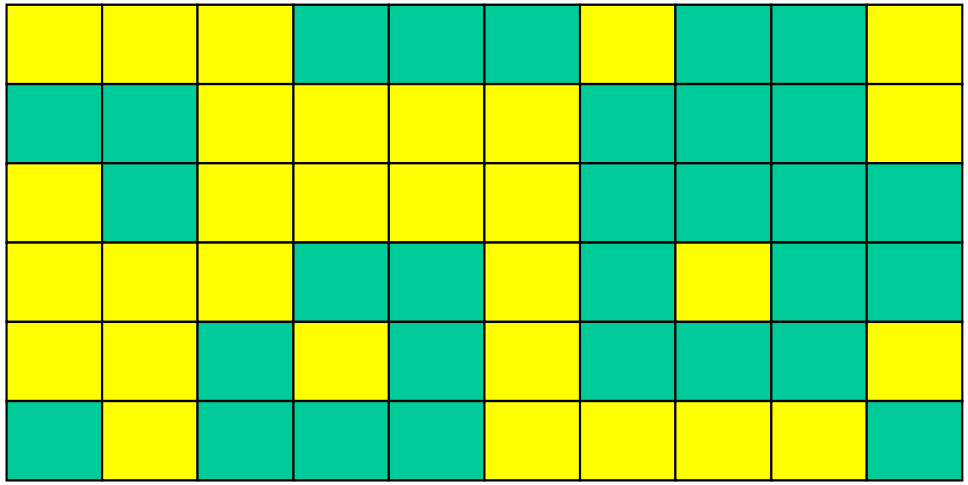
true class labels:

test statistic

2.2

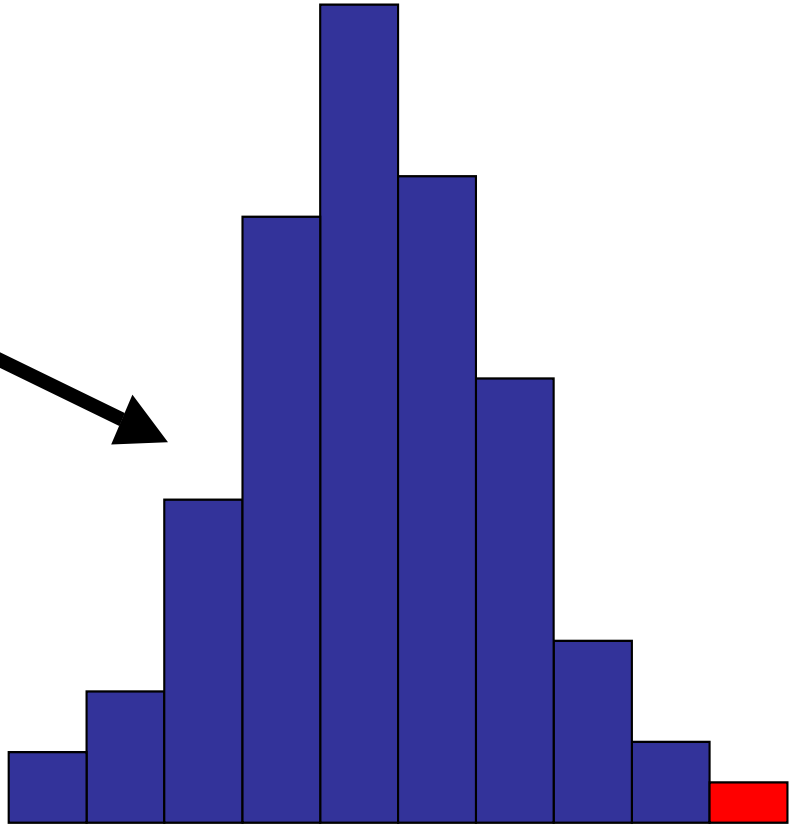
null distribution of test statistic

(random) permutations of class labels:



1.5
-0.4
2.3
0.7
0.2
-1.2

⋮



2.2

Statistical tests: Different settings

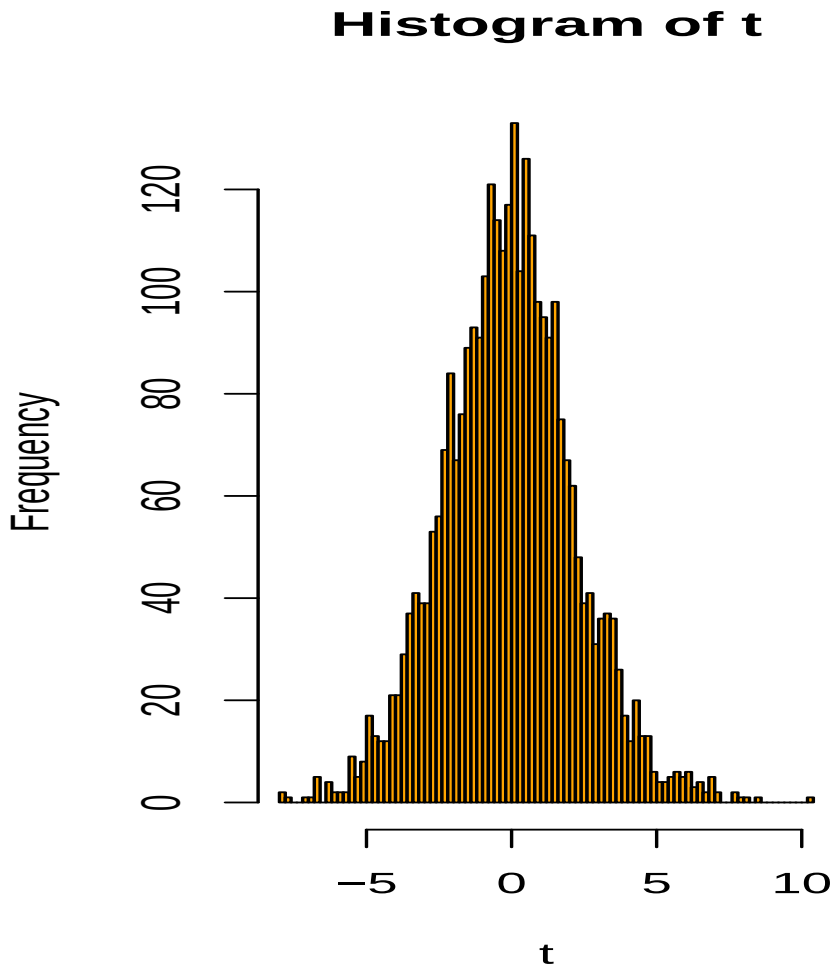
❍ comparison of two classes (e.g. tumor vs. normal), one class, paired observations from two classes: (permutation) t–test, Wilcoxon test

