Statistical tests for differential gene expression

Anja von Heydebreck

Statistical tests

- Suppose we want to find genes that are differentially expressed between different conditions/phenotypes, e.g. two different tumor types.
- We conduct a statistical test for each gene $g = 1, \ldots, m$ (t-test, Wilcoxon test, permutation test, ...).
- 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, $Pr(p_g < \alpha) = \alpha$.
- A low p–value is seen as evidence that the null hypothesis may not be true (i.e., our gene is differentially expressed).

Standard t–test

Assume $X_1, \ldots, X_m$ are from a $N(\mu_1, \sigma^2)$–distribution, and $Y_1, \ldots, Y_n$ from a $N(\mu_2, \sigma^2)$–distribution (equal variances). Compute the pooled variance estimate as

$$s^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right).$$

The t–statistic is given as

$$T(X, Y) = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Under the null hypothesis $\mu_1 = \mu_2$, the t–statistic $T(X, Y)$ follows a t_{m+n-2}–distribution.

Welch t–test

Here we allow for different variances in the two groups. We assume the observations follow distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. The Welch t-statistic for the test of the null hypothesis $\mu_1 = \mu_2$ is defined as

$$T(X, Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}}.$$

Under the null hypothesis $\mu_1 = \mu_2$, it approximately follows a t-distribution with

$$\nu = \frac{(\frac{s_1^2}{m} + \frac{s_2^2}{n})^2}{\frac{(\frac{s_1^2}{m})^2}{m-1} + \frac{(\frac{s_2^2}{n})^2}{n-1}}$$

degrees of freedom.

The Welch t–test is the default in the R function t.test.

Wilcoxon test (Mann–Whitney test, U–test)

- Non–parametric test for equality of two distributions.
- Compute the ranks of observations in the pooled sample.

Observations:	0.3	0.5	0.8	0.9	1.3	2.4
Ranks:	1	2	3	4	5	6
Groups:	1	1	1	2	2	2

- The test statistic is a function of the sum of ranks in group 1; here, $R_1 = 6$.
- For small sample sizes, the null distribution of the test statistic can be computed exactly. For large sample size, a normal approximation is used.
- Advantage: Non–parametric, robust against outliers.

Permutation tests

- Want to test whether observations in two groups follow the same distribution, without making assumptions concerning the distribution (e.g. normality).

- Compute a test statistic T_{obs} for your data (e.g. the t–statistic).

- For $b = 1, \ldots, B$, do
  1. Permute the group labels, giving a new assignment of the observations to the two groups.
  2. Compute the test statistic T_b for the new group assignment.

- This is done either for all possible permutations, or, if this set is too large, for a large number B of random permutations.

