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 test: 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 between group means on the log scale.

❍ To assess the statistical significance of differences, conduct a statistical test for each gene.

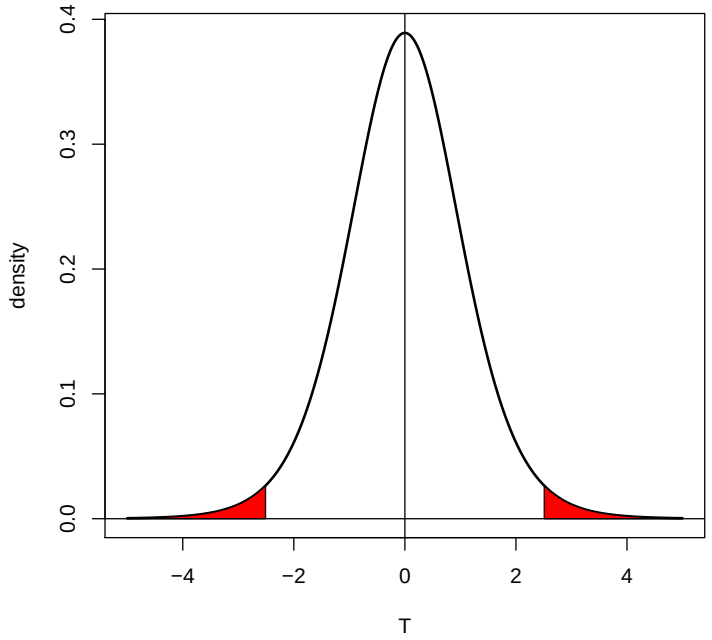
Statistical tests – example

❍ 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.



Statistical tests: Examples

- 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

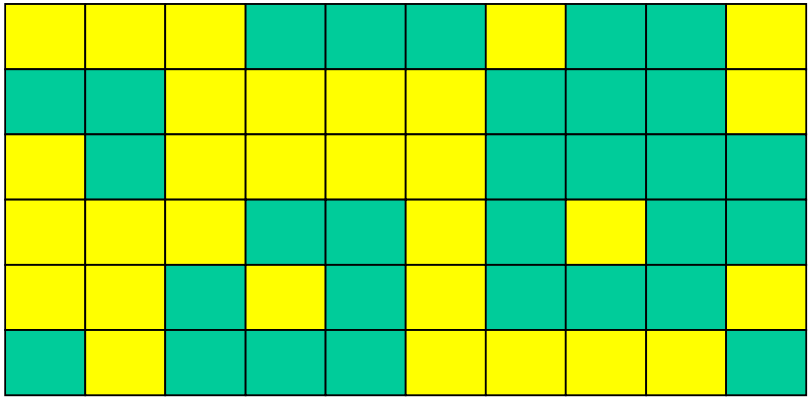
true class labels:

test statistic

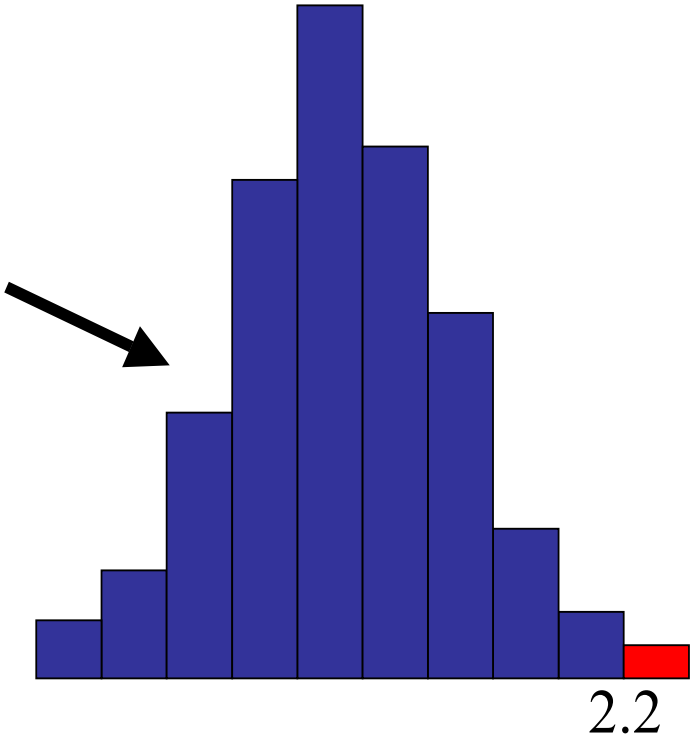
2.2

null distribution of test statistic

(random) permutations of class labels:



1.5
-0.4
2.3
0.7
0.2
-1.2



Statistical tests: Different settings

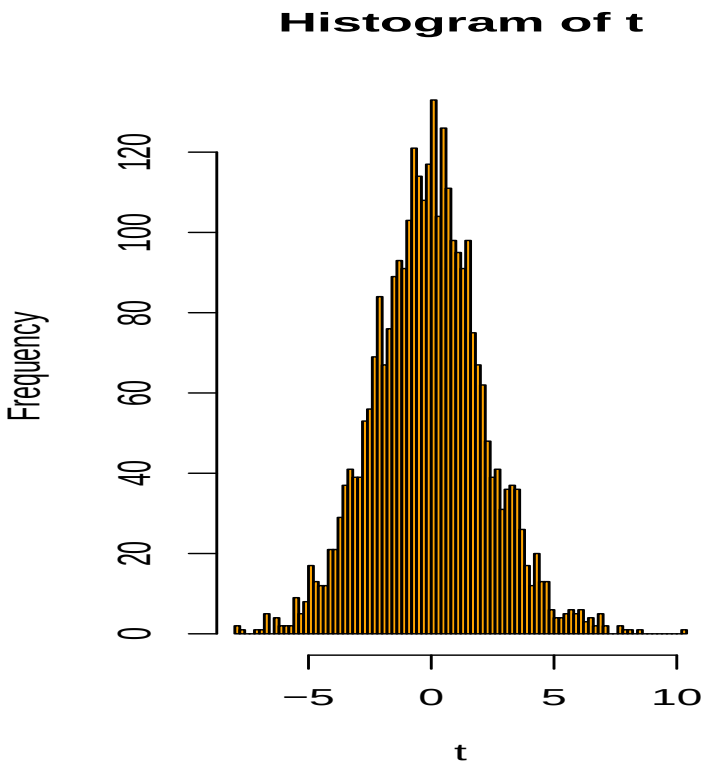
- comparison of two classes (e.g. tumor vs. normal)

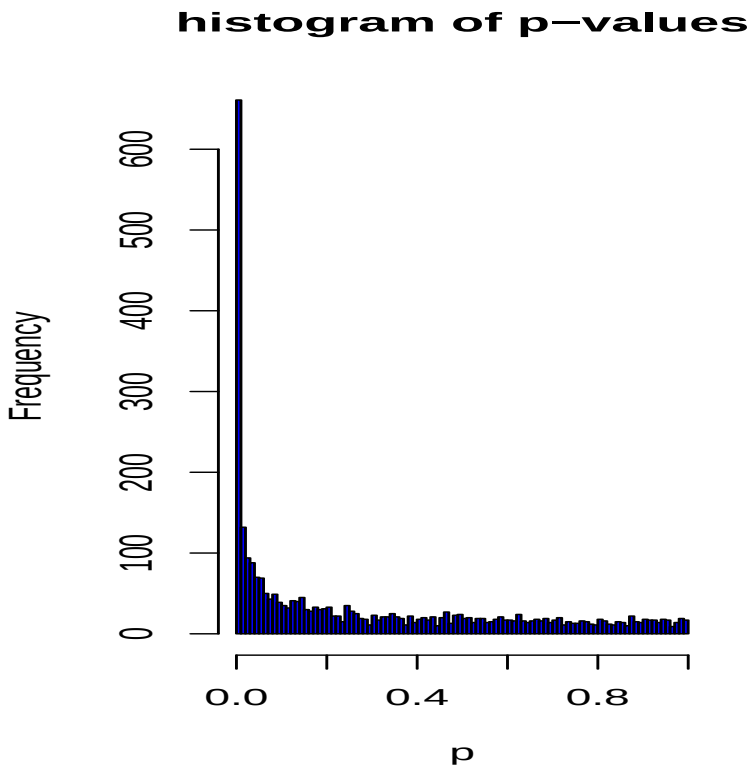
- paired observations from two classes: e.g. the t–test for paired samples is based on the within–pair differences.

- more than two classes and/or more than one factor (categorical or continuous): tests may be based on linear models

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.

Multiple testing methods allow to assess 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) among the genes selected as significant:

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

Type I error rates

2. **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.