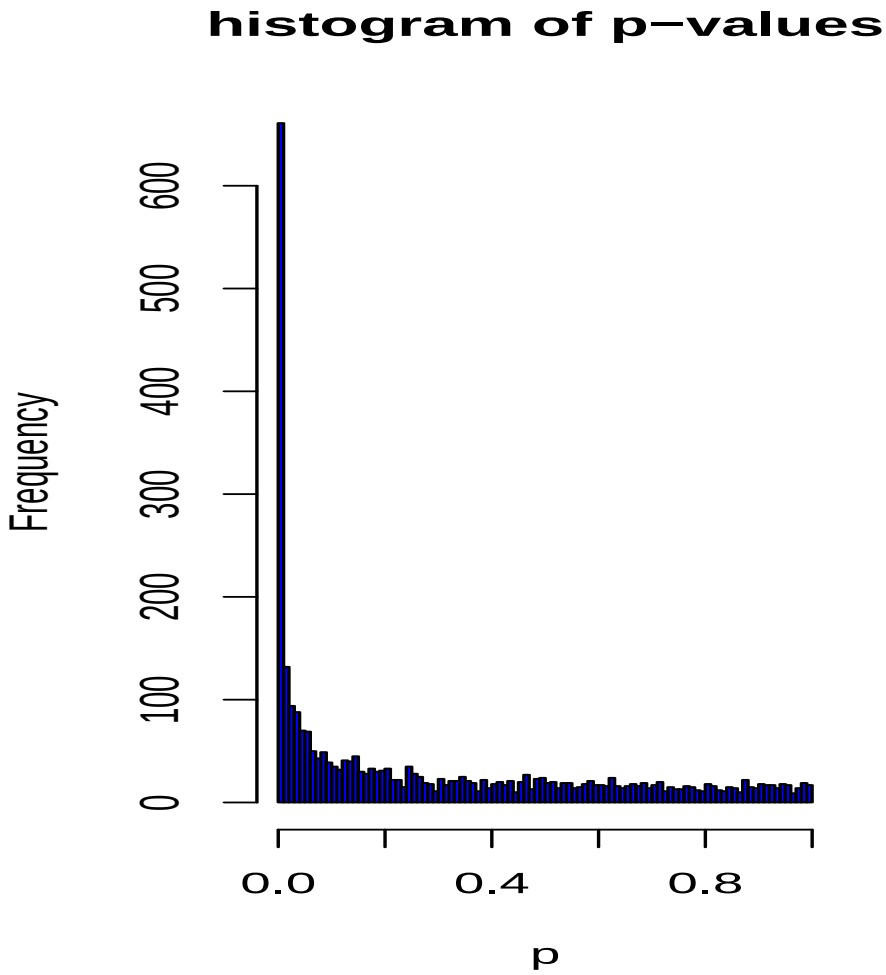
❍ more than two classes and/or more than one factor: tests may be based on ANOVA/linear models

❍ continuous response variable: linear models;
censored survival times: e.g. Cox proportional hazards models

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

	# non–rejected hypotheses	# rejected hypotheses	
# true null hypotheses (non-diff. genes)	U	V Type I error	m_0
# false null hypotheses (diff. genes)	T Type II error	S	m_1
	$m - R$	R	m

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.

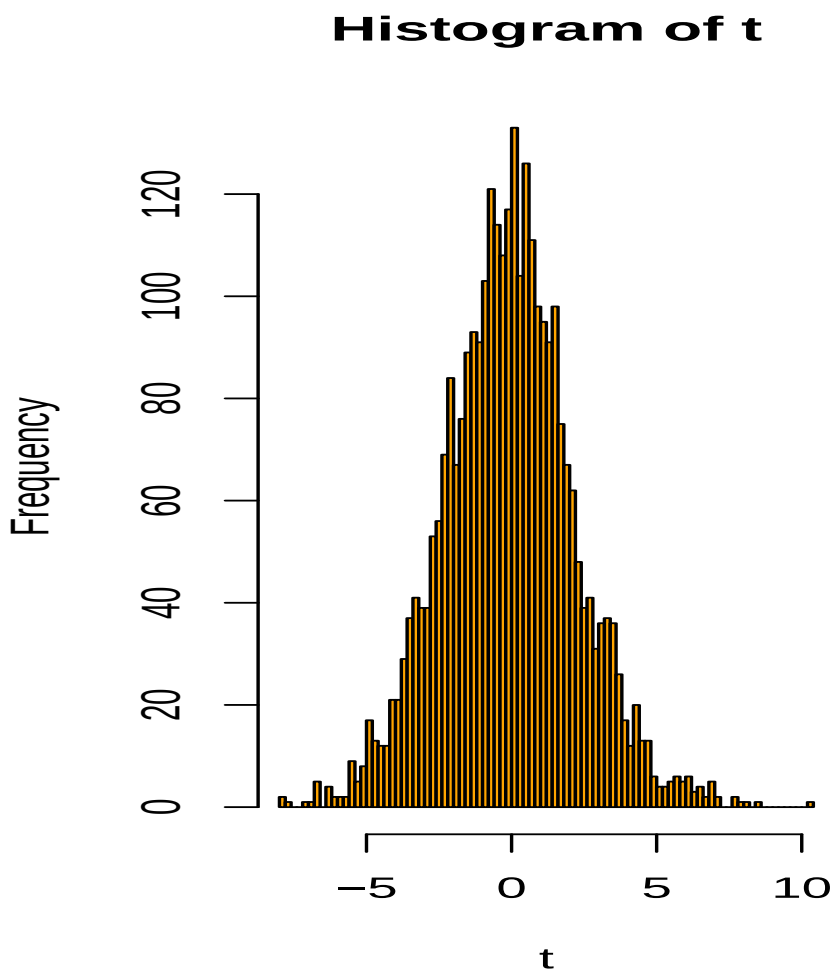
Bonferroni adjusted p–values:

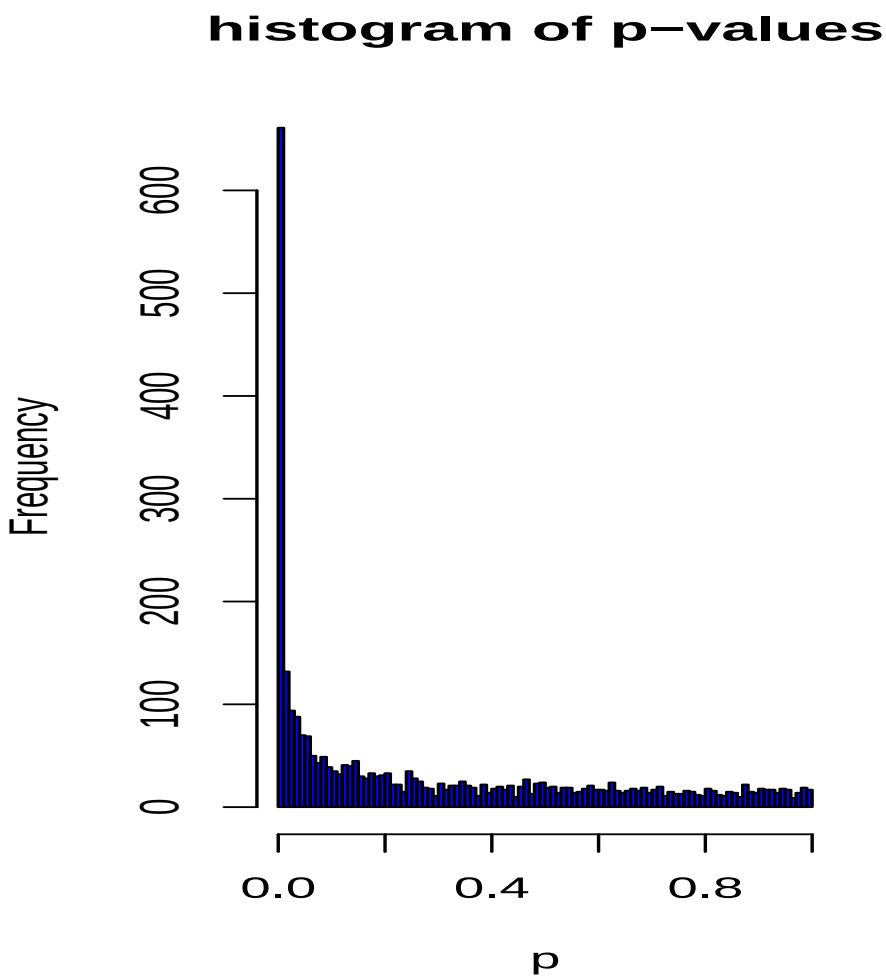
$$\tilde{p}_g = \min\big(mp_g, 1\big).$$

Choosing all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α.

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)

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 $\tilde{p}_g \leq \alpha$ controls the FWER at level α.
- The probabilities $\tilde{p}_g$ are estimated through a permutation scheme.

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.

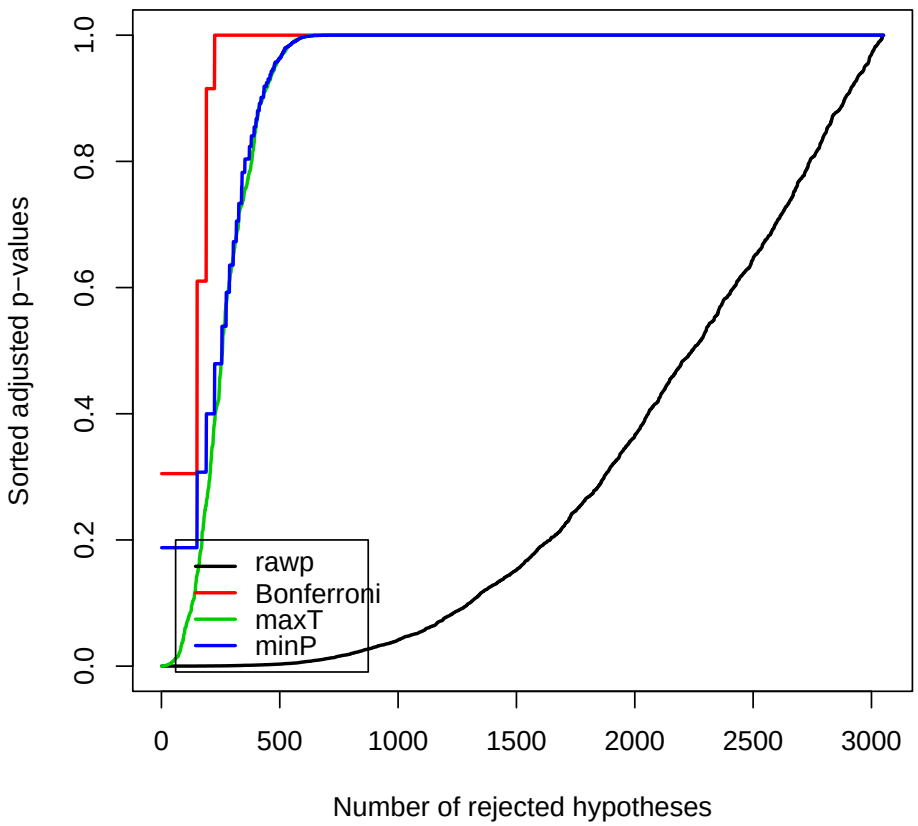
❍ Computationally intensive if the unadjusted p-values arise from permutation tests – two levels of permutations required.

