

**Sensitivity, Specificity, ROC**

**Multiple testing**

**Independent filtering**

**Wolfgang Huber (EMBL)**

# Statistics 101

← bias

accuracy→

dispersion→

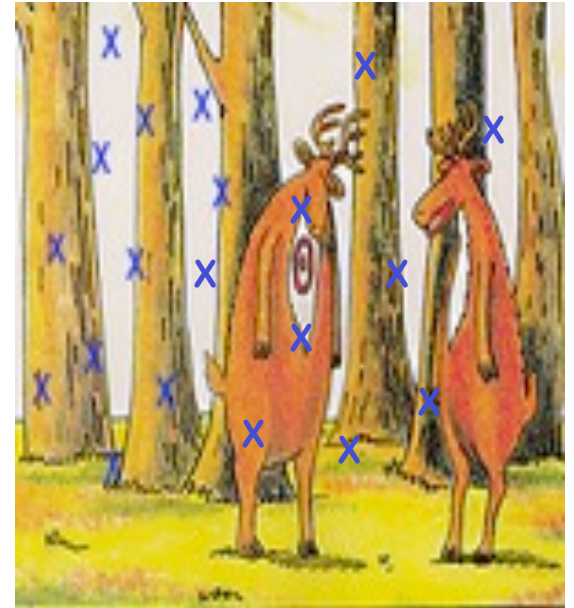
← precision



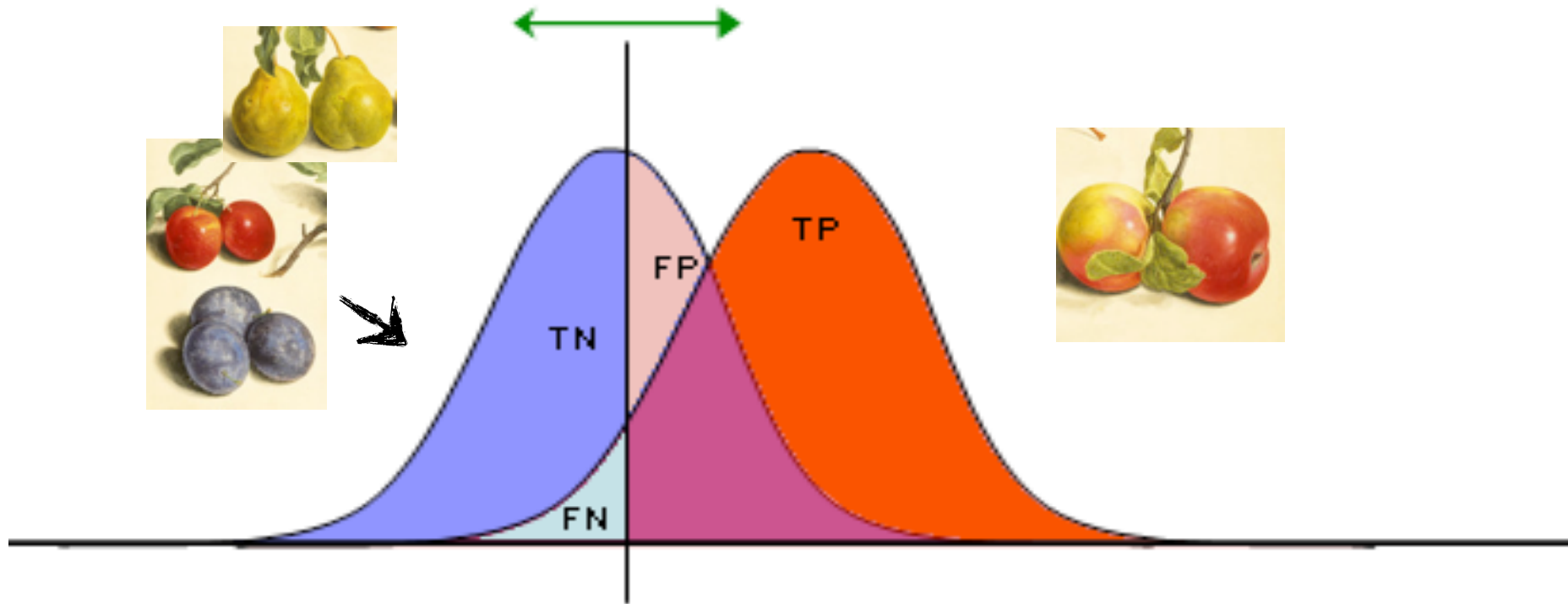
# Basic dogma of data analysis

Can always increase sensitivity on the cost of specificity, or vice versa, the art is to

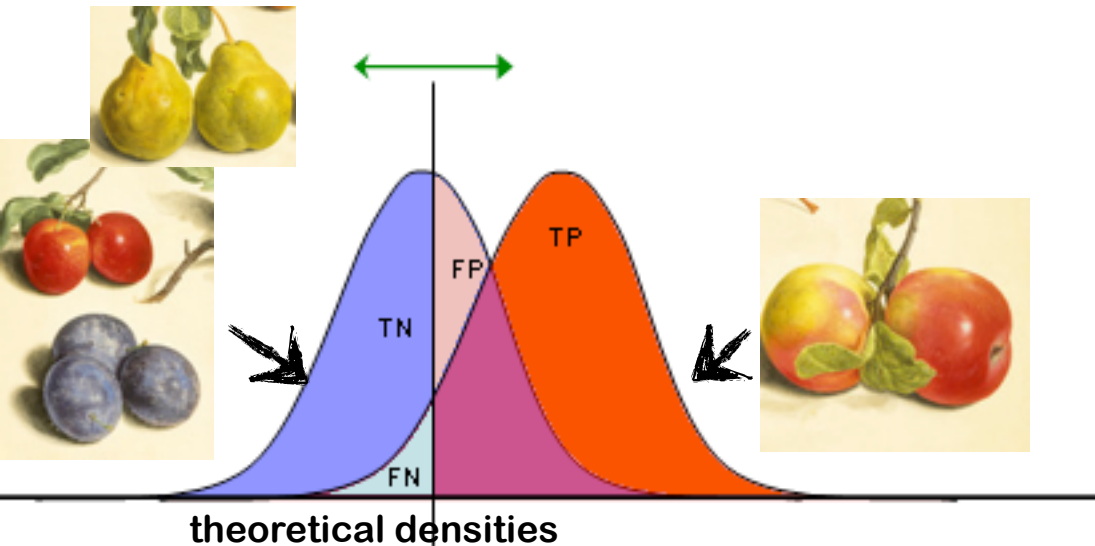
- optimize both
- find the best trade-off