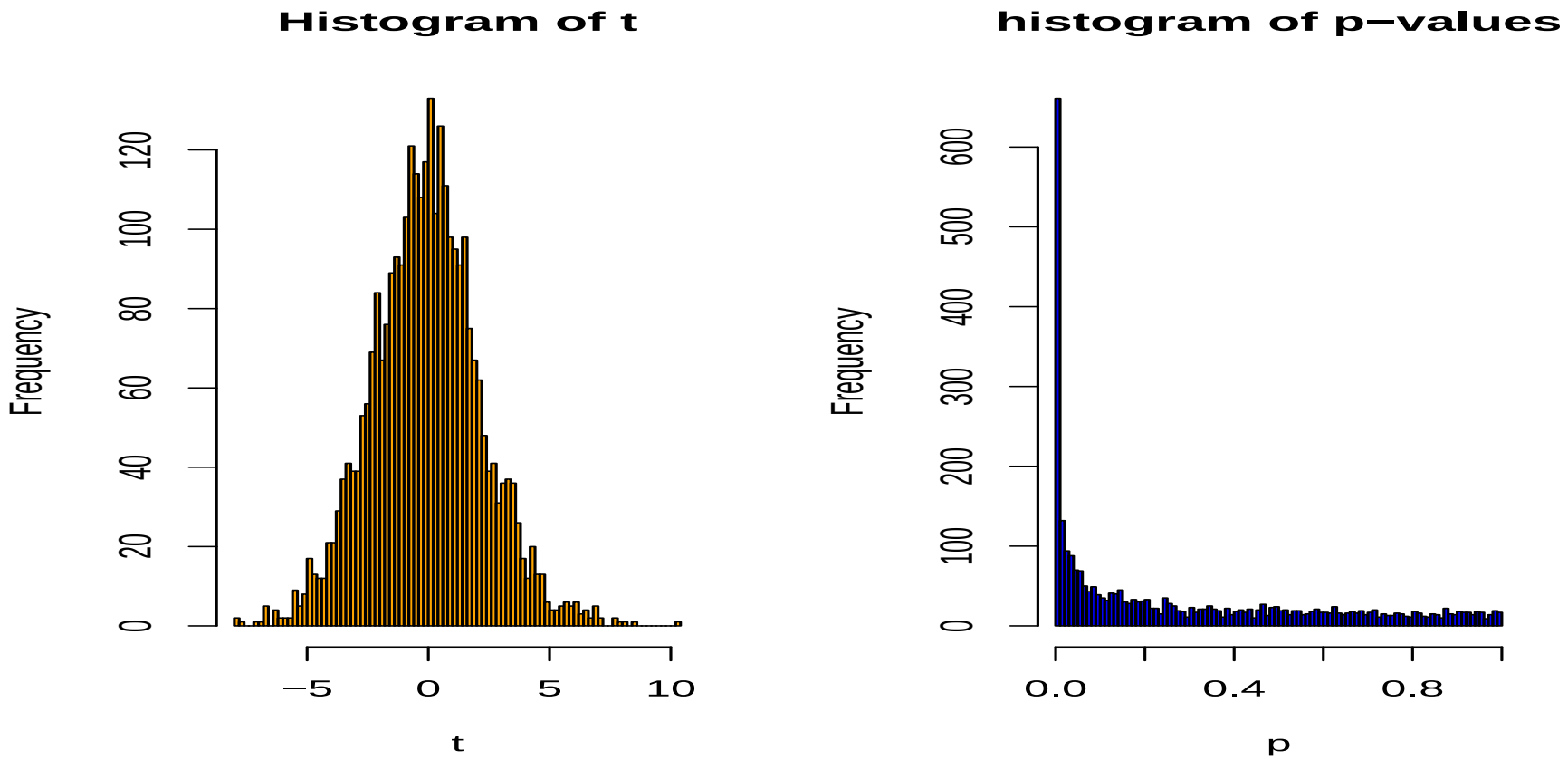
Bonferroni adjusted p–values:

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

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 $\tilde{p}_g < 0.05 \Leftrightarrow p_g < 0.000016$

FWER: Alternatives to Bonferroni

❍ There are alternative methods for FWER p-value adjustment, which can be more powerful.

❍ 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.

More is not always better

❍ Suppose you use a focused 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)

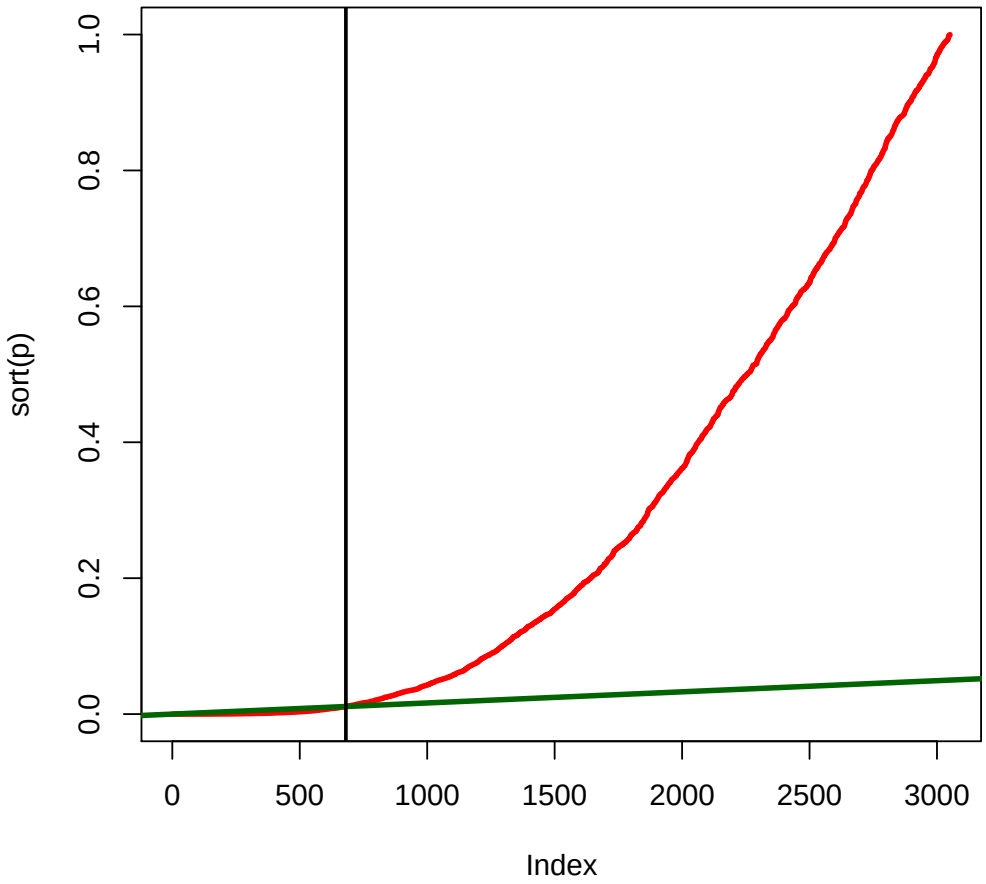
- Ordered unadjusted p–values: $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m}$.
- To control $FDR = E(V/R)$ at level α, let