❍ 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.

More is not always better

❍ Suppose you use 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.

❍ Therefore, it may be worthwile focusing on genes of particular biological interest from the beginning.

Controlling the FDR (Benjamini/Hochberg)

❍ Ordered unadjusted p–values: $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m}$.

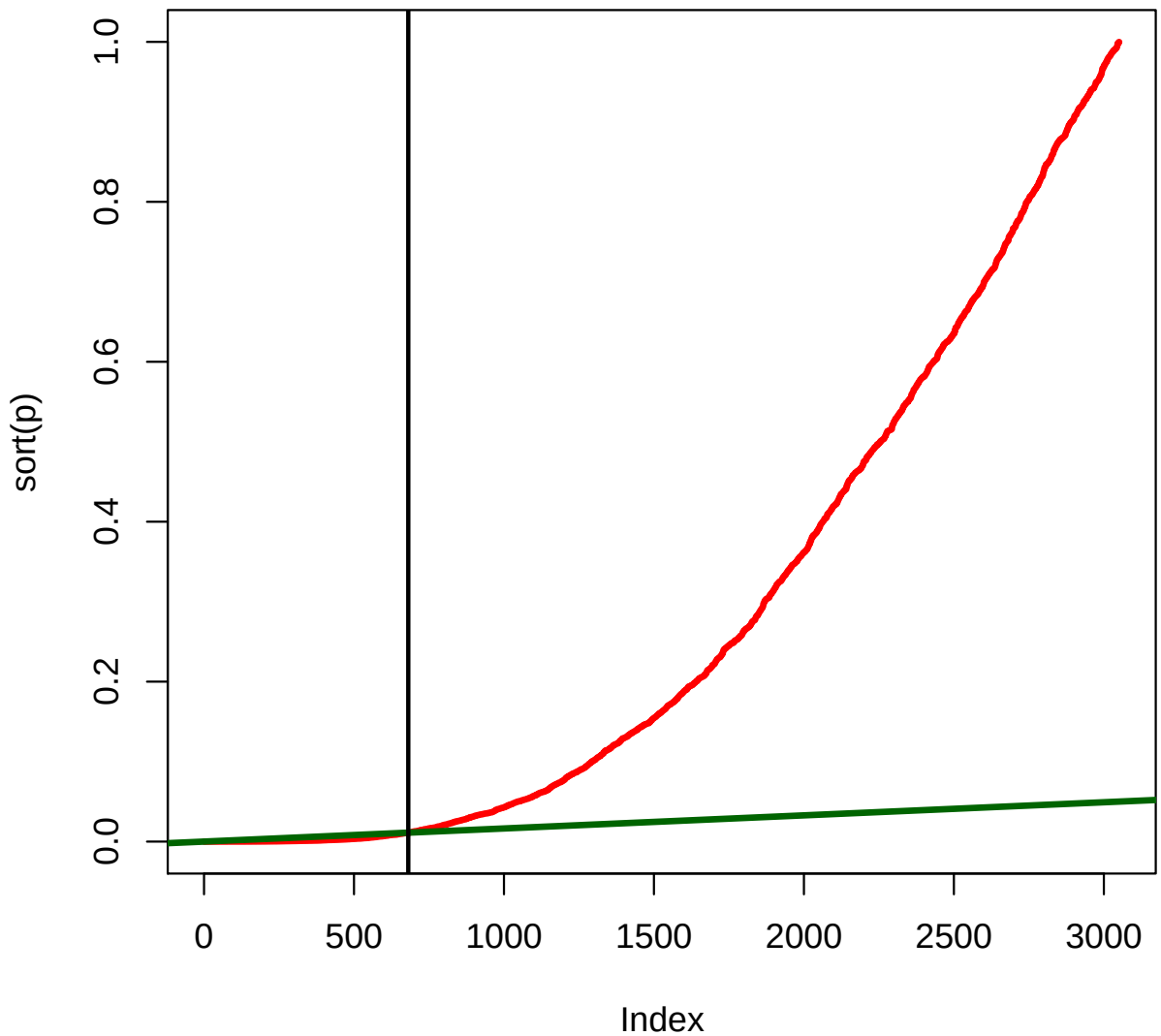
❍ To control $FDR = E(V/R)$ at level α, let

$$j^\star = \max\{j : p_{r_j} \leq (j/m)\alpha\}.$$

Reject the hypotheses H_{r_j} for $j = 1, \ldots, j^\star$.

❍ Works for independent test statistics and for some types of dependence. Tends to be conservative if many genes are differentially expressed. Implemented in multtest.

Controlling the FDR (Benjamini/Hochberg)



Golub data: 681 genes with BH–adjusted $p < 0.05$.

Estimation of the FDR (SAM, Storey/Tibshirani 2003)

Idea: Depending on the chosen cutoff-value 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. Estimate the expected number of false positives under the complete null hypothesis, $E(V_0)$, through resampling.
   Then, $\widehat{E(V)} = \frac{\hat{m}_0}{m}\widehat{E(V_0)}$ (because only the non-diff. genes may yield false positives).

