Differential gene expression

Anja von Heydebreck
Dept. of Bio– and Chemoinformatics,
Merck KGaA

anja.von.heydebreck@merck.de

Slides partly adapted from S. Dudoit and A. Benner

Outline

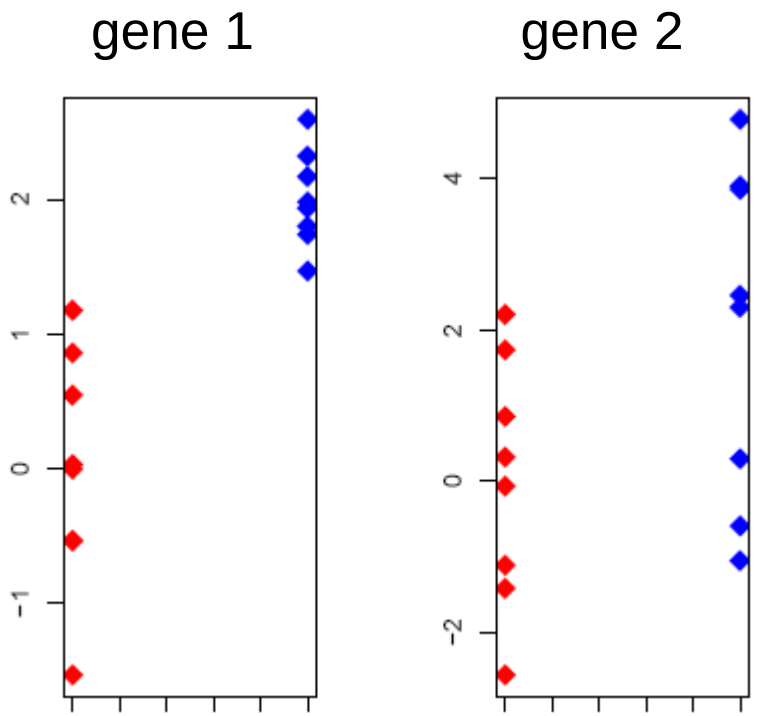
- Statistical tests: introduction
- Multiple testing
- Prefiltering of genes
- Linear models
- Gene screening using ROC curves

Identifying differentially expressed genes

- Aim: find genes that are differentially expressed between different conditions/phenotypes, e.g. two different tumor types.
- Estimate effects/differences between groups by (generalized) log–ratio, i.e., the difference on the log scale: log(X/Y) = log(X) – log(Y).
- Logs of ratios are symmetric around zero: The average of log(2) and log(1/2) is 0.
- If replicated measurements are available, first compute the within-group average on the log scale.

Identifying differentially expressed genes

- But what is a significant change?
- Depends on the variability within groups, which may be different from gene to gene.
- To assess the statistical significance of differences, conduct a statistical test for each gene.



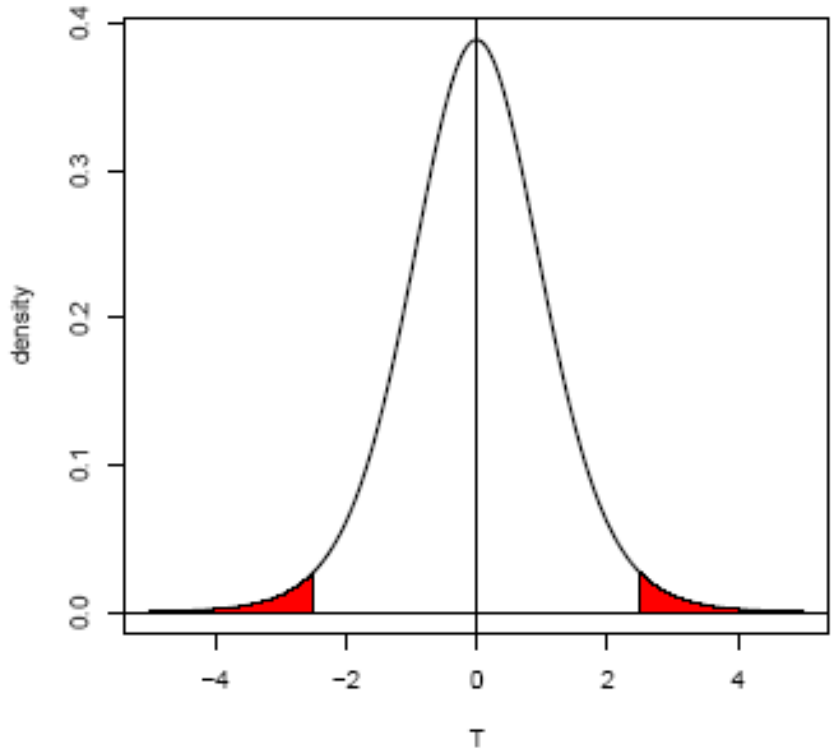
Statistical tests

- 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

- Standard t-test: assumes normally distributed data in each class (almost always questionable, but may be a good approximation), 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 is given as the fraction of permutations yielding a test statistic that is at least as extreme as the observed one.

Permutation tests

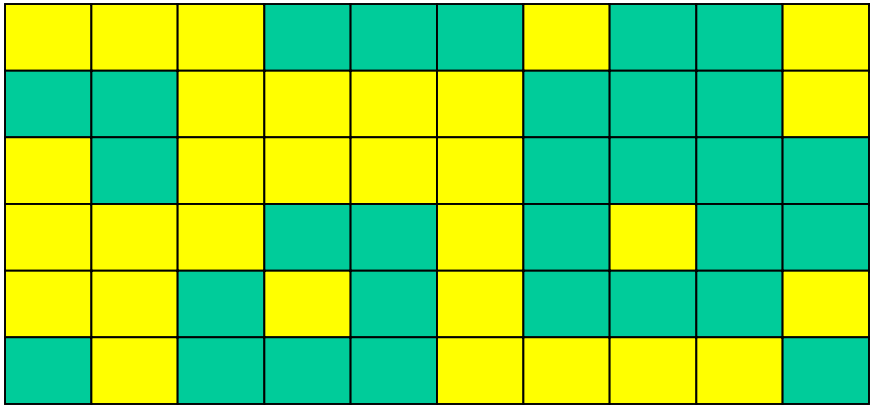
true class labels:

test statistic

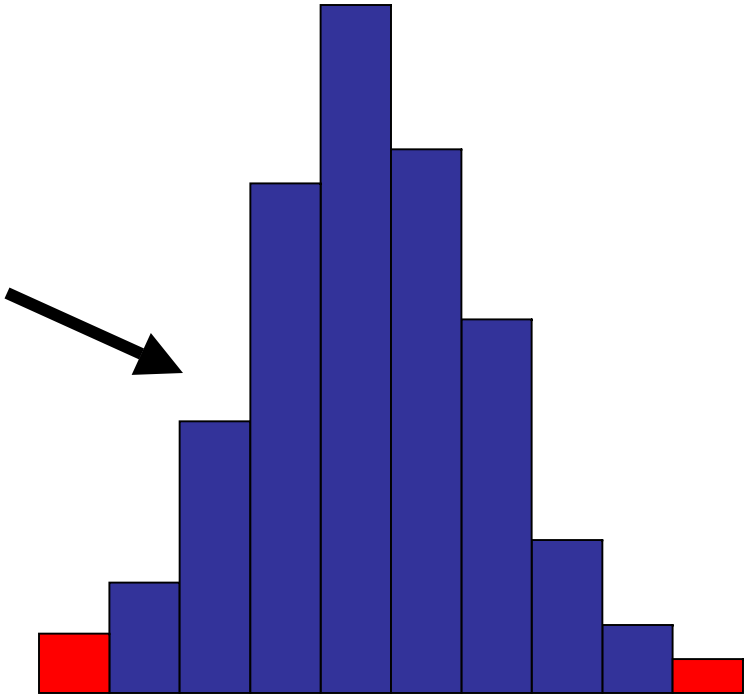
2.2

null distribution of test statistic

(random) permutations of class labels:



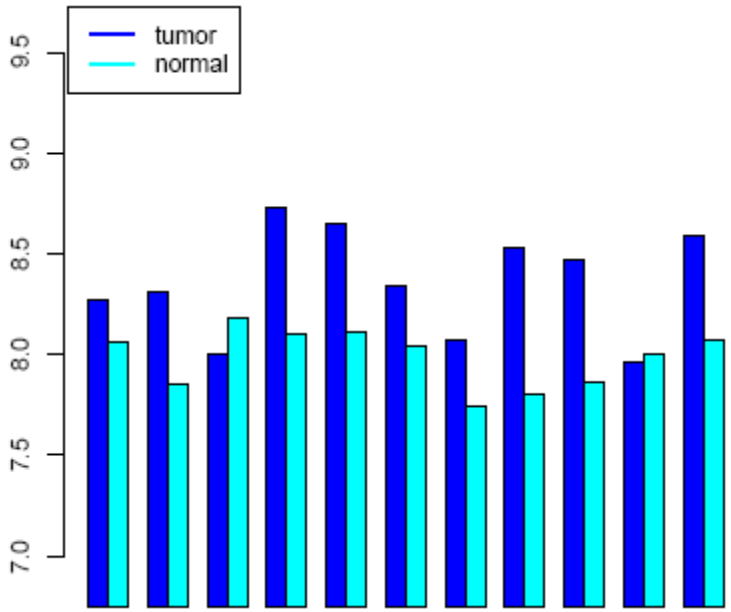
1.5
-0.4
2.3
0.7
0.2
-1.2



2.2

Statistical tests: Different settings

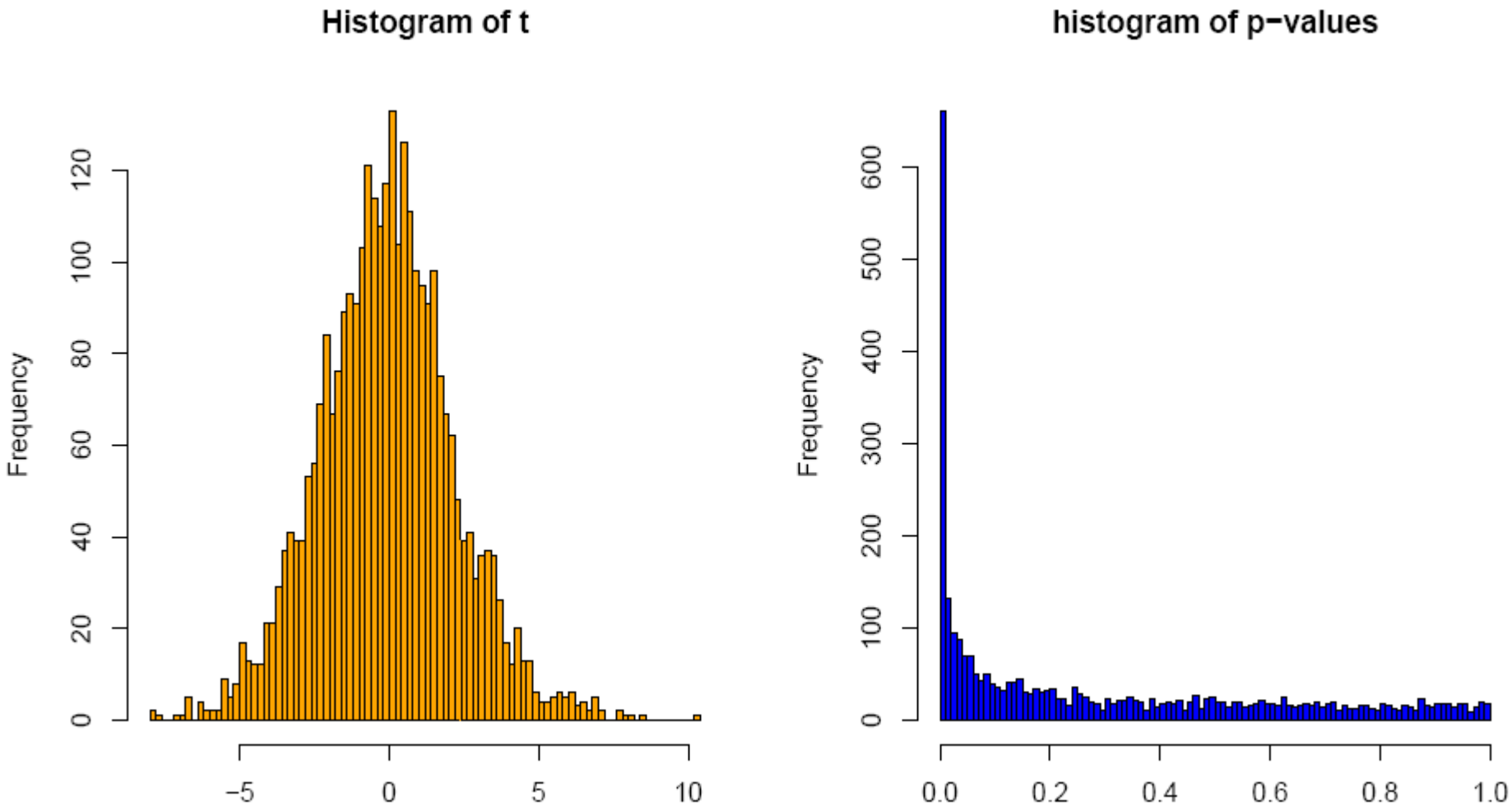
- comparison of two classes (e.g. tumor vs. normal)
- paired observations from two classes: e.g. the t–test for paired samples is based on the within–pair differences.
- more than two classes and/or more than one factor (categorical or continuous): tests may be based on linear models