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

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

- 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.

Controlling the FDR (Benjamini/Hochberg)



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

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).

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.

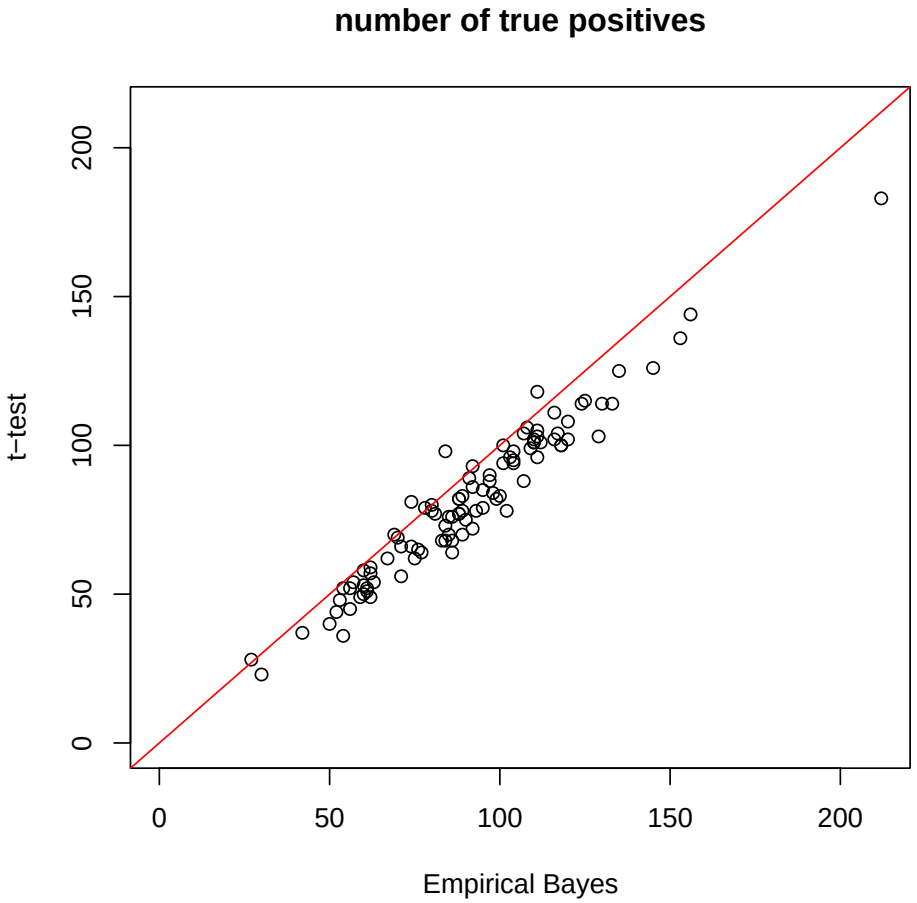
❍ In a moderated t–statistic, the estimated gene–specific variance s_g^2 is augmented with s_0^2, a global variance estimator obtained from pooling all genes. This gives an interpolation between the t–test and a fold–change criterion.

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

❍ Bioconductor packages limma, siggenes.