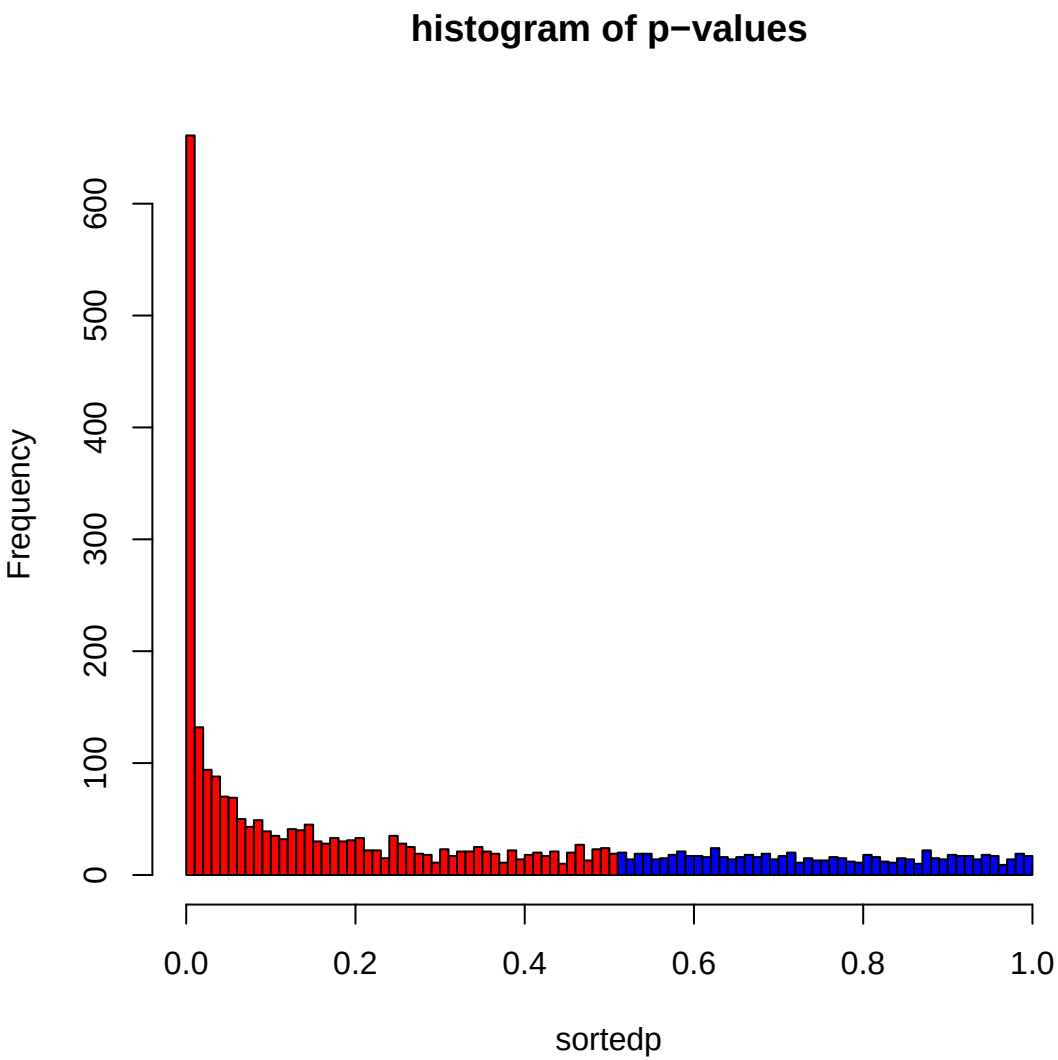
3. Estimate $FDR = E(V/R)$ by $\widehat{E(V)}/R$.

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

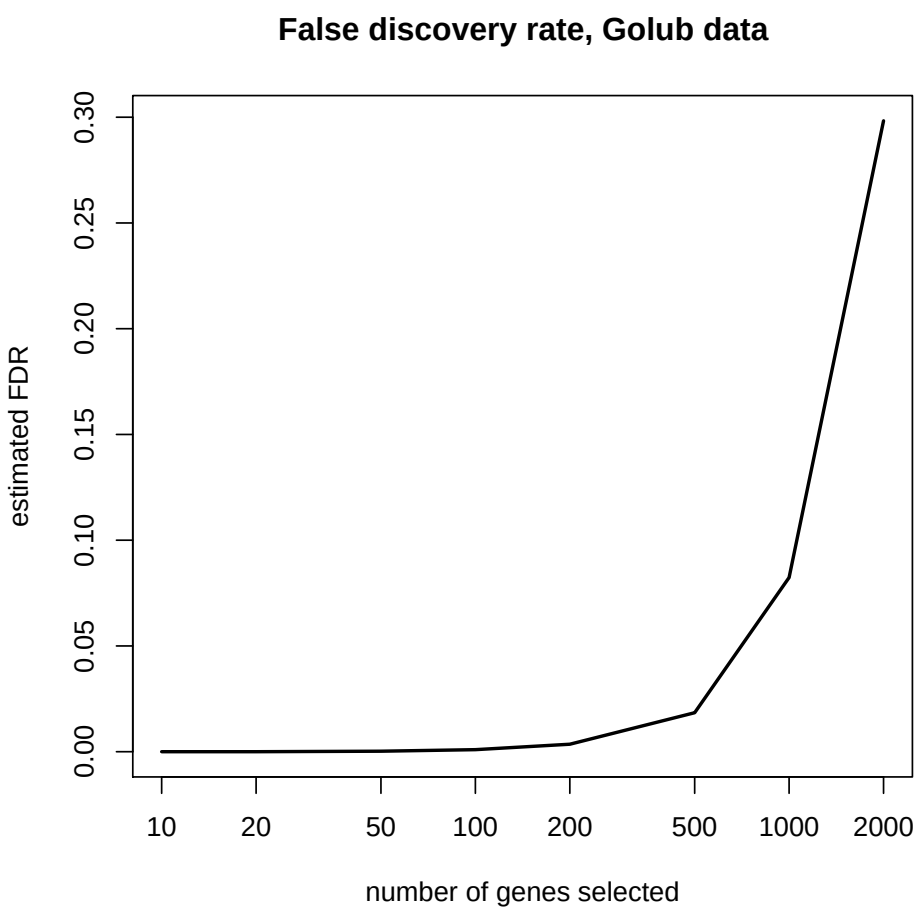
❍ For $b = 1, \ldots, B$, (randomly) permute the sample labels, compute test statistics T_{gb} corresponding to the complete null hypothesis.