# Problem: detecting apples from other fruit



# The apple detection assay and the receiver operating characteristic curve

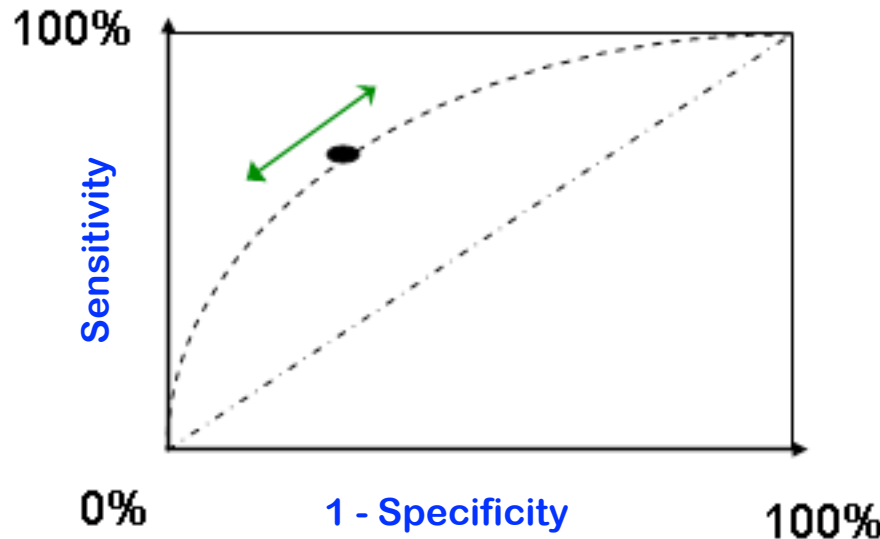


apples	other fruit
TP	FP
FN	TN
P	N

empirical results

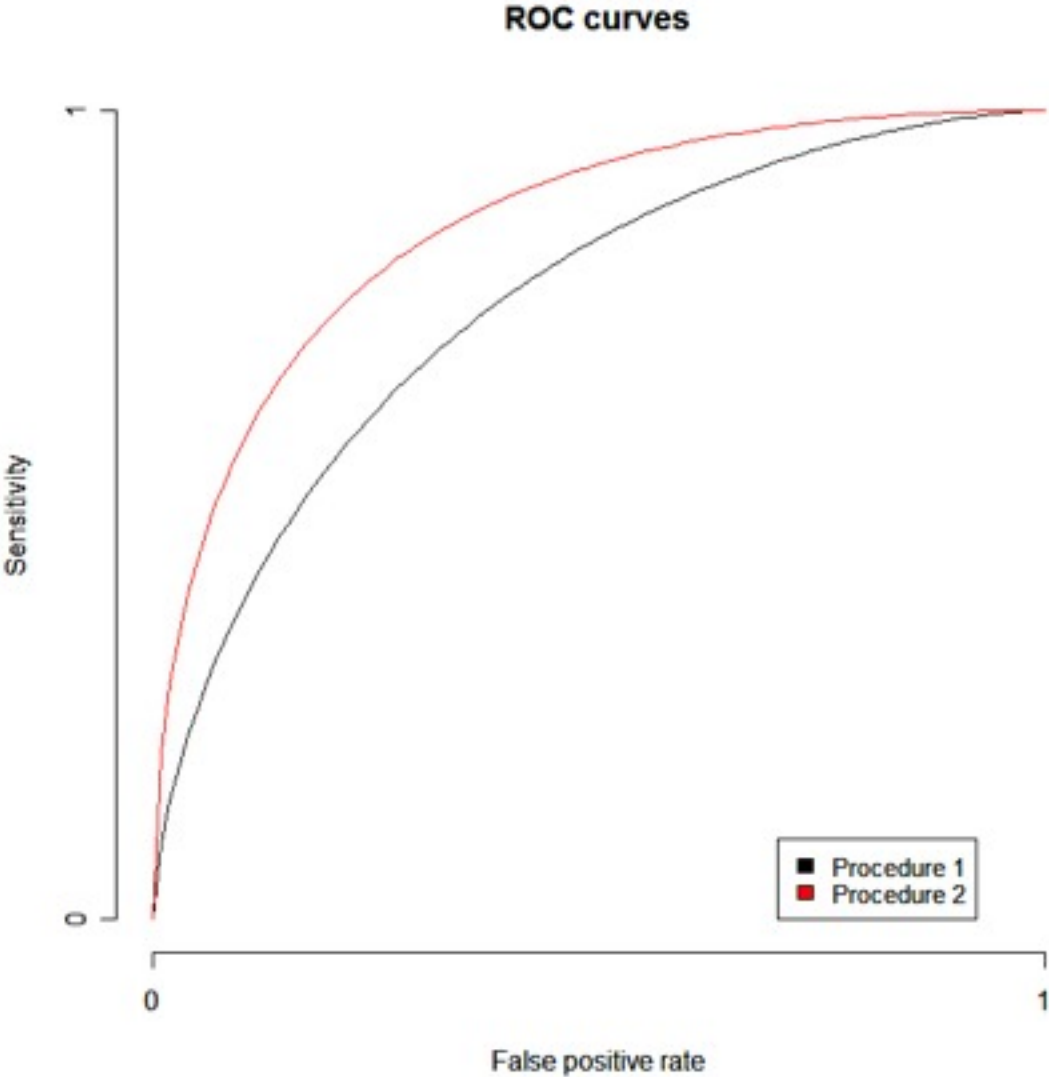
apple detection assay

**Sensitivity:**  
Probability that a detected object is really an apple. Estimated by  $TP / P$ .



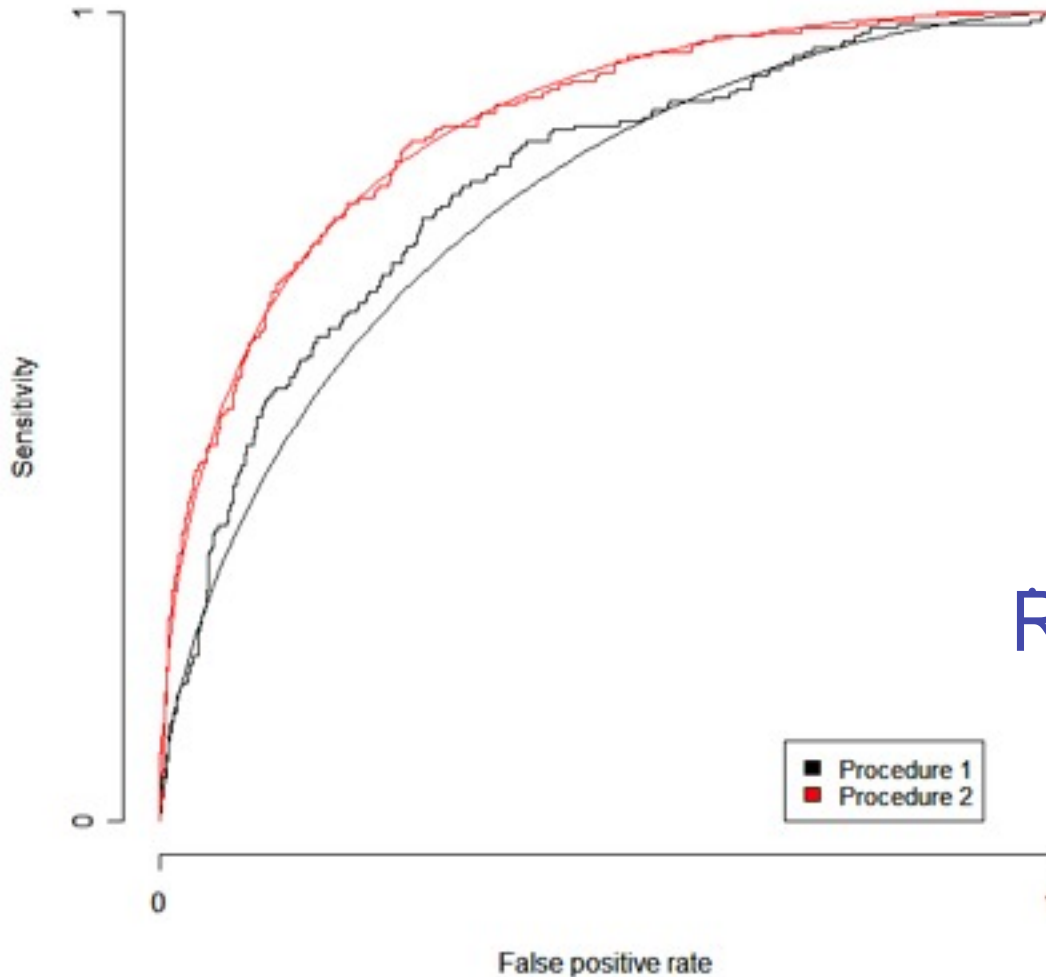
**Specificity:**  
Probability that a non-detected object is really not an apple. Estimated by  $TN / N$ .