Moderated t–statistic

Repeatedly draw 4 ALL and 4 AML samples out of the total 38 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

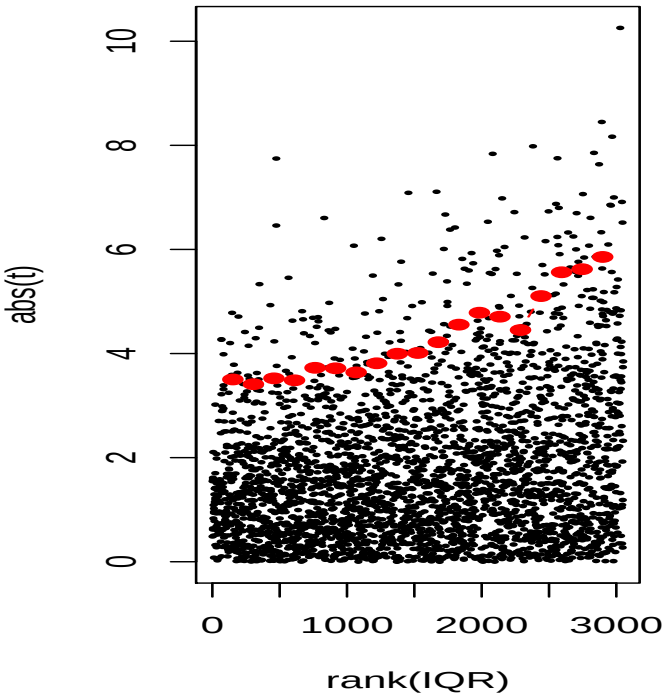
❍ What about prefiltering genes (according to intensity, variance etc.) to reduce the proportion of false positives?

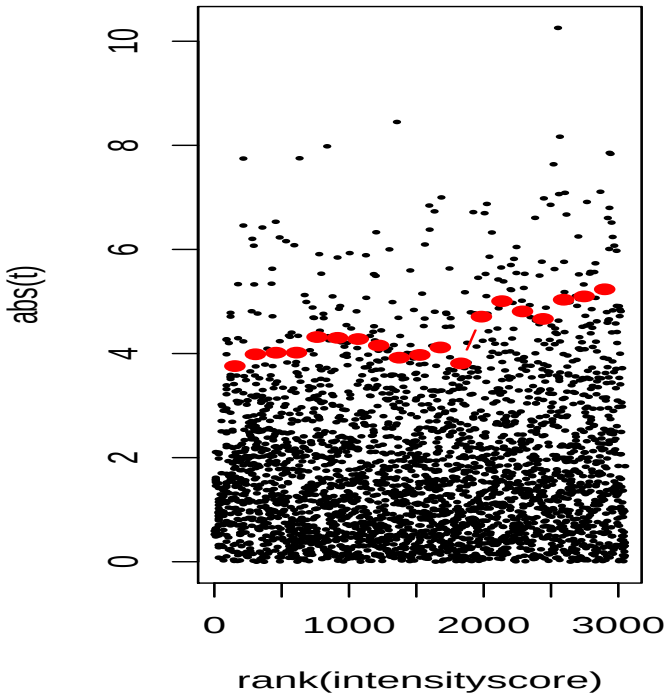
❍ Can be useful: Genes with low intensities in most of the samples or low variance across the samples are less likely to be interesting.

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





Linear models

❍ Linear models are a flexible framework for assessing the effects of phenotypic variables on gene expression.

❍ The expression y_i of a given gene in sample i is modeled as linearly depending on one or several attributes (factors; could be cell type, treatment, etc., encoded in x_{ij}) of the sample:

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

❍ 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:

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

where $x_i j = 1$ if tumor sample i is of type j, and 0 otherwise.
The coefficients $\hat{a}_i$ estimated by least squares 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 arrays, and the columns to the coefficients of the linear model.

❍ Contrasts are extracted after fitting the linear model.