- The permutation p–value is given as

$$p = \#\{b : |T_b| \geq |T_{obs}|\}/B.$$

Statistical tests: Different settings

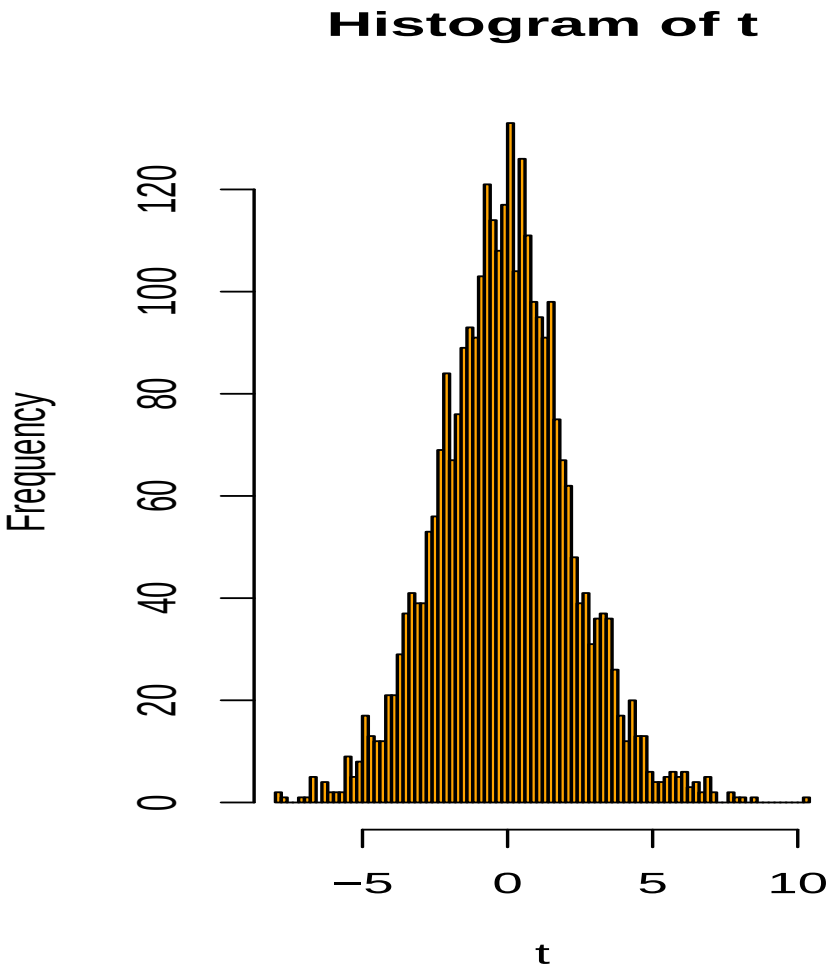
- comparison of two classes (e.g. tumor vs. normal), one class, paired observations from two classes: (permutation) t–test, Wilcoxon test

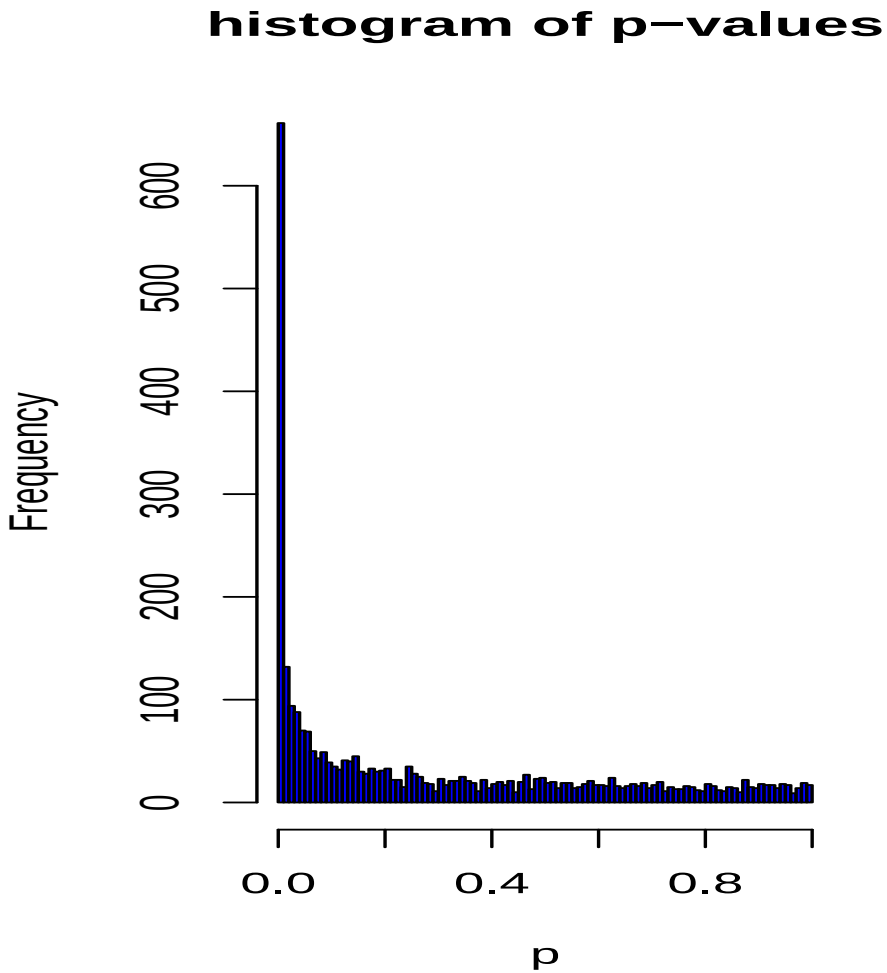
- more than two classes and/or more than one factor: tests may be based on ANOVA/linear models

- 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 $\leq \alpha$.