# ROC curves for method comparison



# Empirical estimation of ROC curves

Estimated ROC curves (empirical CDFs)



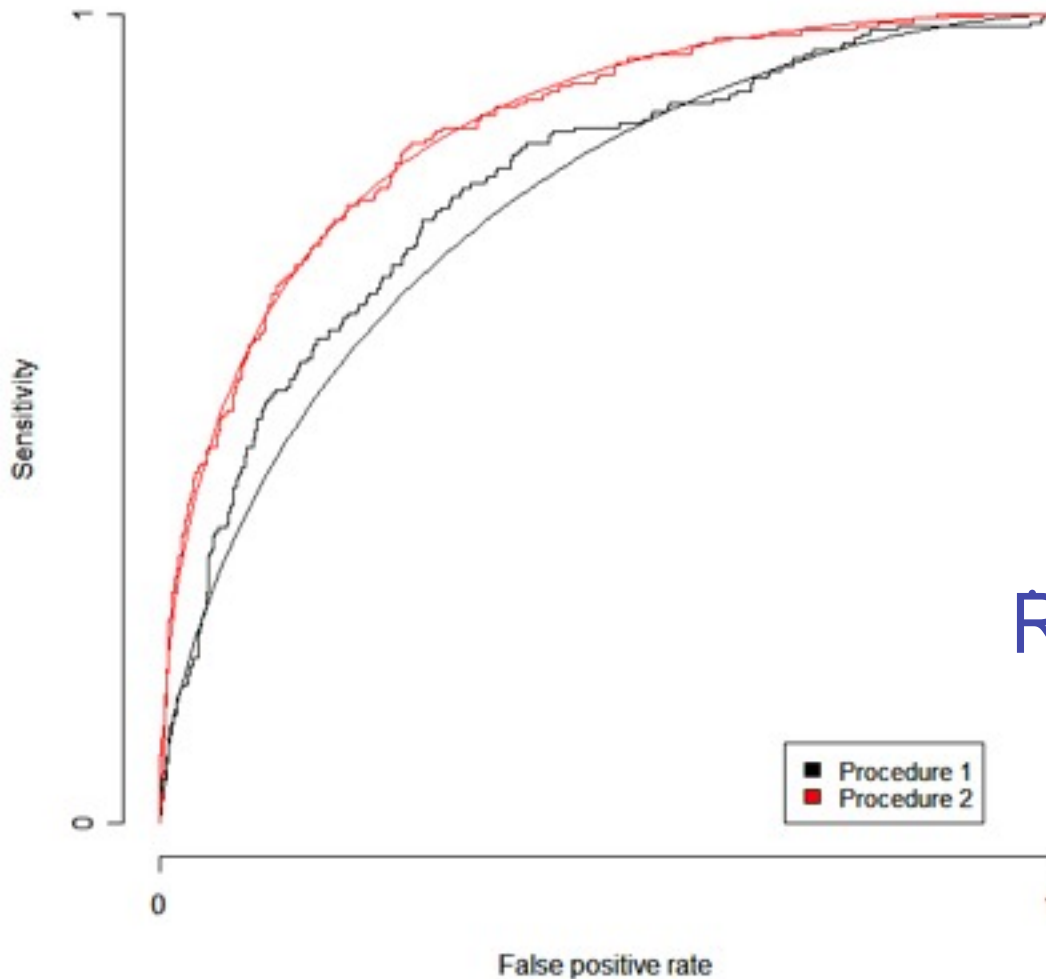
$$\hat{F}(t) = \frac{1}{m} \sum_{i=1}^m 1\{X_i \leq t\}$$

$$\hat{G}(t) = \frac{1}{n} \sum_{i=1}^n 1\{Y_i \leq t\}$$

$$\text{ROC} \equiv \left\{ \left( 1 - \hat{F}(t), 1 - \hat{G}(t) \right) \right\}_{t \in \mathcal{E}_i}$$

# Empirical estimation of ROC curves

Estimated ROC curves (empirical CDFs)



$$\hat{F}(t) = \frac{1}{m} \sum_{i=1}^m 1\{X_i \leq t\}$$

This assumes that we know the “ground truth”. Can we still do it if we don’t?

ROC

$t\}$

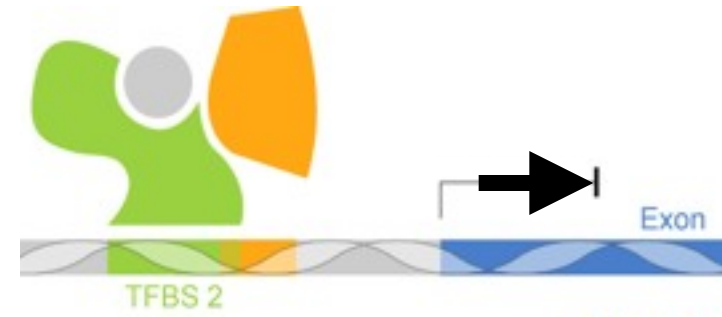
$t\}$



# Example: identification of transcription factor binding sites

$H_0 = \{\text{regions with no binding site}\}$

$H_1 = \{\text{regions with a binding site}\}$



**True positives?**

Small numbers of known sites for most factors.

Even the real sites are not active under all conditions.

**True negatives?**

Non-canonical / unexpected locations can hold real sites.

# True ROC curve

$H_0 = \{\text{regions with no binding site}\}$

$H_1 = \{\text{regions with a binding site}\}$

Test statistic	Set	Distribution
$X_1, \dots, X_m$	$H_0$	<b>F</b>
$Y_1, \dots, Y_n$	$H_1$	<b>G</b>

# “Pseudo-ROC” curve

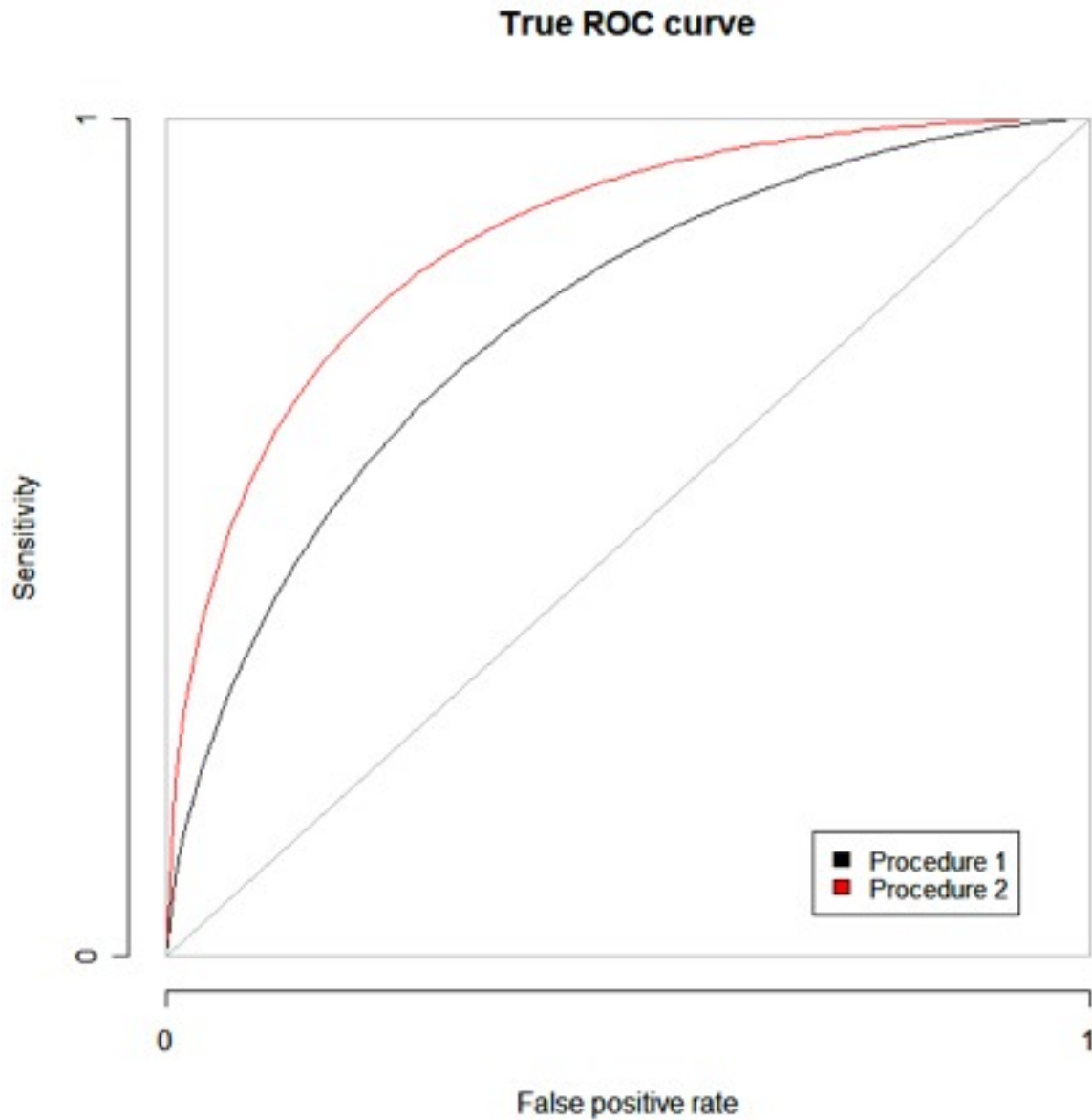
$S_0 = \{\text{regions less likely to have a binding site}\}$