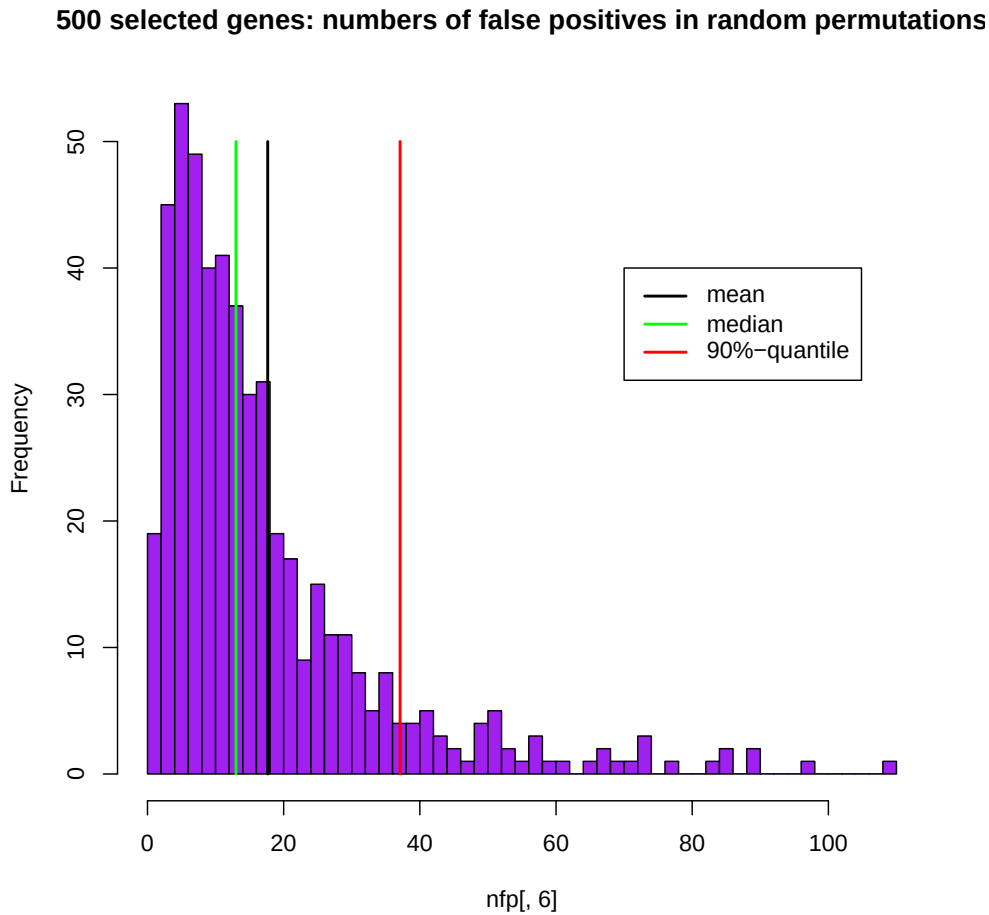
❍ For any threshold t_0 of the test statistic, compute the numbers V_b of genes with $T_{gb} > t_0$ (numbers of false positives).

❍ 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, as the actual proportion of false positives may be much larger than the mean.