- The type I error is defined with respect to a given configuration of true and false null hypotheses.

- Weak control of type I error: only under the assumption that all null hypotheses are true (*complete null hypothesis*, H_0^C).

- Strong control of type I error: for all possible configurations of true and false null hypotheses.

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

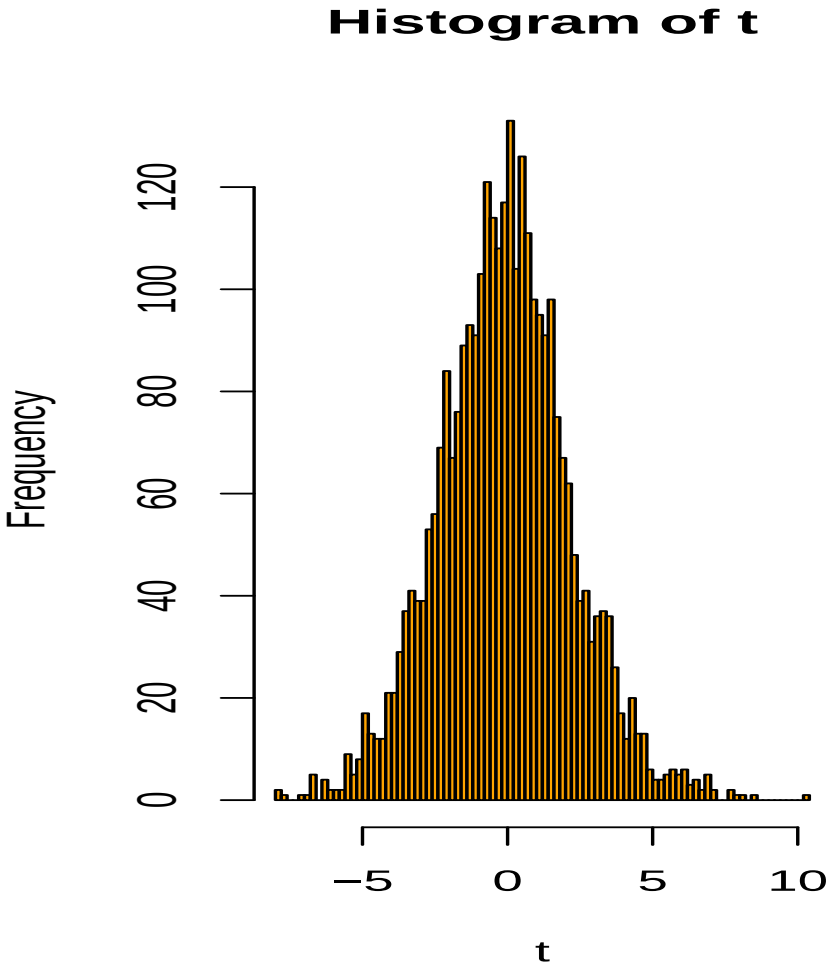
Choosing all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α. Under the complete null hypothesis H_0^C that no gene is differentially expressed, we have:

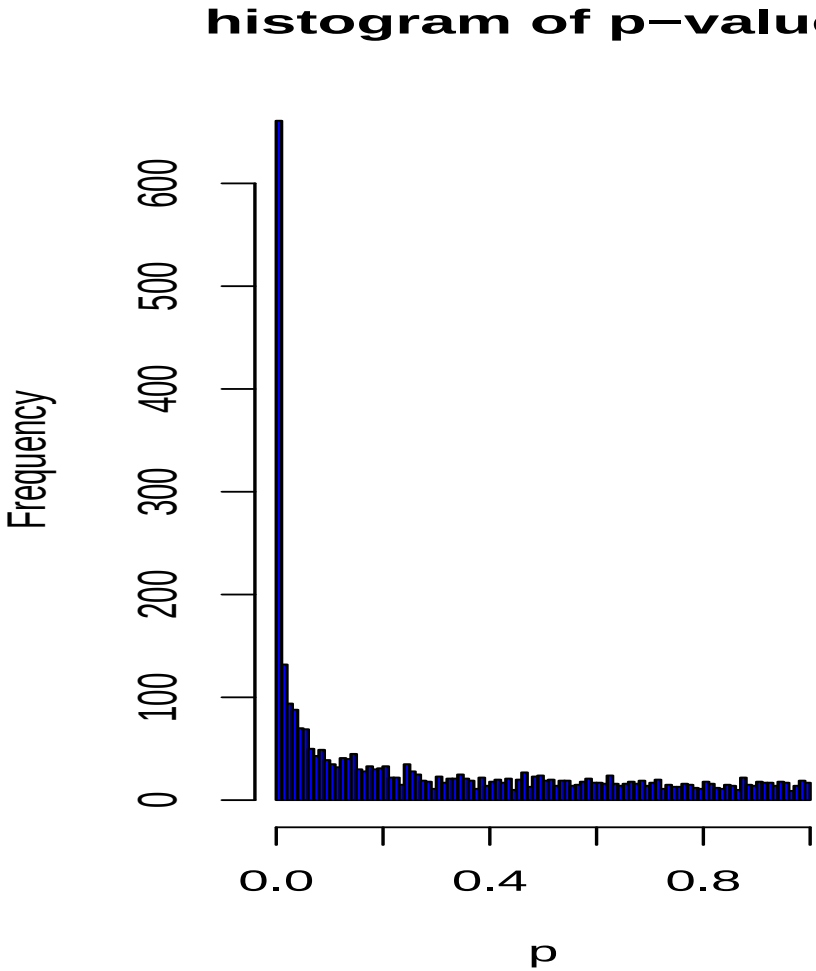
$$\begin{aligned} FWER = Pr(V > 0) = & \quad Pr(\text{at least one } \tilde{p}_g \leq \alpha | H_0^C) \\ = & \quad Pr(\text{at least one } p_g \leq \alpha/m | H_0^C) \\ \leq & \quad \sum_{g=1}^{m} Pr(p_g \leq \alpha/m | H_0^C) \\ = & \quad m * \alpha/m = \alpha. \end{aligned}$$

(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)

FWER: Step–down procedures (Holm)

- Modified Bonferroni correction. Same adjustment for the smallest p–value, successively smaller adjustment for the following ones:

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

- To control the FWER at level α, let

$$j^\star = \min\{j : p_{r_j} > \alpha/(m-j+1)\}.$$

Reject the hypotheses H_{r_j} for $j = 1, \ldots, j^\star - 1$. Thus, the adjusted p–values are

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \{\min((m-k+1)p_{r_k}, 1)\}$$

FWER: Improvements to Bonferroni (Westfall/Young)

- The minP adjusted p-values (Westfall and Young):
- $\tilde{p}_g = Pr(\min_{k=1,\ldots,m} P_k \leq p_g | H_0^C)$.
- Choosing all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α.
- Consider the complete null hypothesis H_0^C:

$$\begin{aligned} FWER = Pr(V > 0) = & \quad Pr(\text{at least one } \tilde{p}_g \leq \alpha | H_0^C) \\ = & \quad Pr(\min \tilde{p}_g \leq \alpha | H_0^C) \\ = & \quad Pr(\min p_g \leq c_\alpha | H_0^C) \\ = & \quad \alpha. \end{aligned}$$

- 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 $\tilde{p}_g = Pr(\min_{k=1,\ldots,m} P_k \leq p_g | H_0^C)$ by

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

Westfall/Young: 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): same adjustment for the smallest p-value, successively smaller adjustment for larger ones:

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \{ \min_{l \in \{r_k,\ldots,r_m\}} P_l \leq p_{r_k} | H_0^C \}$$

Westfall/Young FWER control

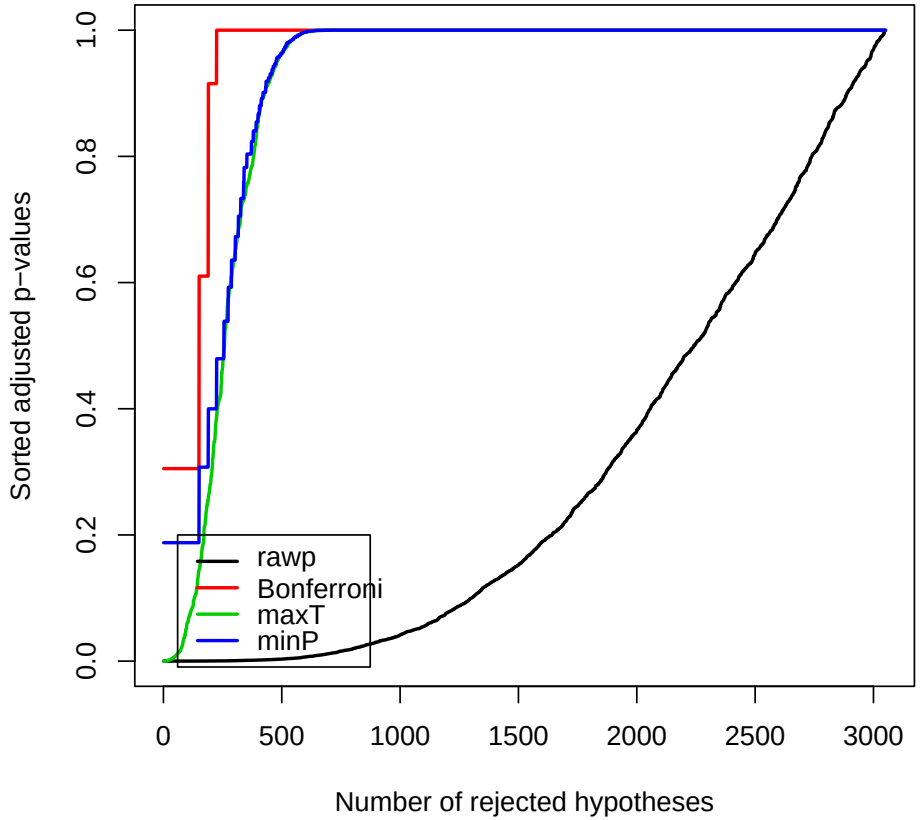
- Computationally intensive if the unadjusted p-values arise from permutation tests.

- 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.

Controlling the FDR (Benjamini/Hochberg)