paired samples

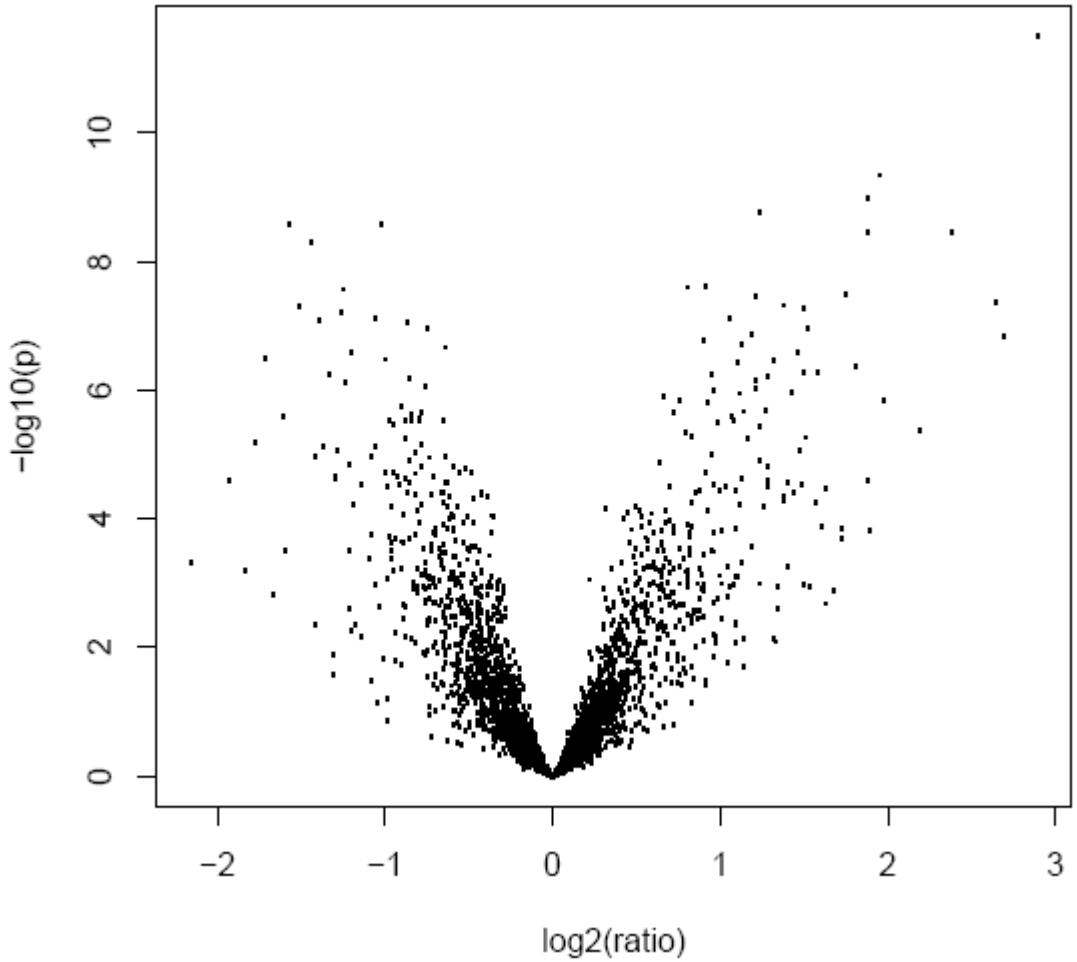
Example

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



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

The volcano plot: log-ratio vs. -log(p-value)



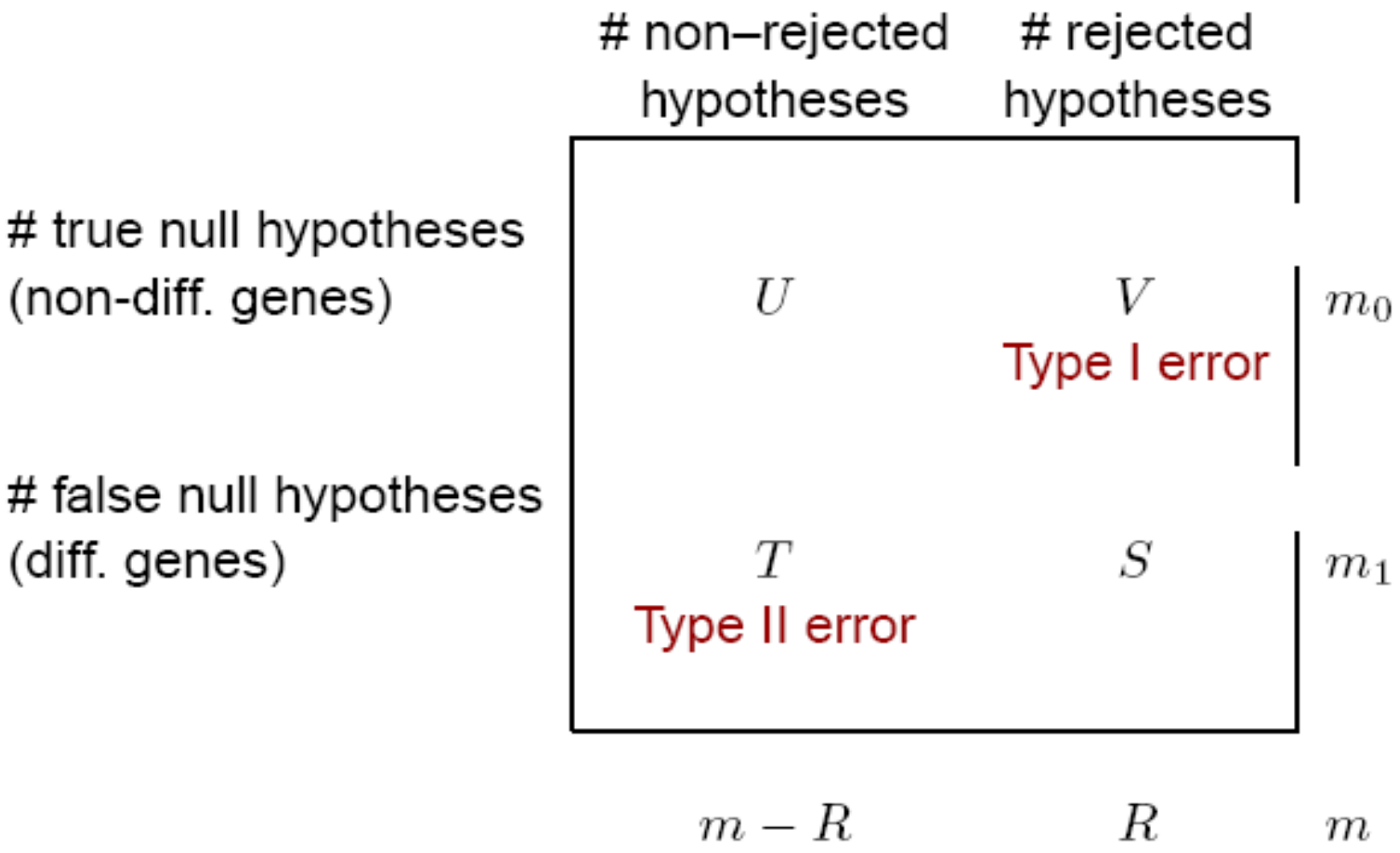
Multiple testing: the problem

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.

Multiple testing methods allow to assess the statistical significance of findings.

Multiple hypothesis testing



From Benjamini & Hochberg (1995).

Type I error rates

- Family–wise error rate (FWER). The FWER is defined as the probability of at least one Type I error (false positive) among the genes selected as significant:

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

Type I error rates

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

$$FDR = E(Q),$$

with

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

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.