$S_1 = \{\text{regions more likely to have a binding site}\}$

Test statistic	Set	Distribution
$X_1, \dots, X_m$	$H_0$	<b>F</b>
$Y_1, \dots, Y_n$	$H_1$	<b>G</b>
$X'_1, \dots, X'_m$	$S_0$	$(1 - \kappa)F + \kappa G$
$Y'_1, \dots, Y'_m$	$S_1$	$(1 - \lambda)F + \lambda G$

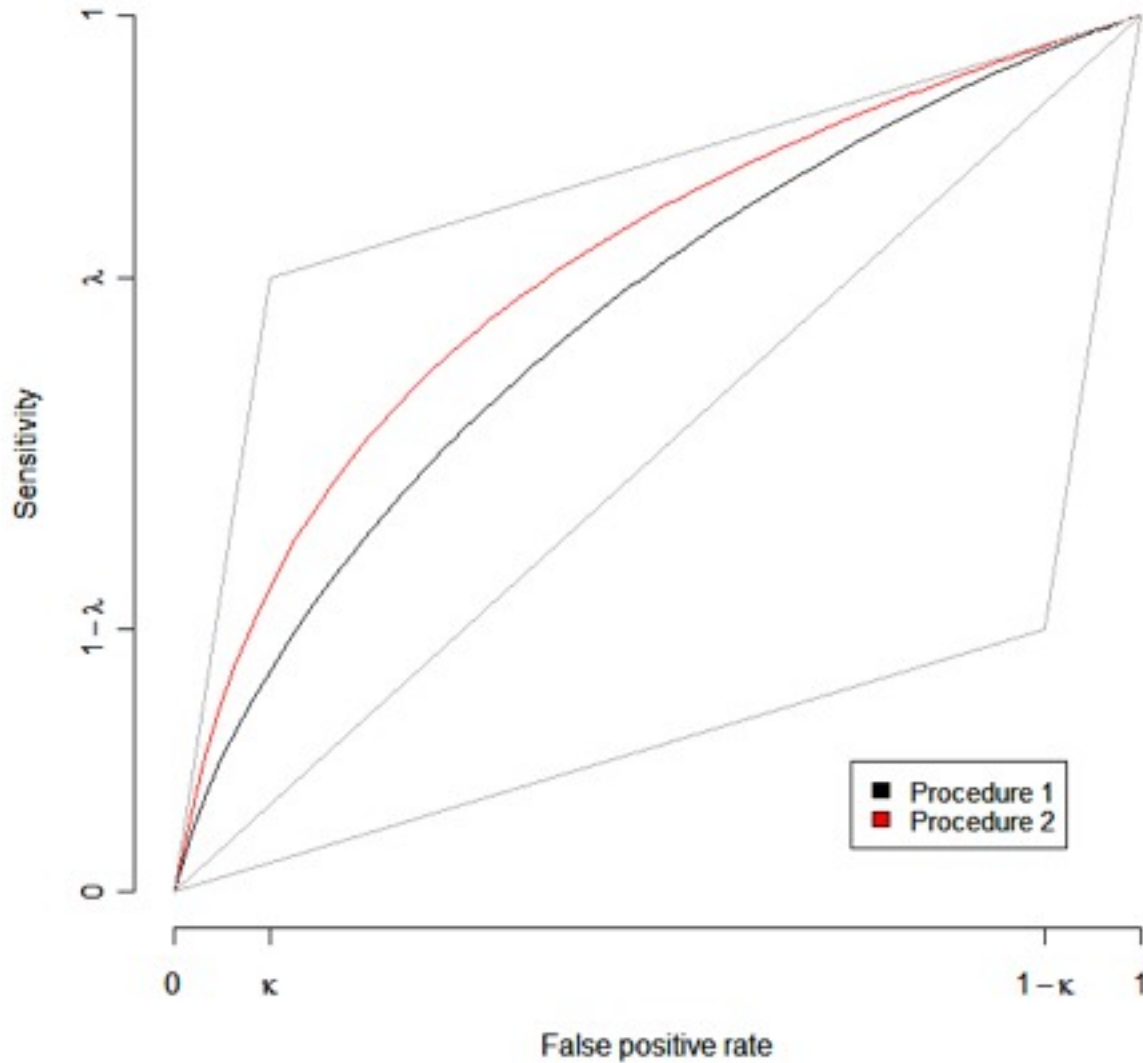
If  $\kappa = 0$  and  $\lambda = 1$ , test data are correctly classified.

# Correctly classified test data



# Contaminated test data

Pseudo-ROC curve



# Linear transform

$$\text{ROC}' = \left\{ (1 - F'(t), 1 - G'(t)) : t \in i \right\}$$

$$= \left\{ (1 - (1 - \kappa)F(t) - \kappa G(t), 1 - (1 - \lambda)F(T) - \lambda G(t)) : t \in i \right\}$$

$$= \left\{ (1 - F(t), 1 - G(t)) \begin{pmatrix} 1 - \kappa & 1 - \lambda \\ \kappa & \lambda \end{pmatrix} : t \in i \right\}$$

$$= \left\{ (p, q) \begin{pmatrix} 1 - \kappa & 1 - \lambda \\ \kappa & \lambda \end{pmatrix} : (p, q) \in \text{ROC} \right\}$$

# Comparing two methods

$$\text{ROC}' = \left\{ (p, q) \begin{pmatrix} 1-\kappa & 1-\lambda \\ \kappa & \lambda \end{pmatrix} : (p, q) \in \text{ROC} \right\}$$

The transformation depends on the contamination fractions only, not  $F_1$  and  $G_1$ , or  $F_2$  and  $G_2$ .

Assuming  $\kappa < \lambda$ , the transformation preserves the ordering of curves and of the area under them (AUC).