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

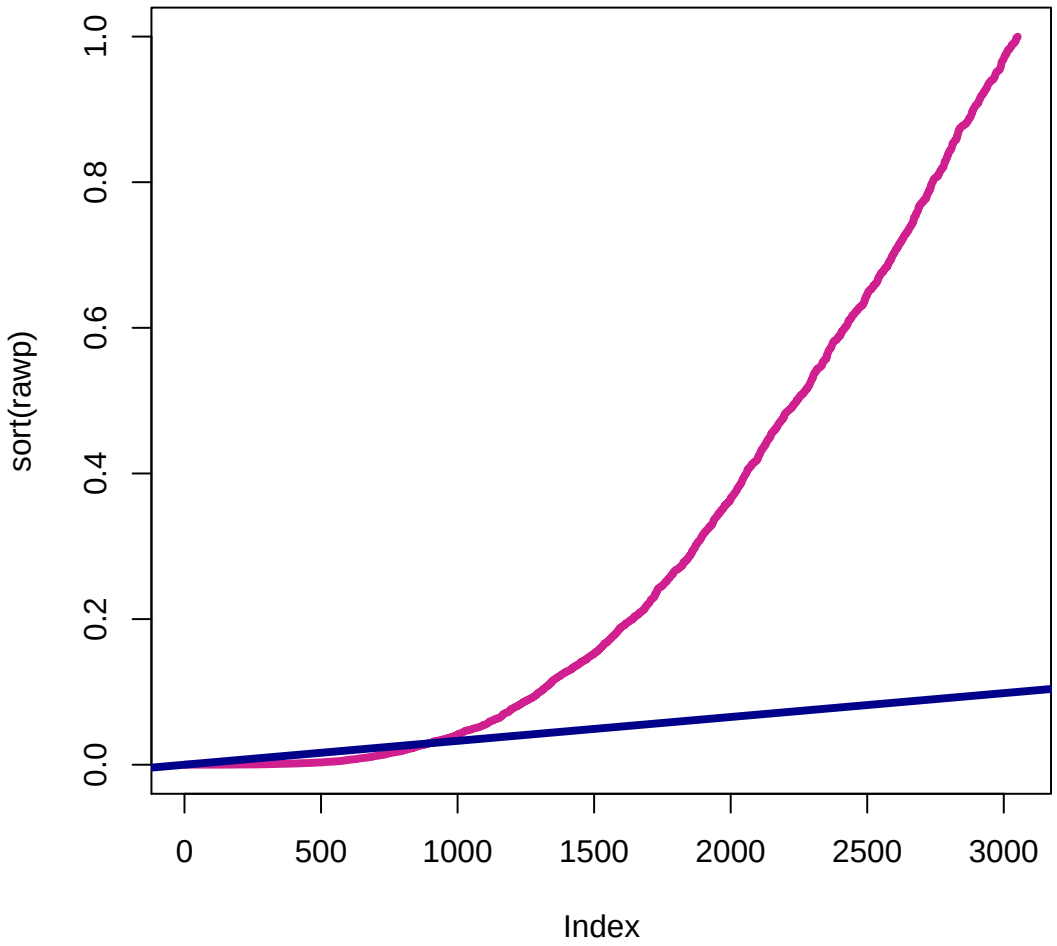
- To control $FDR = E(V/R)$ at level α, let

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

  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)



Compare sorted p–values with line through the origin of slope α/m (here, $\alpha = 0.1$).

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.

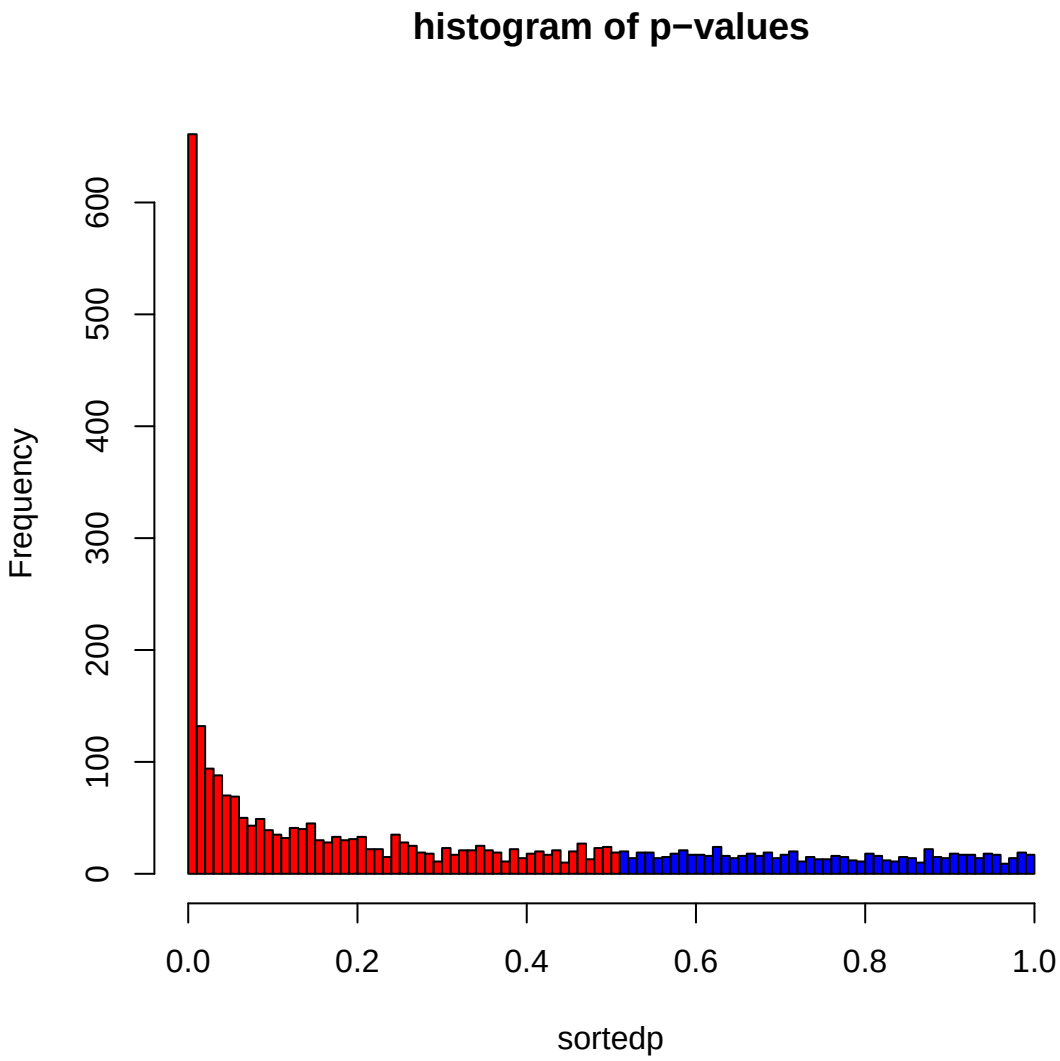
1. Estimate the number m_0 of non-diff. genes.