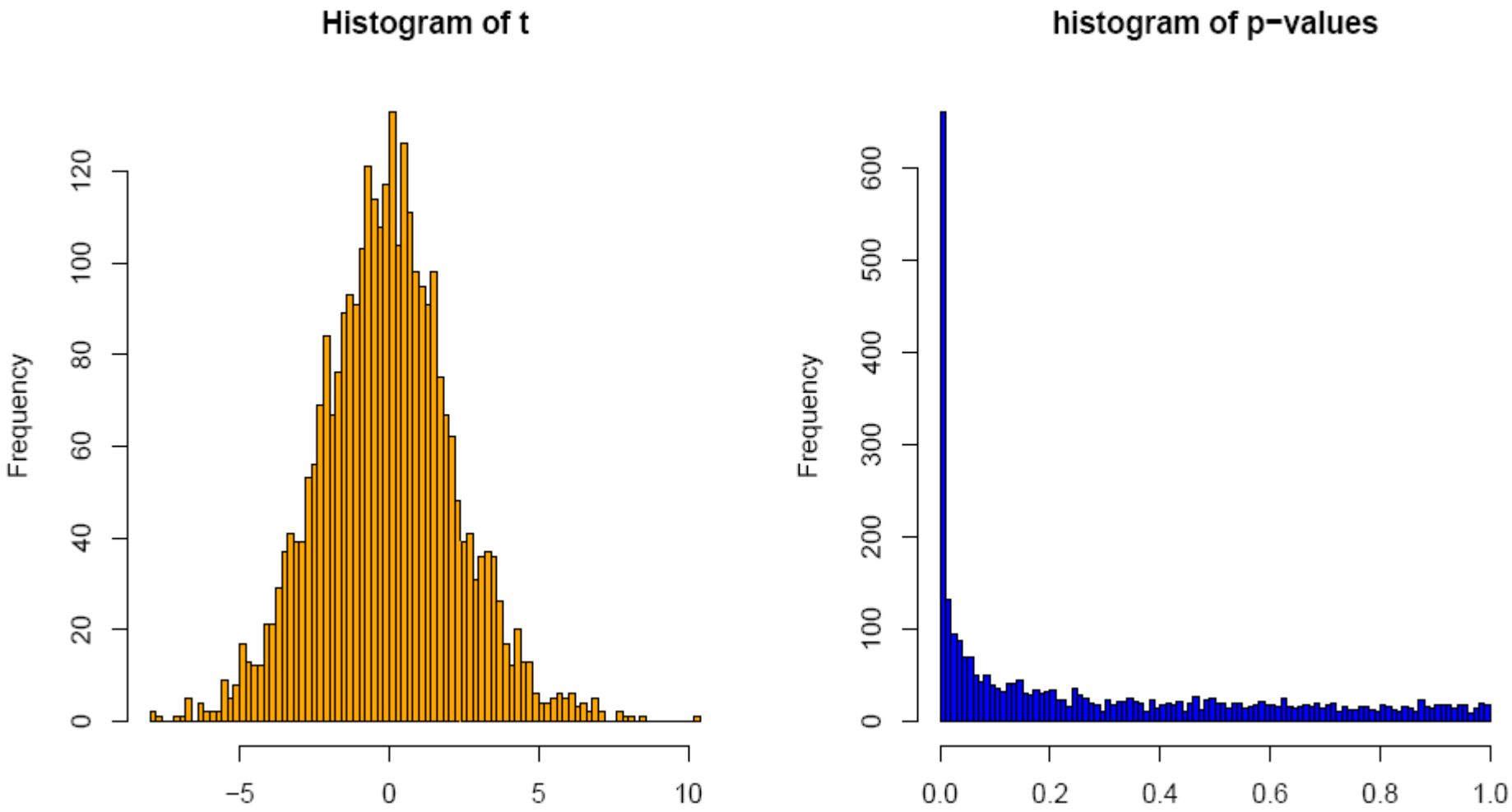
Bonferroni adjusted p–values:

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

Selecting all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α, that is, $Pr(V > 0) \leq \alpha$.

Example

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



98 genes with Bonferroni-adjusted $p < 0.05$, $p_{raw} < 0.000016$

FWER: Alternatives to Bonferroni

- There are alternative methods for FWER p-value adjustment.
- The permutation–based Westfall-Young method takes the correlation between genes into account and is typically more powerful for microarray data.
- See the Bioconductor package multtest.

Controlling the FDR (Benjamini/Hochberg)

- FDR: the expected proportion of false positives among the significant genes.

- 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 $j = 1, \ldots, j^\star$.

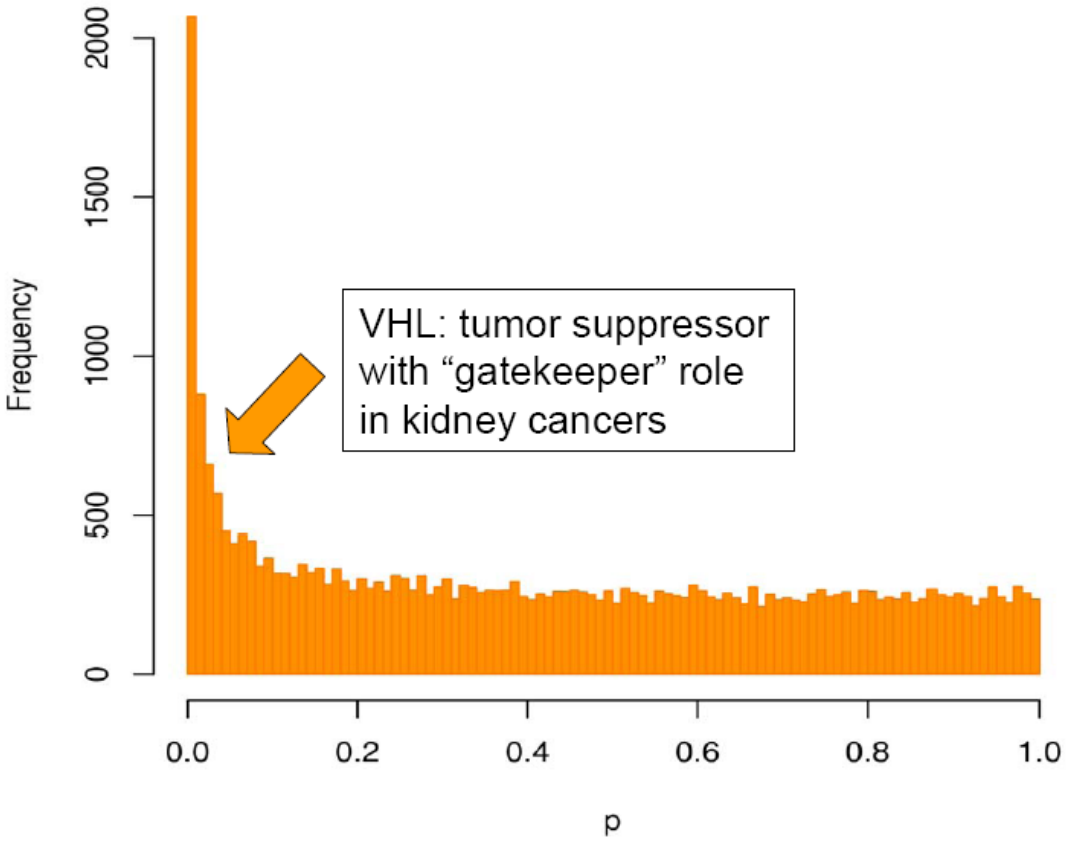
- Is valid for independent test statistics and for some types of dependence. Tends to be conservative if many genes are differentially expressed. Implemented in multtest.