The area between (and under) the curves is compressed — more severely as  $\lambda \rightarrow 0$  or  $\kappa \rightarrow 1$ .

# Summary

If, for both procedures being compared,

- correctly and incorrectly classified true positives have the same statistical properties, and
- correctly and incorrectly classified true negatives have the same statistical properties, then

the pseudo-ROC and true ROC select the same procedure as superior.



# Multiple testing

Many data analysis approaches in genomics rely on item-by-item (i.e. multiple) testing:

Microarray or RNA-Seq expression profiles of “normal” vs “perturbed” samples: gene-by-gene

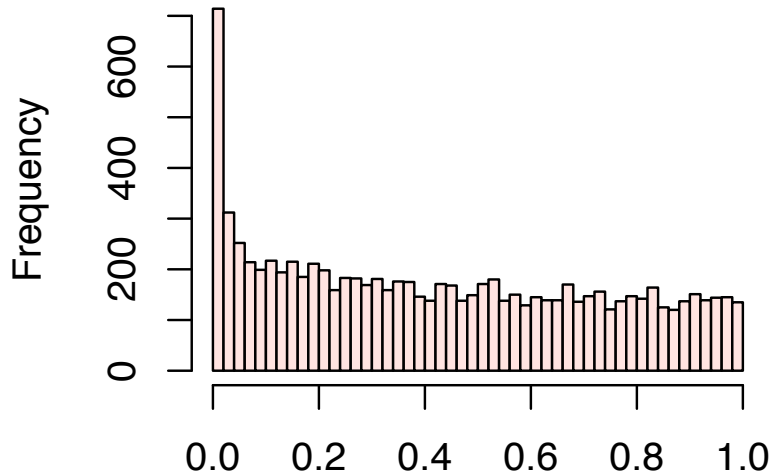
ChIP-chip: locus-by-locus

RNAi and chemical compound screens

Genome-wide association studies: marker-by-marker

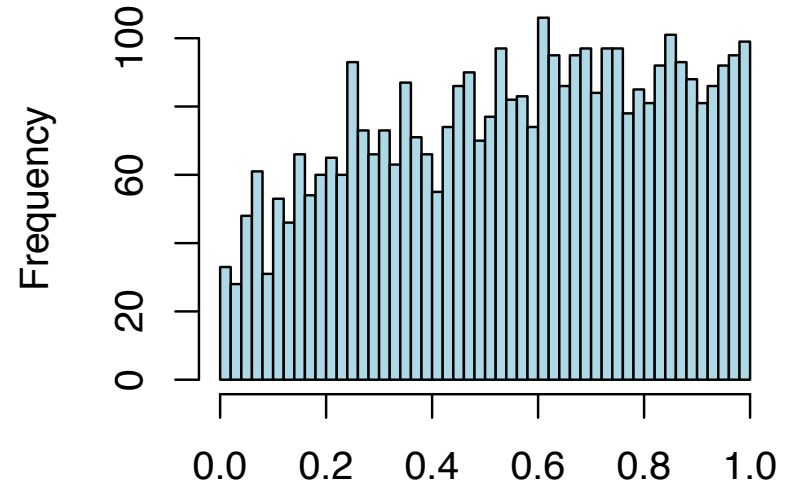
QTL analysis: marker-by-marker and trait-by-trait

# Diagnostic plot: the histogram of p-values



Observed p-values are a mix of samples from

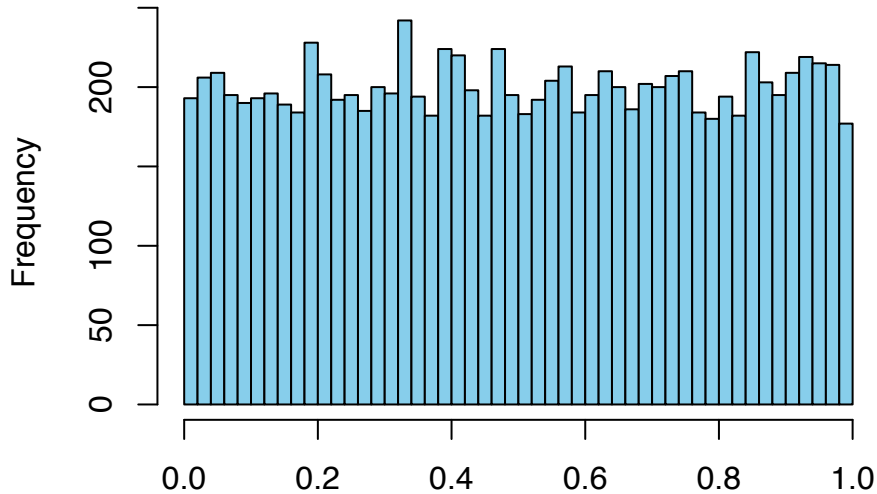
- a uniform distribution (from true nulls) and
- from distributions concentrated at 0 (from true alternatives)



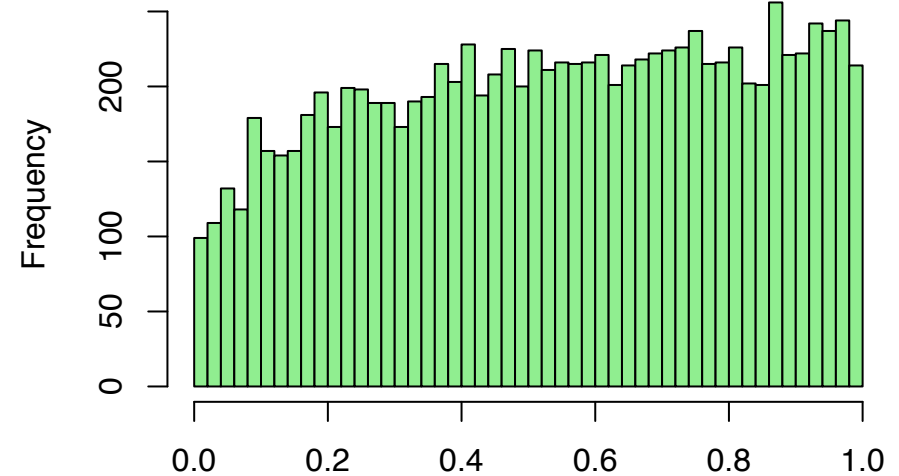
Depletion of small p can indicate the presence of confounding hidden variables (“batch effect”)

# Batch effects or “latent variables”

Histogram of `rt1$p.value`



Histogram of `rt2$p.value`



**n = 10000**

**m = 20**

```
x = matrix(rnorm(n*m), nrow=n, ncol=m)
```

```
fac = factor(c(rep(0, 10), rep(1, 10)))
```

```
rt1 = rowttests(x, fac)
```

```
x[, 6:15] = x[, 6:15]+1
```

```
rt2 = rowttests(x, fac)
```

*sva* package; Leek JT, Storey JD.  
Capturing heterogeneity in gene  
expression studies by surrogate  
variable analysis. PLoS Genet. 2007

Stegle O, Parts L, Durbin R, Winn J.  
A Bayesian framework to account for  
complex non-genetic factors in gene  
expression levels greatly increases  
power in eQTL studies. PLoS Comput  
Biol. 2010.

# Multiple testing

## Classical hypothesis test:

null hypothesis  $H_0$ , alternative  $H_1$

test statistic  $X \mapsto t(X) \in \mathbb{R}$

$\alpha = \mathbf{P}(t(X) \in \Gamma_{\text{rej}} \mid H_0)$  type I error (false positive)

$\beta = \mathbf{P}(t(X) \notin \Gamma_{\text{rej}} \mid H_1)$  type II error (false negative)

When  $n$  tests are performed, what is the extent of type I errors, and how can it be controlled?