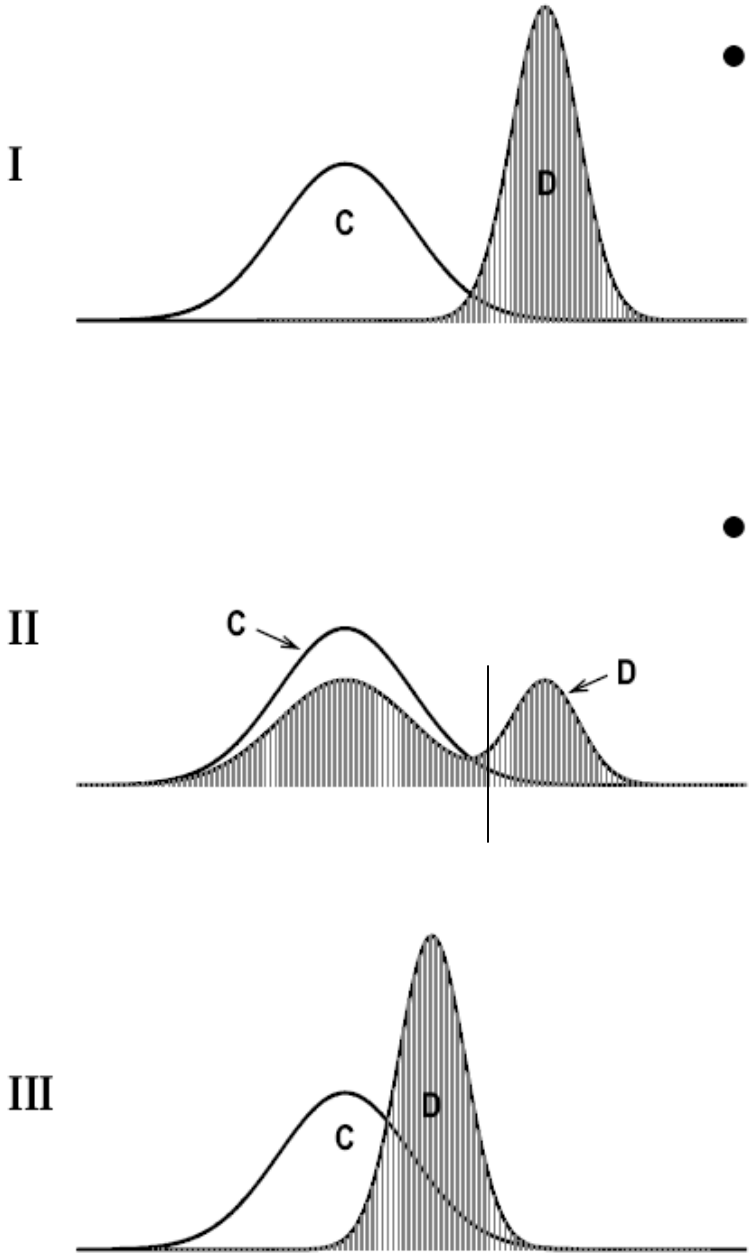
❍ The significance of contrasts is assessed with a moderated t–statistic.

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.

Gene screening using ROC curves

- 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.



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

  Classify with almost 100% accuracy.

- **Panels II and III:**
  Overlapping distributions.

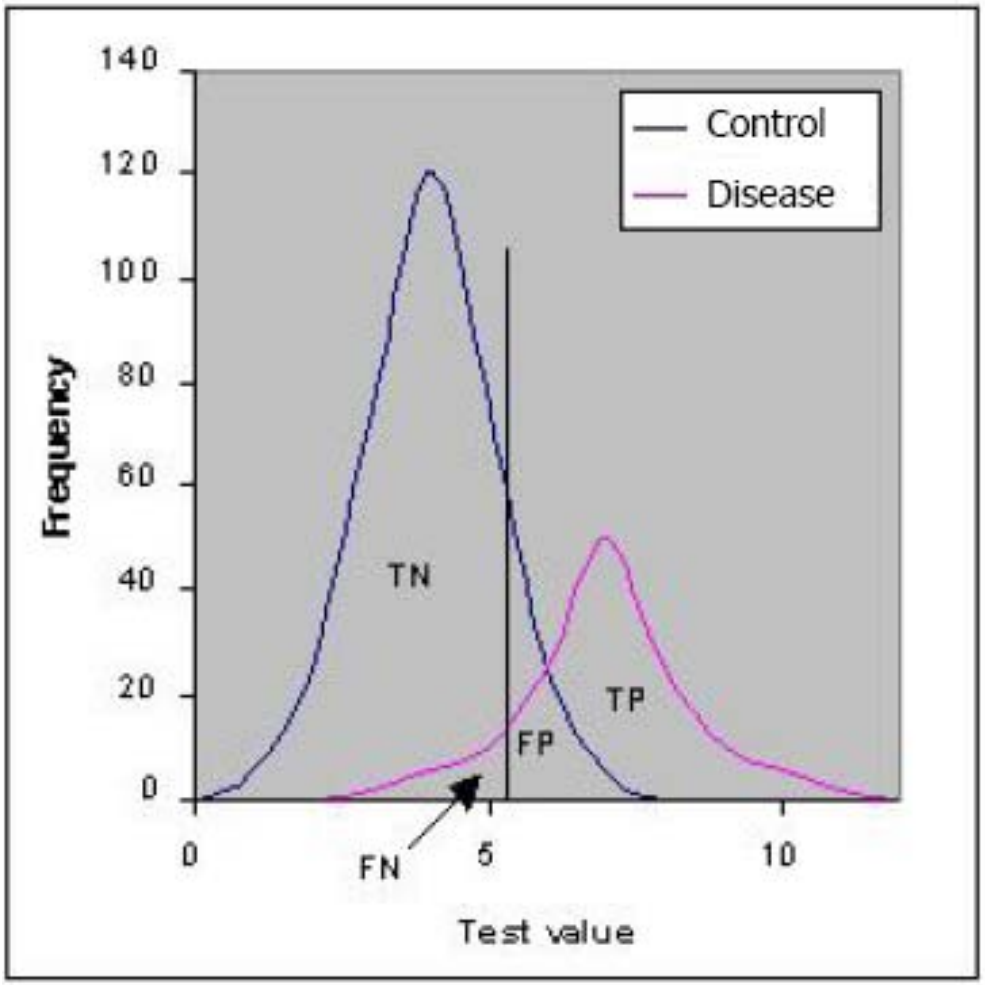
  Cancer screening:

  Panel II is of more practical interest than panel III.