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.
- For some applications, even the unadjusted p–values may be most appropriate (e.g. comparison of functional categories of affected vs. unaffected genes).

More is not always better

- On a genome-wide array with, say, 50,000 genes/ESTs, 50 genes can be expected to have a p-value below 0.001 by chance.
- Furthermore, the most significant genes are not necessarily the most biologically relevant ones.
- Therefore, it may be worthwile focusing on genes of particular biological interest from the beginning.



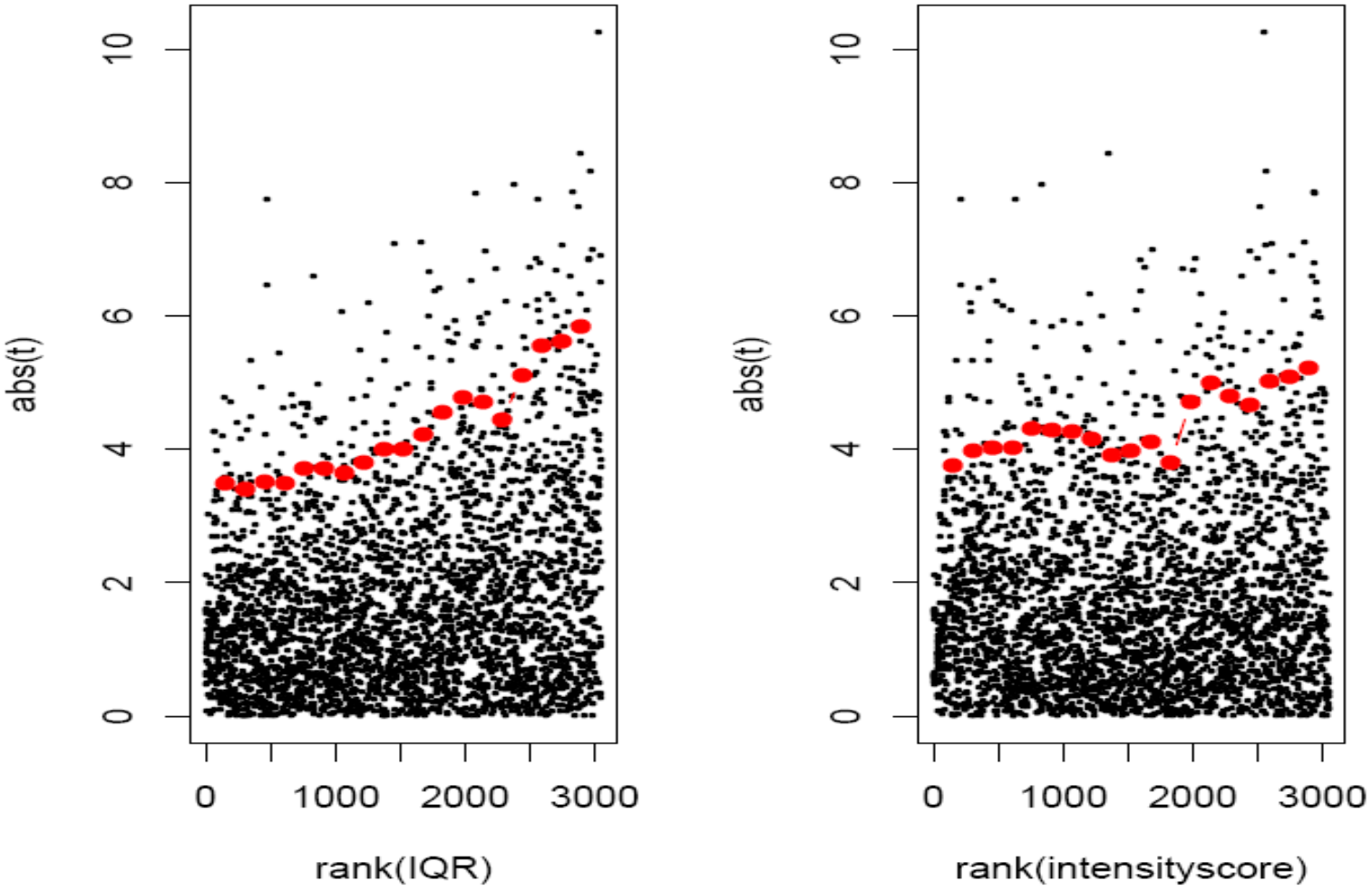
Boer et al., Genome Res. 2001: kidney tumor/normal profiling study

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 (-> use global criteria that are independent of phenotype distinctions).

Prefiltering by intensity and variability

Golub data. Ranks of interquartile range and 75%–quantile of intensities vs. absolute *t*–statistic.



Few replicates – moderated t–statistics

- With the *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), the variance estimates are unstable.
- In a moderated t–statistic, the estimated gene–specific variance s^2_g is augmented with s^2_0, a global variance estimator obtained from pooling all genes. This gives an interpolation between the *t*–statistic and a fold–change criterion:

$$T_g \sim \frac{\bar{X}_{g1} - \bar{X}_{g2}}{\sqrt{\mu s_g^2 + \lambda s_0^2}}.$$

- Bioconductor packages limma, siggenes.

Linear models

- Linear models are a flexible framework for assessing the associations of phenotypic variables with gene expression.
- The expression y_i of a given gene in sample i is modeled as linearly depending on one or several factors (e.g. cell type, treatment, encoded in x_{ij}) of the sample:

$$y_i = a_1 x_{i1} + \ldots + a_m x_{im} + \varepsilon_i.$$

- Estimated coefficients a_j and their standard errors are obtained using least squares, assuming normally distributed errors ε_i (R function lm); or with a robust method (R function rlm).

Linear models

- Contrasts, that is, differences/linear combinations of the coefficients, express the differences between phenotypes and can be tested for significance (*t*–test).
- Example: Consider a study of three different types of kidney cancer. For each gene set up a linear model:

$$y_i = a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \varepsilon_i,$$

  where $x_{ij} = 1$ if tumor sample *i* is of type j, and 0 otherwise.
- The least squares estimates of the coefficients a_i are the mean expression levels in the classes.
- The contrast $a_1 - a_2$ expresses the mean difference between class 1 and 2.