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.

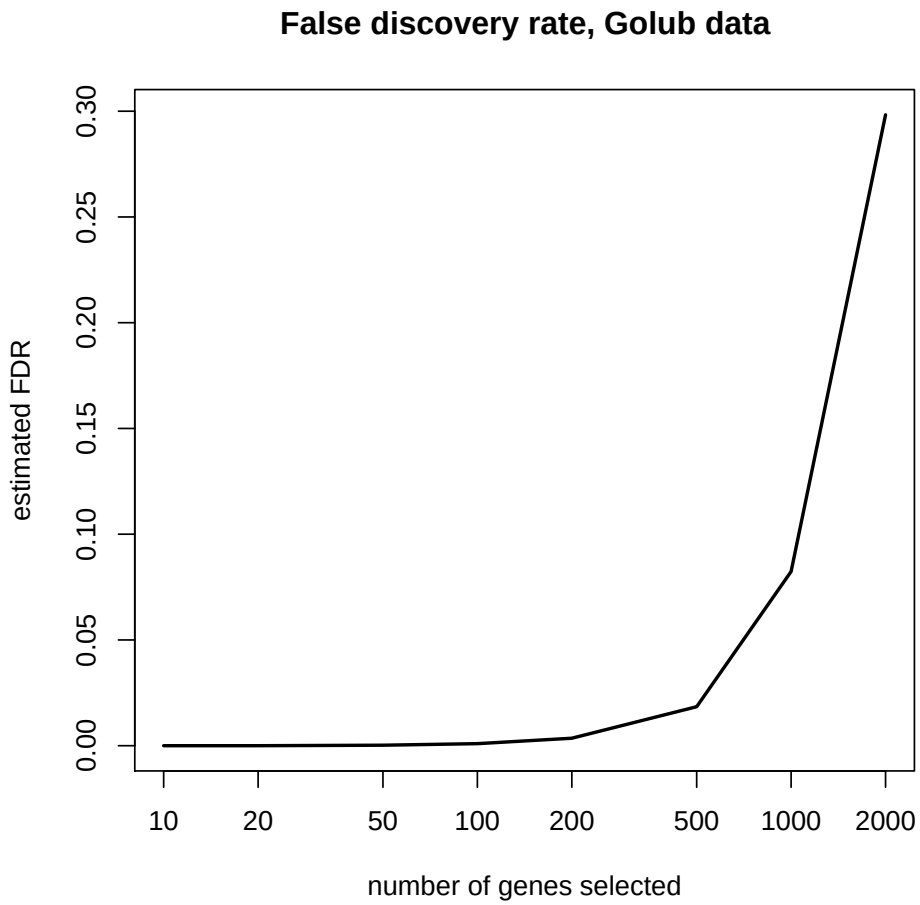
- For any threshold t_0 of the test statistic, compute the numbers V_b of genes with $T_{gb} > t_0$ (numbers of false positives).