  Panel II: clearly distinguishes a subset of D from C

  Panel III: values for D are entirely within the range of those for C.

(Pepe et al., Biometrics 2003)



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

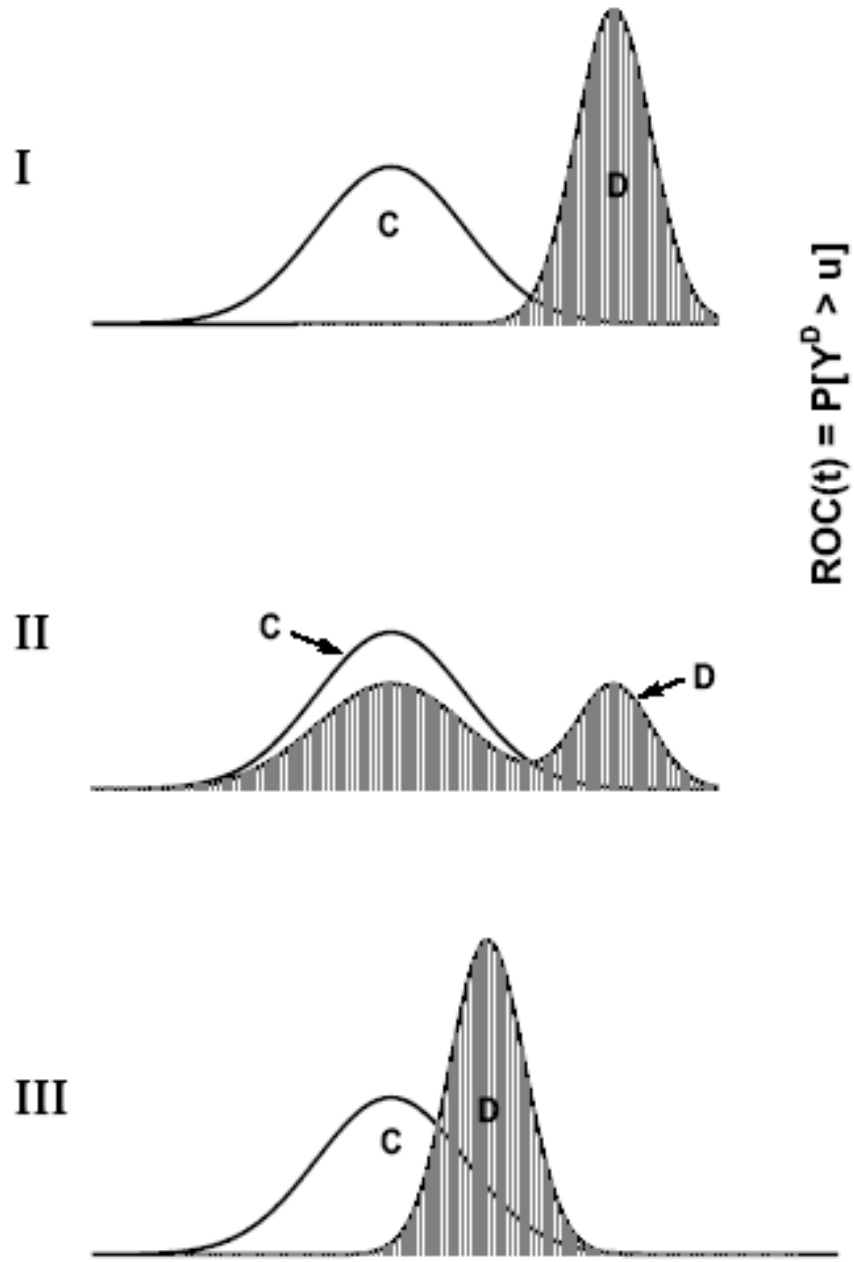
Let Y_g^i denote the relative expression level of gene g in sample $i=C,D$ after normalization.