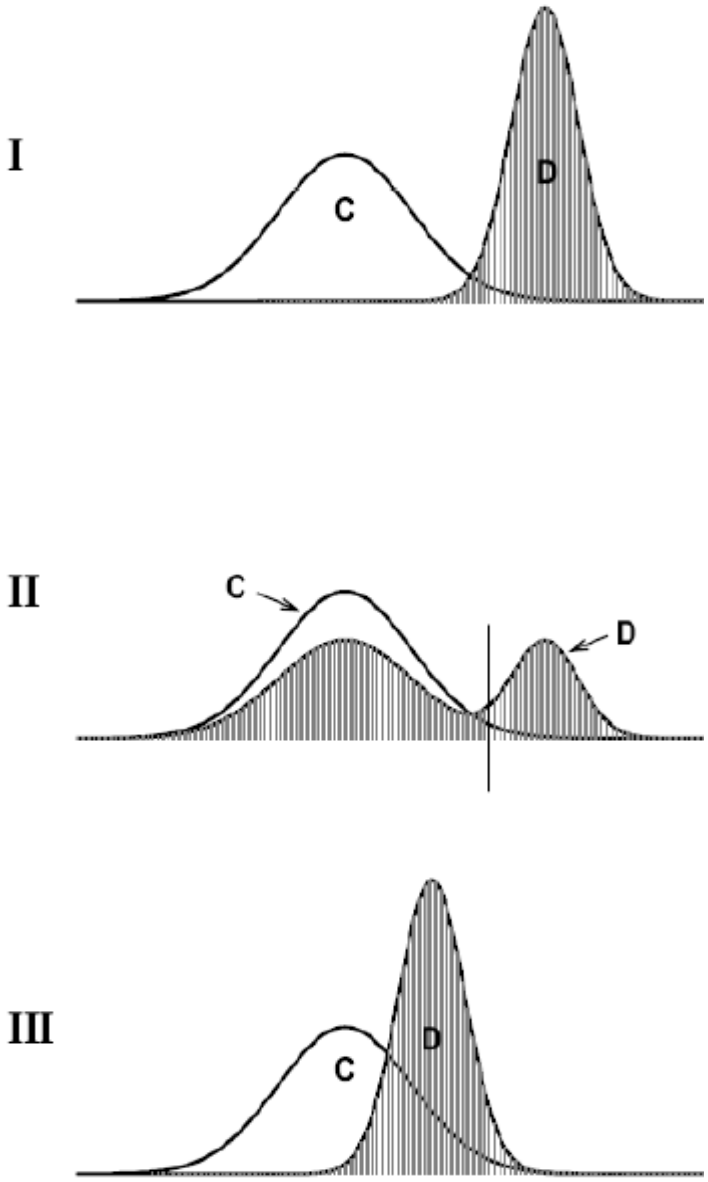
Linear model analysis with the Bioconductor package limma

- The phenotype information for the samples is to be entered as a design matrix (x_{ij} from the above formula). The rows of the matrix correspond to the samples, and the columns to the coefficients of the linear model.
- Contrasts are extracted after fitting the linear model.
- The significance of contrasts is assessed with a moderated *t*–statistic.

Gene screening using ROC curves

- Screening for biomarkers: rank genes according to their ability to distinguish between two phenotypes (e.g. disease and control).
- ROC: receiver operating characteristic
- Pepe et al., Biometrics 2003.

One gene in two groups



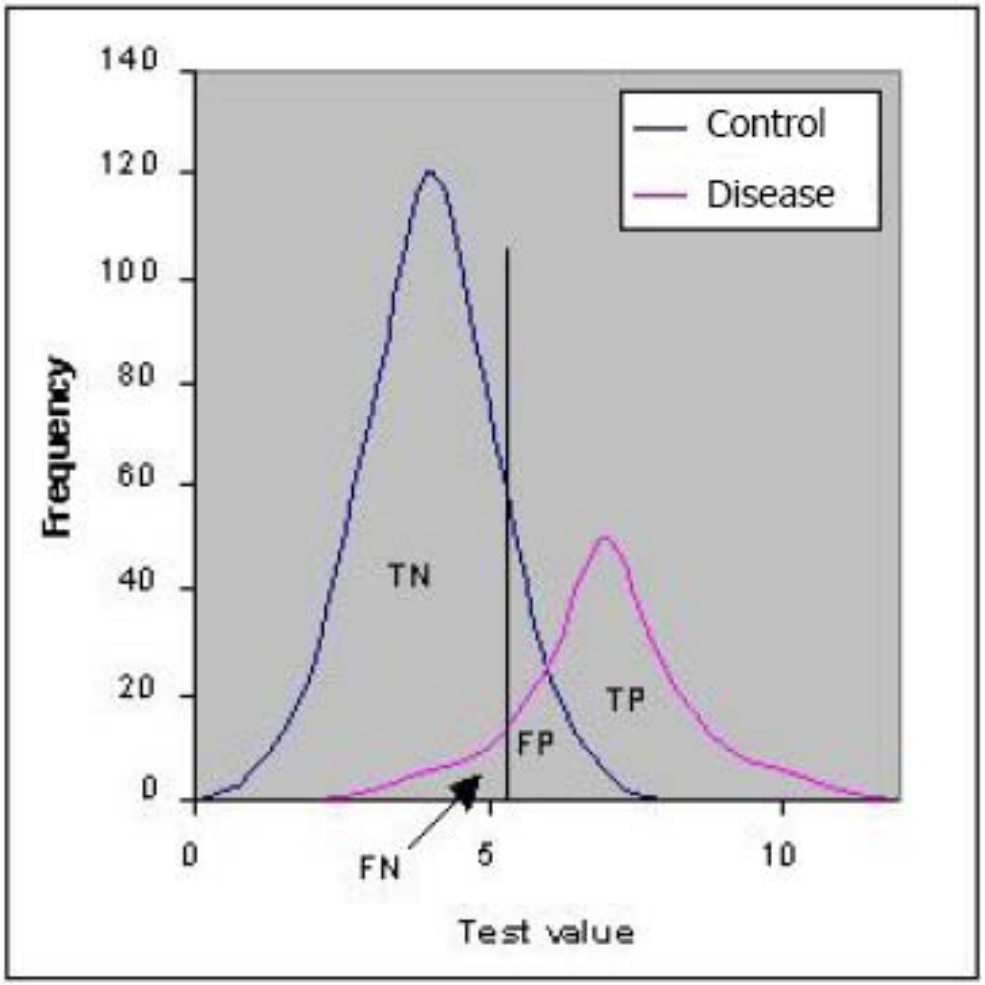
- **Panel I**: Almost complete separation between the distributions of controls (C) and disease (D).
- **Panel II and III**: Overlapping distributions.

Cancer screening:
Panel II is of more practical interest than panel III.

Panel II: clearly distinguishes a subset of D from C.
Panel III: The values of D are entirely within the range of those for C.