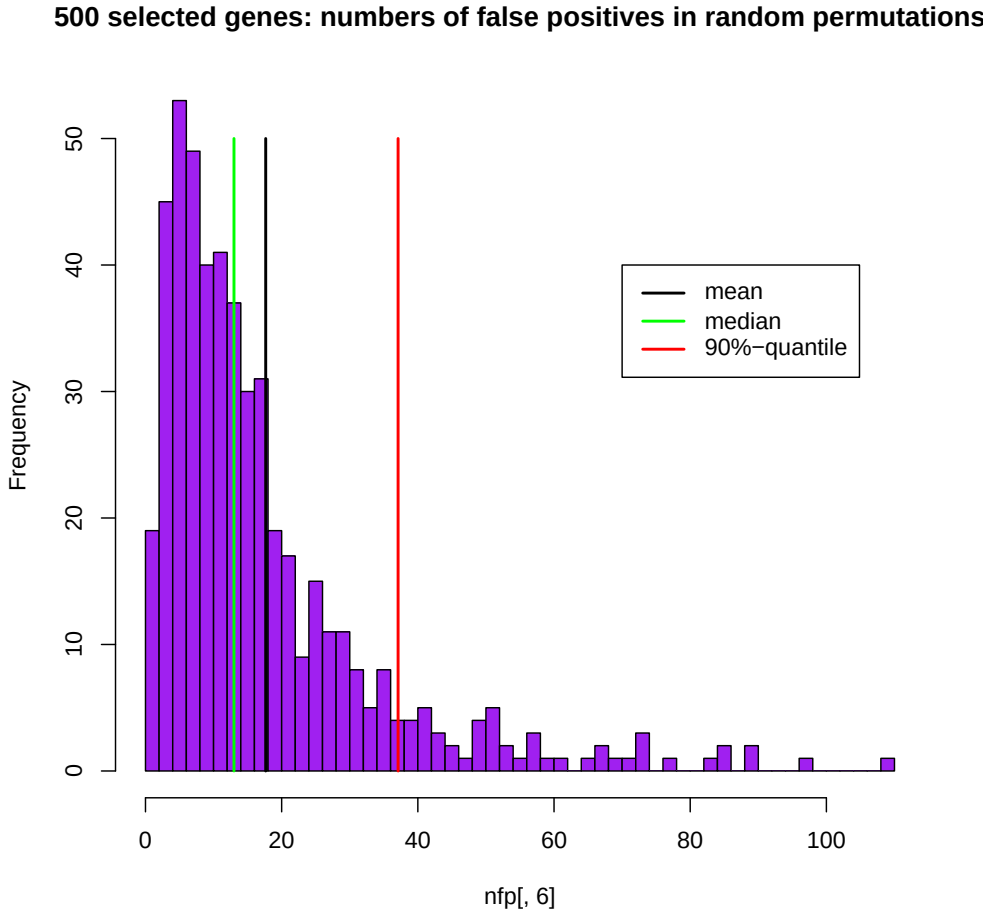
- The estimation of the FDR is based on the mean of the V_b. However, a quantile of the V_b may also be interesting, as the actual proportion of false positives may be much larger than the mean.

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

Estimation of the FDR: Example

Golub data





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.

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.

- 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.

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 small sample sizes, one may use a regularized t–statistic. The gene–specific variance/standard deviation is augmented by adding a constant (e.g. used in SAM, limma).

- 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.