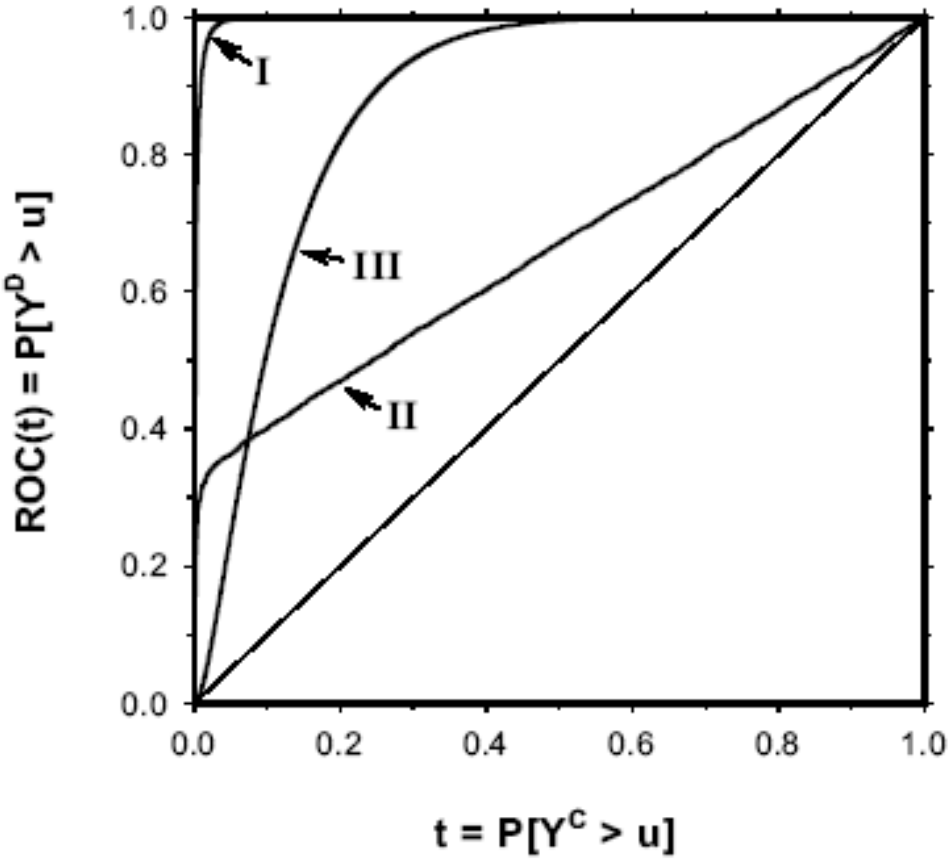
Each point on the ROC- curve, $\{t, ROC(t)\}$, corresponds to a different gene expression level u with

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

and

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



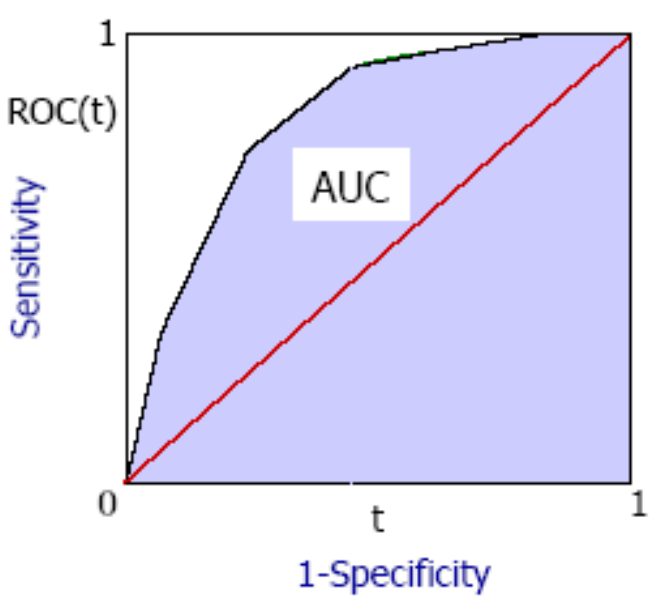


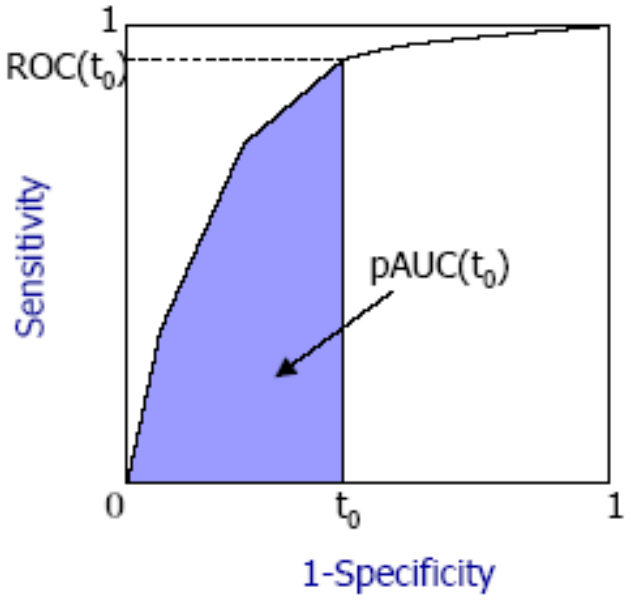
Resulting ROC curves for panels I- III

Pepe et al., Biometrics 2003

- AUC (~Mann-Whitney statistic) scores for discrimination ability (and equals 0.5 for a random classifier)

- Besides AUC, the area under the full ROC curve, more interest is on the ROC curve at low values of *t*, corresponding to a maximum tolerable false positive rate t_0.





- 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 $t \leq t_0$.

ROC curve screening with the 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).