Pepe et al., Biometrics 2003

Sensitivity vs. Specificity



TN: true negative (specificity)
FP: false positive (1-spec.)
FN: false negative (1-sens.)
TP: true positive (sensitivity)

	Null hypothesis H_0	
	true	false
H_0 rejected	FP (α)	TP (1-β)
H_0 accepted	TN	FN

Gene screening by ROC analysis

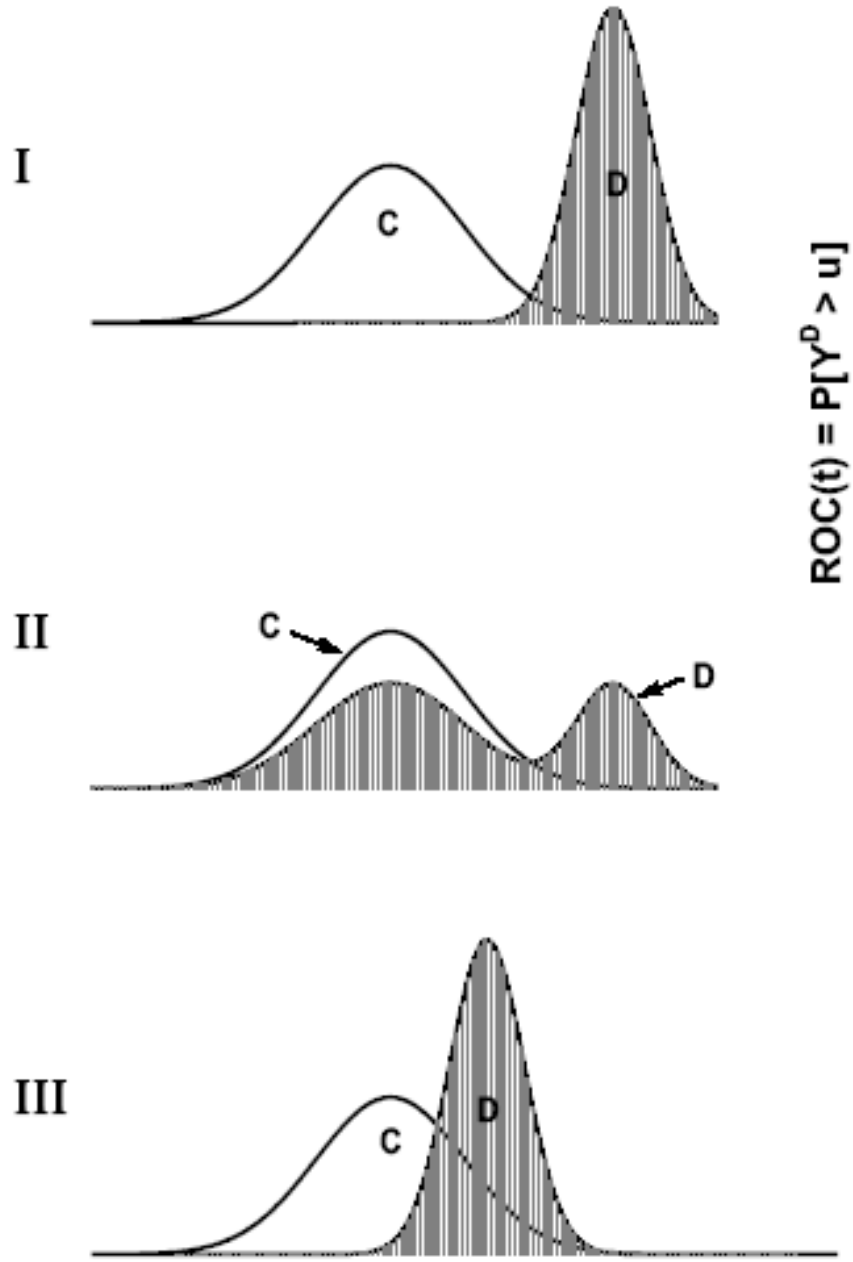
Let Y_g^i denote the relative expression level of gene g in sample $i=C,D$ after normalization.

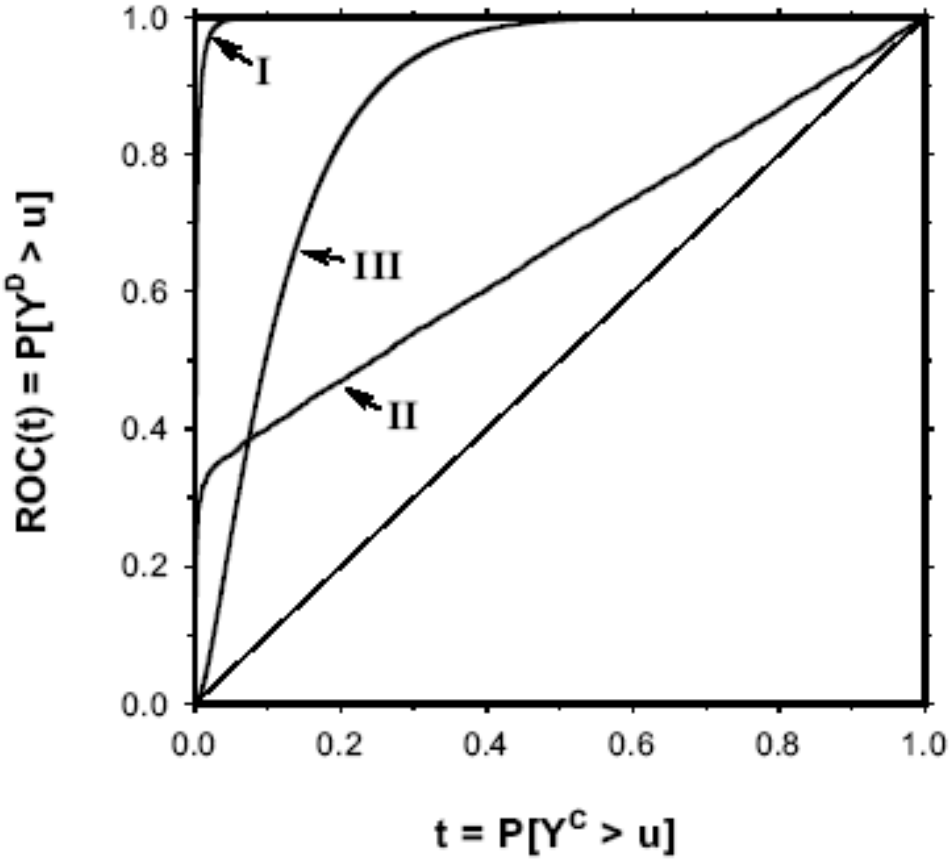
Each point on the ROC- curve, $\{t, ROC(t)\}$, corresponds to a different gene expression level u with