E.g.: 20,000 tests at  $\alpha=0.05$ , all with  $H_0$  true: expect 1,000 false positives

# Experiment-wide type I error rates

	Not rejected	Rejected	Total
True null hypotheses	U	V	$m_0$
False null hypotheses	T	S	$m_1$
Total	$m - R$	R	m

**Family-wise error rate:**  $P(V > 0)$ , the probability of one or more false positives. For large  $m_0$ , this is difficult to keep small.

**False discovery rate:**  $E[ V / \max\{R, 1\} ]$ , the expected fraction of false positives among all discoveries.

## FWER: The Bonferroni correction

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

an observed test statistic:  $T_g$

an unadjusted  $p$ -value:  $p_g$ .

Bonferroni adjusted  $p$ -values:

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

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

## Controlling the FDR (Benjamini/Hochberg)

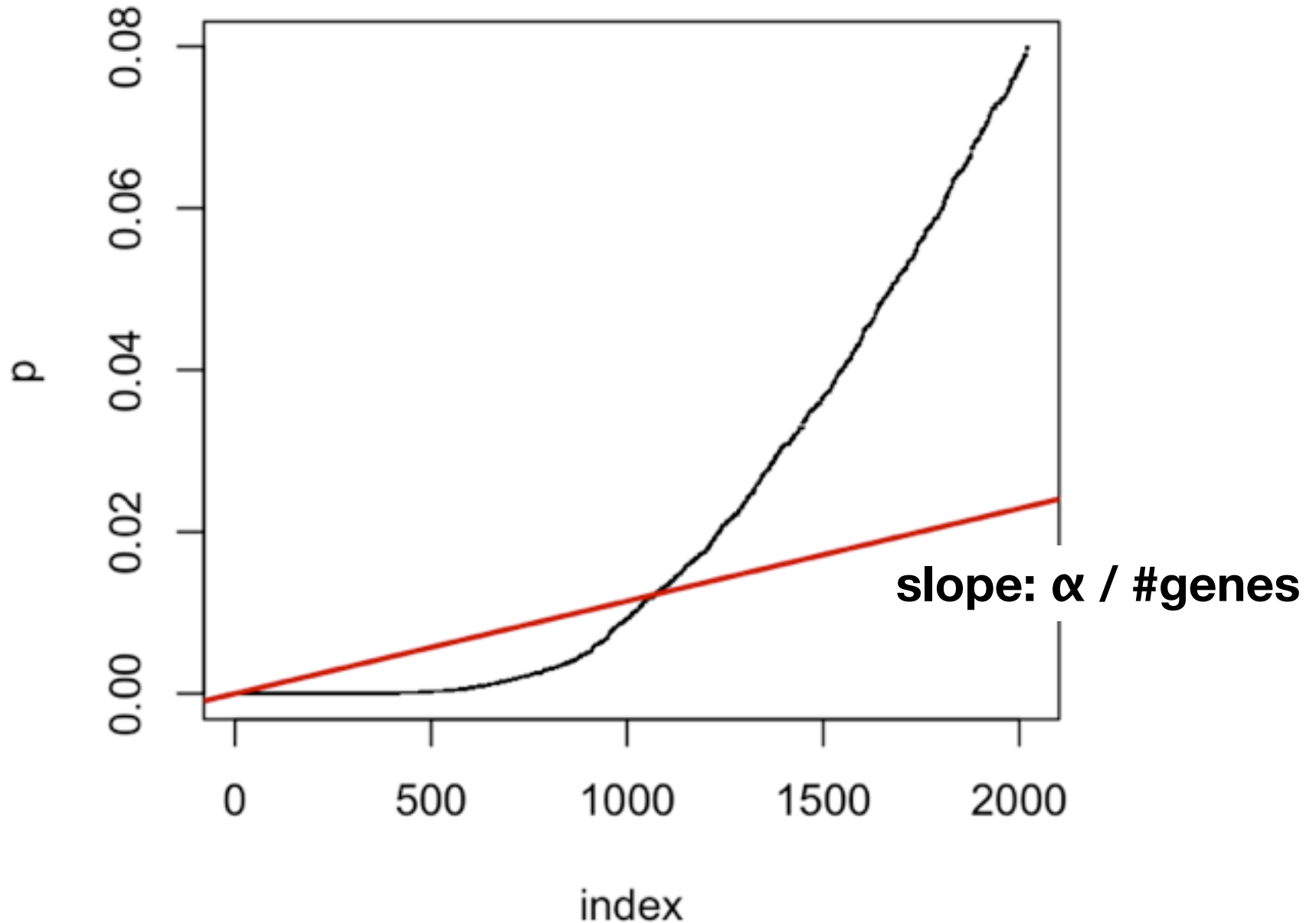
- FDR: the expected proportion of false positives among the significant genes.
- Ordered unadjusted  $p$ -values:  $p_{r_1} \leq p_{r_2} \leq \dots \leq p_{r_m}$ .
- To control  $FDR = E(V/R)$  at level  $\alpha$ , let

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

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

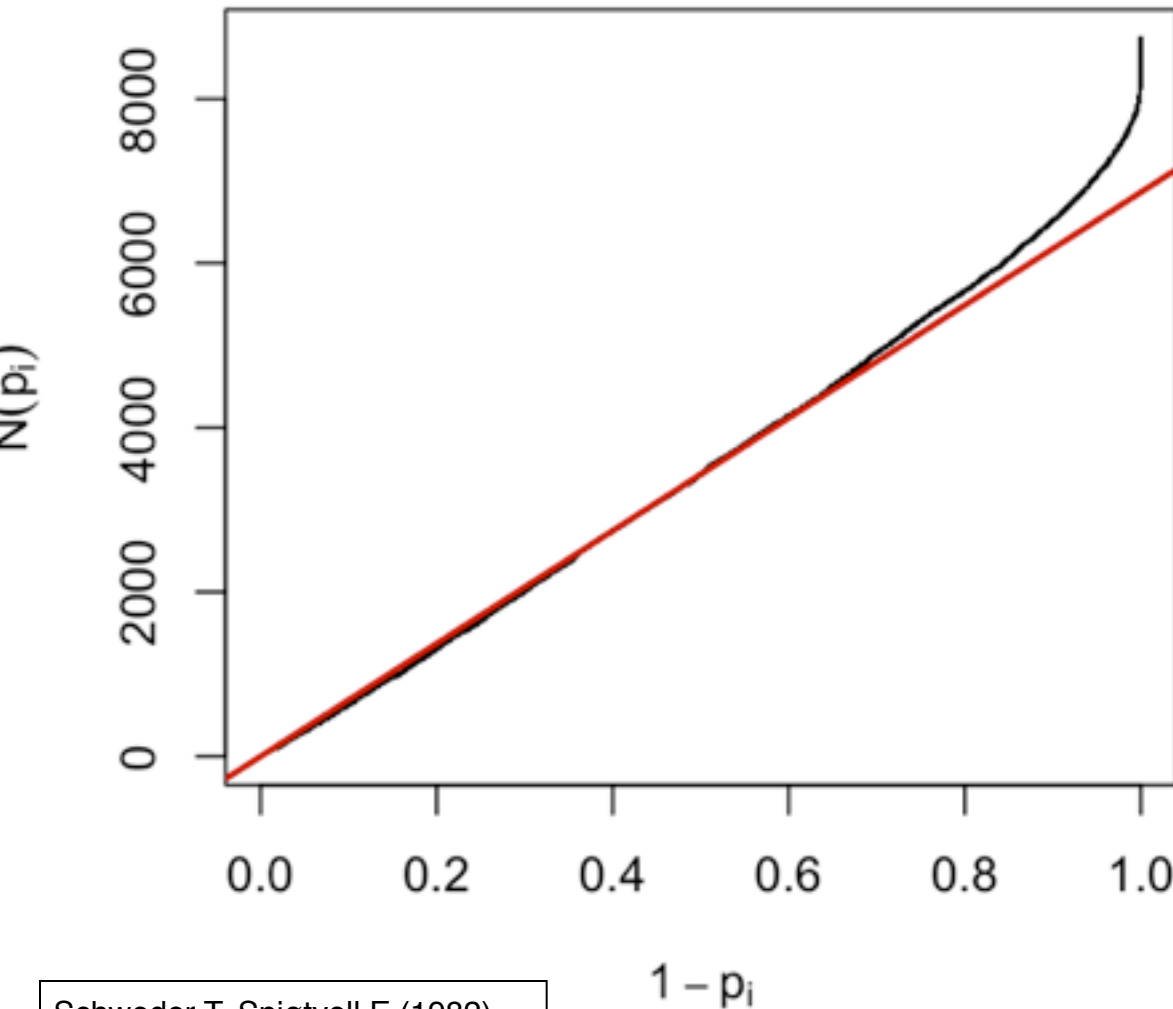
- Is valid for independent test statistics and for some types of dependence.