Estimation of the FDR: Example

Golub data



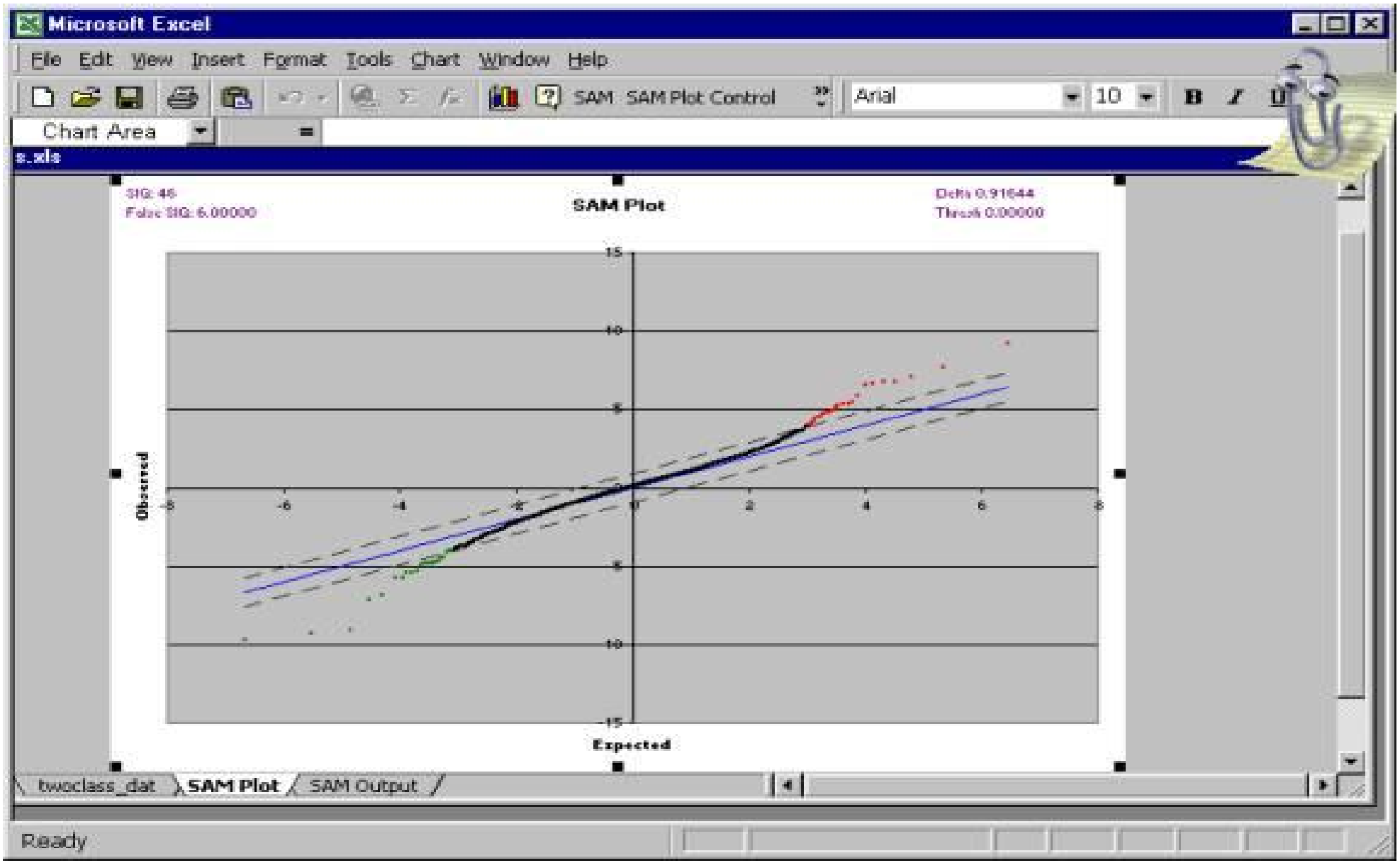


Estimation of the FDR

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

❍ In SAM, the q-value of a gene is defined as the minimal estimated FDR at which it appears significant.

The SAM plot


Microsoft Excel
File Edit View Insert Format Tools Chart Window Help
SAM SAM Plot Control
Arial
10
Chart Area
s.xls
SIG: 46
False SIG: 6.00000
SAM Plot
Delta 0.91644
Thresh 0.00000
15
10
5
-5
-10
-15
-6
-4
-2
2
4
6
Observed
Expected
twoclass_dat
SAM Plot
SAM Output
Ready

FWER or FDR?

❍ Choose 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 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.

Few replicates – moderated t–statistics

❍ When doing a t–test, we estimate the variance of each gene individually. This is fine if we have enough replicates, but with few replicates (say 2–5 per group), these variance estimates are highly variable.

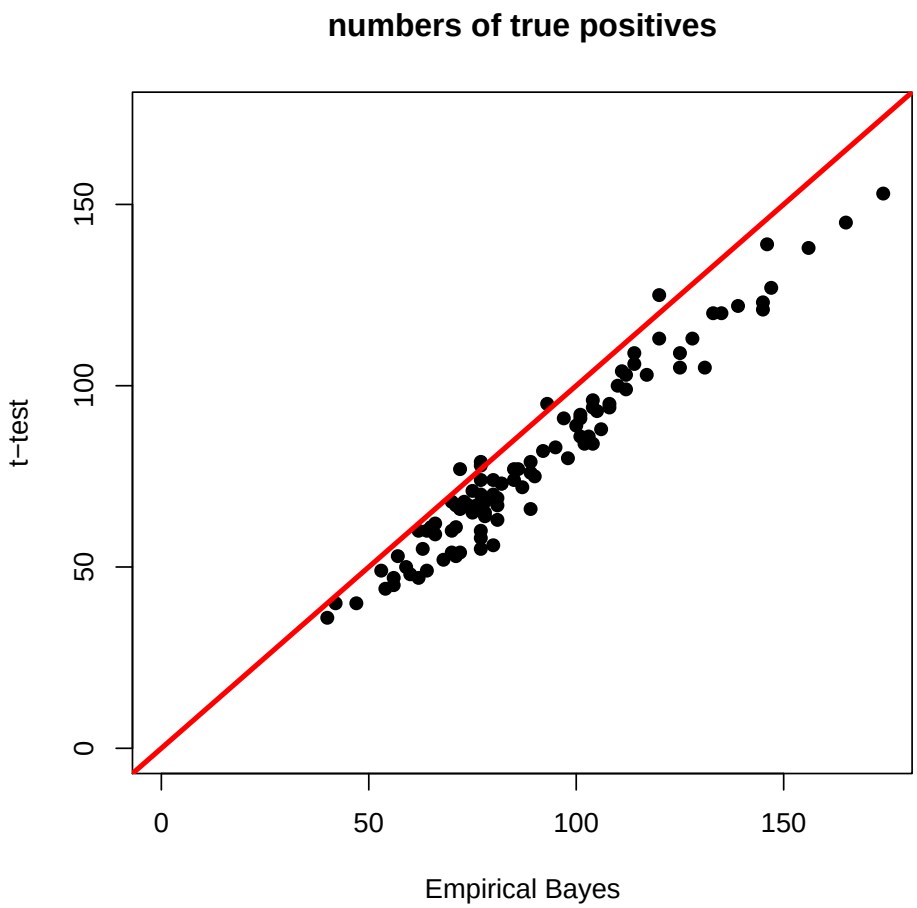
❍ In a moderated t–statistic, the estimated gene–specific variance s_g^2 is replaced by a weighted average of s_g^2 and s_0^2, which is a global variance estimator obtained from pooling all genes.

❍ This gives an interpolation between the t–test and a fold–change criterion.

❍ Examples: packages limma, siggenes in Bioconductor, SAM.

Moderated t–statistic

Repeatedly draw 4 ALL and 4 AML samples and apply the usual and moderated t–test (Bioconductor package limma) to them. Using a cut–off of $p < 0.05$, "true positives" are defined on the basis of the analysis of the whole data set (681 genes with FDR < 0.05).