$$t = 1 - P[Y_g^C < u] \quad \text{(1-specificity/false positive)}$$

and

$$ROC(t) = P[Y_g^D \geq u] \quad \text{(sensitivity/true positive).}$$

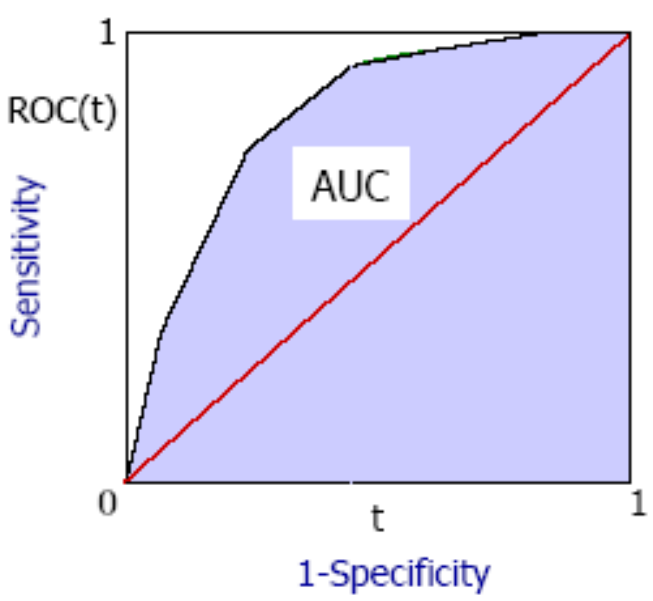


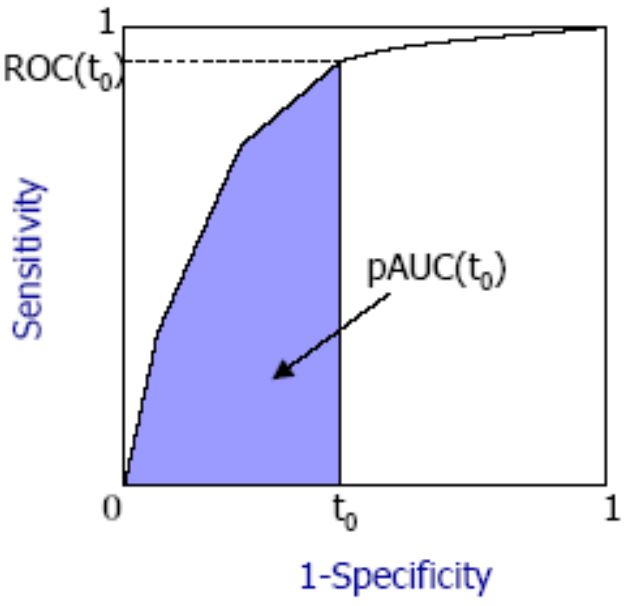


Resulting ROC curves for panels I- III

Pepe et al., Biometrics 2003

• The area under the curve (AUC, ~ Mann-Whitney statistic) scores for discrimination ability.
• Besides AUC, special interest is on the ROC curve at low values of *t*, corresponding to a maximum tolerable false positive rate t_0, or on the corresponding partial area under the curve, pAUC(t_0).





ROC curve screening in Bioconductor: package ROC

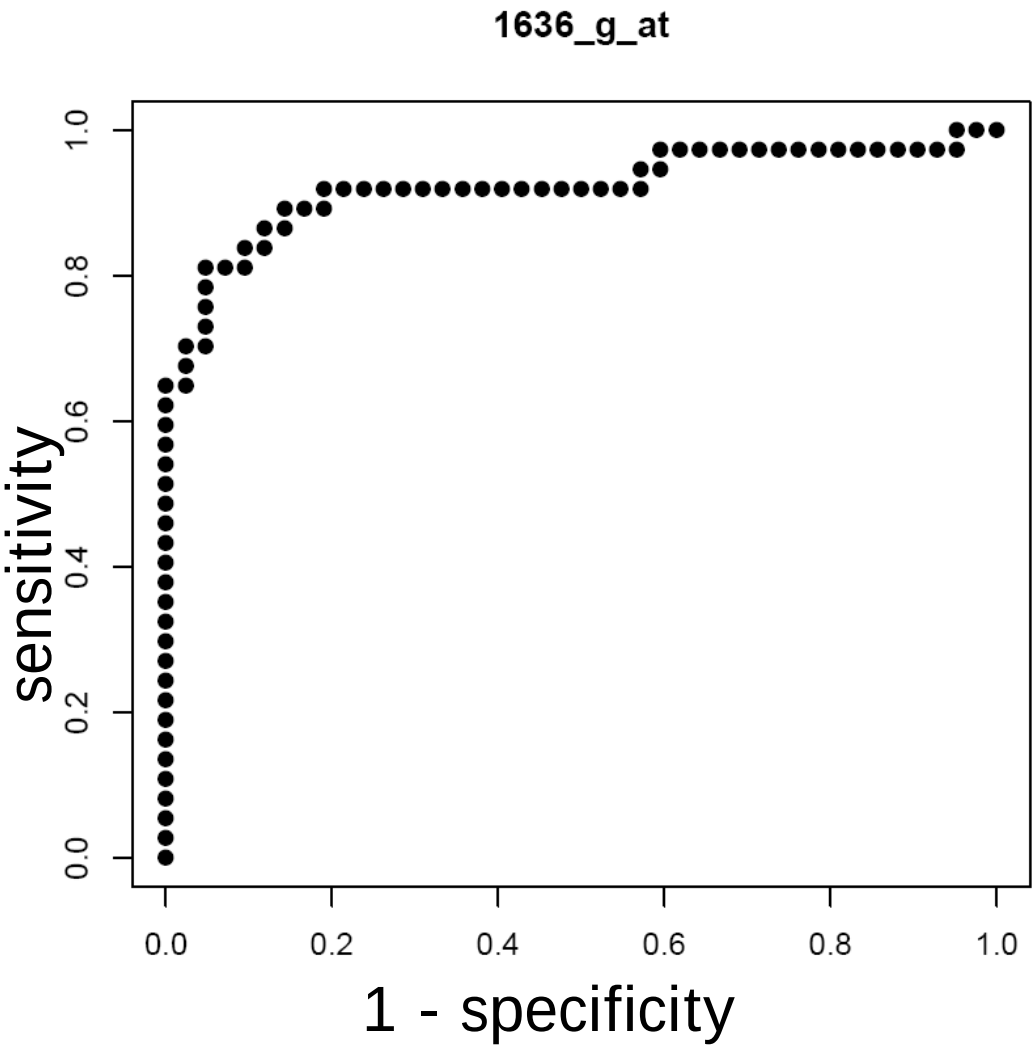
Suppose we have an *exprSet* object **eset** and a binary phenotype variable **labels** for the samples. We can compute the partial area under the ROC curve as follows:

```
> library(ROC)
> mypauc1 <- function(x) {
+   pAUC(rocdemo.sca(truth = labels, data = x, rule =
+               dxrule.sca), t0=0.1)
+ }
> pAUC1s <- esApply(eset, 1, mypauc1)
```

Example: B-cell ALL with/without the BCR/ABL translocation

Bioconductor data package ALL.
'Disease' class: samples with BCR/ABL translocation.

The probe set 1636_g_at, which represents the ABL1 gene, has the highest value of pAUC(0.1).



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.
- M. Pepe et al. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*, Vol. 59, 133–142.

- Summary measures are defined by $AUC = \int_0^1 ROC(t)\,dt$,

$ROC(t_0) = P[Y_g^D \geq y_{(1-t_0)}^C]$ and $pAUC(t_0) = \int_0^{t_0} ROC(t)\,dt$

where t_0 is a given false positive rate and $y_{(1-t_0)}^C$ is the corresponding $(1\text{-}t_0)$ quantile of the distribution of Y_g^C.

The value $ROC(t_0)$ gives the proportion of target samples with expression levels above the $(1\text{-}t_0)$ quantile of control samples.

The partial area under the curve, $pAUC(t_0)$, averages this proportion across values of $t \leq t_0$.