# Benjamini Hochberg multiple testing adjustment





# Schweder and Spjøtvoll p-value plot



For a series of hypothesis tests  $H_1 \dots H_m$  with p-values  $p_i$ , plot

$(1 - p_i, N(p_i))$  for all  $i$

where  $N(p)$  is the number of p-values greater than  $p$ .

Schweder T, Spjøtvoll E (1982)  
Plots of P-values to evaluate  
many tests simultaneously.  
Biometrika 69:493–502.

# Example: differential expression testing

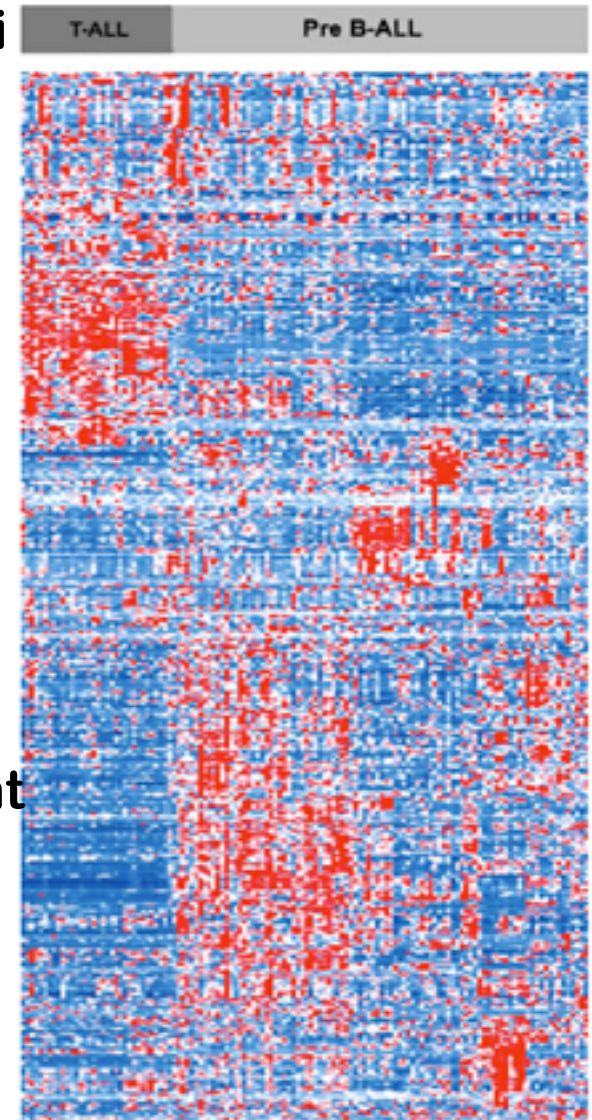
Acute lymphocytic leukemia (ALL) data, Chiaretti et al., Clinical Cancer Research 11:7209, 2005

Immunophenotypic analysis of cell surface markers identified

- T-cell derivation in 33,
- B-cell derivation in 95 samples

Affymetrix HG-U95Av2 3' transcript detection arrays with ~13,000 probe sets

Chiaretti et al. selected probesets with “sufficient levels of expression and variation across groups” and among these identified 792 differentially expressed genes.



*Clustered expression data for all 128 subjects, and a subset of 475 genes showing evidence of differential expression between groups*

# Independent filtering

From the set of 13,000 probesets,

**first** filter out those that seem to report negligible signal (say, 40%),

**then** formally test for differential expression on the rest.

Conditions under which we expect negligible signal :

1. Target gene is absent in both samples. (Probes will still report noise and cross-hybridization.)
2. Probe set fails to detect the target.

Literature: von Heydebreck et al. (2004)

McClintick and Edenberg (BMC Bioinf. 2006) and references therein

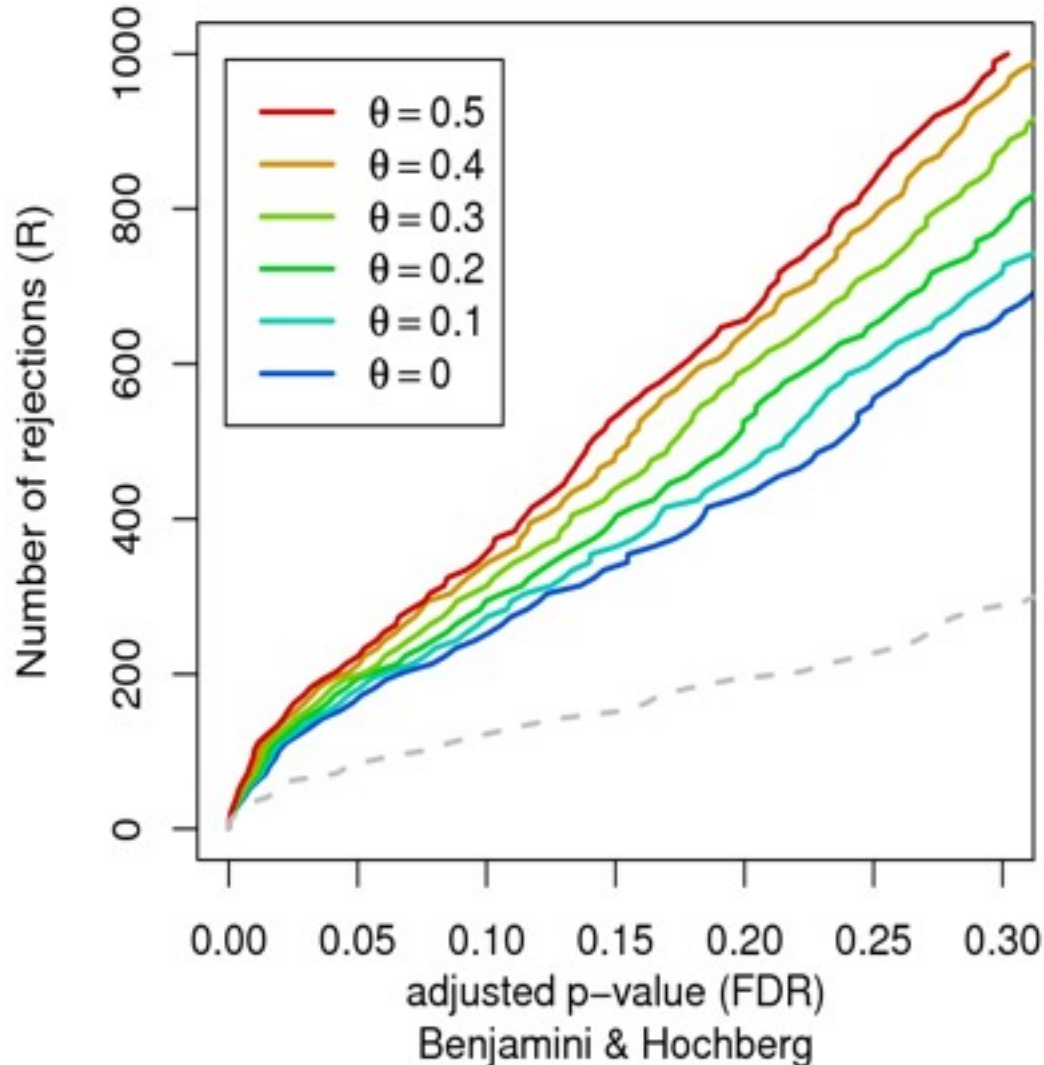
Hackstadt and Hess (BMC Bioinf. 2009)

Many others.