Prefiltering

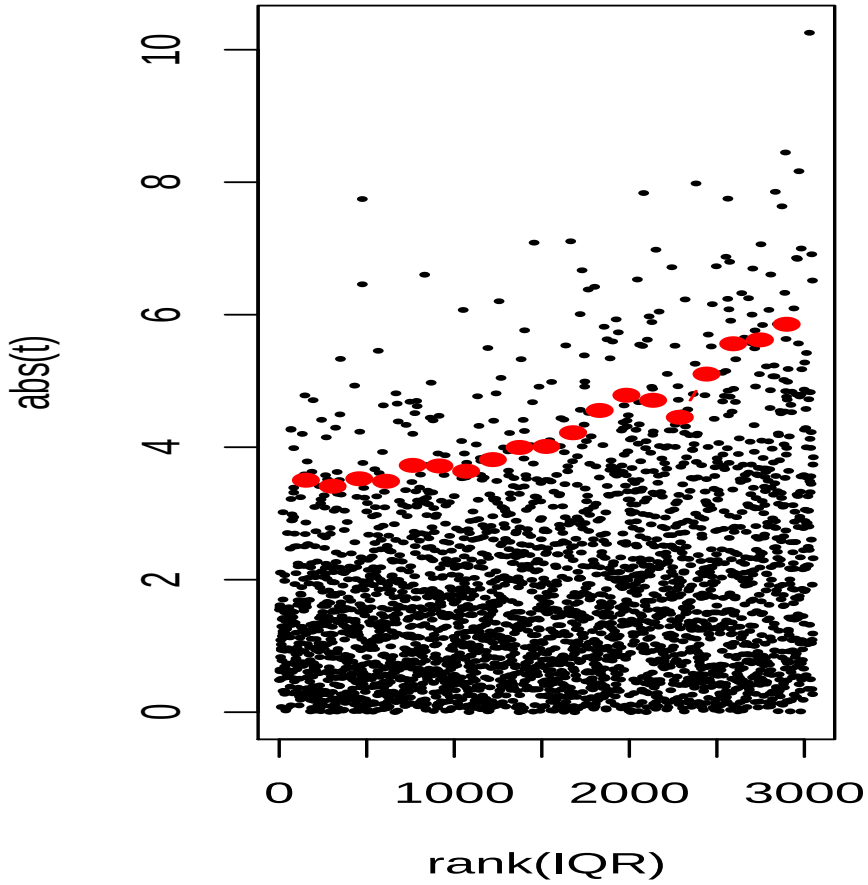
❍ What about prefiltering genes (according to intensity, variance etc.) to reduce the proportion of false positives?

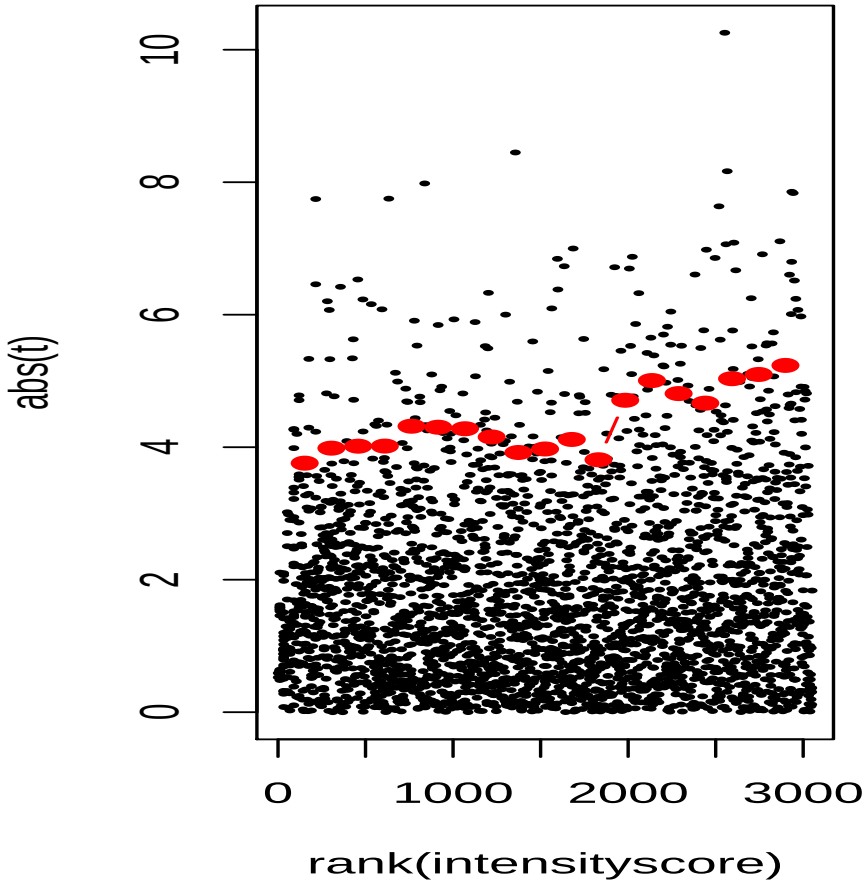
❍ Can be useful: Genes with low intensities in most of the samples or low variance across the samples are less likely to be interesting.

❍ In order to maintain control of the type I error, the criteria have to be independent of the distribution of the test statistic under the null hypothesis.

Prefiltering by intensity and variability

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





Ranks of interquartile range and 75%–quantile of intensities versus absolute t–statistic.

What else?

❍ Statistical tests rely on independent observations. For example, if you have 6 biological samples with 2 replicate hybridizations each, a t–test based on all 12 observations is not appropriate. Here, one may either i) average over the technical replicates or ii) use special methods (mixed effects models, see e.g. Bioconductor package limma for the case of duplicate spots).

❍ The Bioconductor package globaltest by J. Goeman provides a test whether a group of genes (e.g. a GO category) contains any differentially expressed genes.

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.

❍ S. Dudoit, J.P. Shaffer, J.C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, Vol. 18, 71–103.

❍ J.D. Storey and R. Tibshirani (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: *The analysis of gene expression data: methods and software.* Edited by G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger. Springer, New York.

❍ 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.