# Increased detection rates

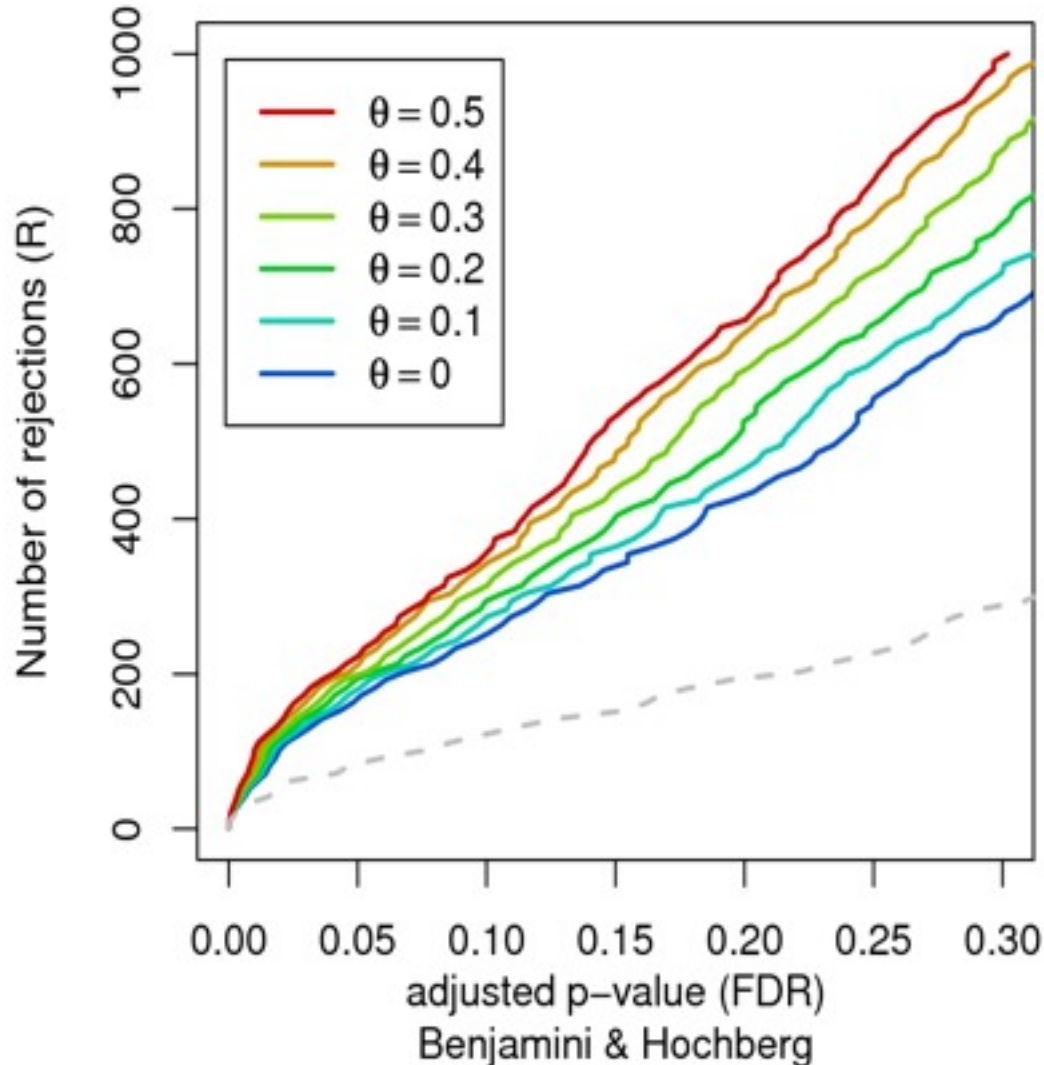
Stage 1 filter: compute variance, across samples, for each probeset, and remove the fraction  $\theta$  that are smallest

Stage 2: standard two-sample t-test



# Increased power?

Increased detection rate implies increased power only if we are still controlling type I errors at the same level as before.



# Increased power?

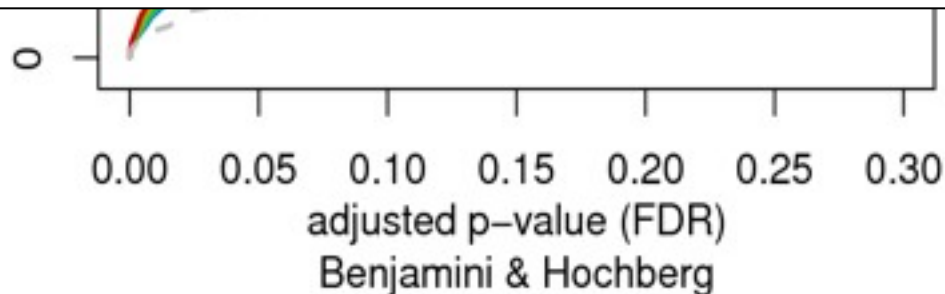
Increased detection rate implies increased power  
only if we are still controlling type I errors at the same level as  
before.

## Concerns:

- Have we thrown away good genes?
- Use a data-driven criterion in stage 1, but do type I error consideration only on number of genes in stage 2

Informal justification:

Filter does not use covariate information



# What do we need for type I error control?

- I. For each individual (per gene) test statistic, we need to know its correct null distribution
- II. If and as much as the multiple testing procedure relies on certain (in)dependence structure between the different test statistics, our test statistics need to comply.

I.: one (though not the only) solution is to make sure that by filtering, the null distribution is not affected - that it is the same before and after filtering

II.: See later

# Result: independence of stage 1 and stage 2 statistics under the null hypothesis

For genes for which the null hypothesis is true ( $X_1, \dots, X_n$  exchangeable),  $f$  and  $g$  are statistically independent in both of the following cases:

- **Normally distributed data:**

$f$  (stage 1): overall variance (or mean)

$g$  (stage 2): the standard two-sample t-statistic, or any test statistic which is scale and location invariant.

- **Non-parametrically:**

$f$ : any function that does not depend on the order of the arguments. E.g. overall variance, IQR.

$g$ : the Wilcoxon rank sum test statistic.

Both can be extended to the multi-class context: ANOVA and Kruskal-Wallis.



# Derivation

## Non-parametric case:

Straightforward decomposition of the joint probability into product of probabilities using the assumptions.

## Normal case:

Use the spherical symmetry of the joint distribution,  $p$ -dimensional  $N(0, 1\sigma^2)$ , and of the overall variance; and the scale and location invariance of  $t$ .

This case is also implied by Basu's theorem

( $V$  complete sufficient for family of probability measures  $P$ ,  $T$  ancillary  $\Rightarrow T, V$  independent)

# What do we need for type I error control?

The distribution of the test statistic under the null.

I. **Marginal**: for each individual (per gene) test statistic

II. **Joint**: some (though not all) multiple testing procedures relies on certain independence properties of the joint distribution

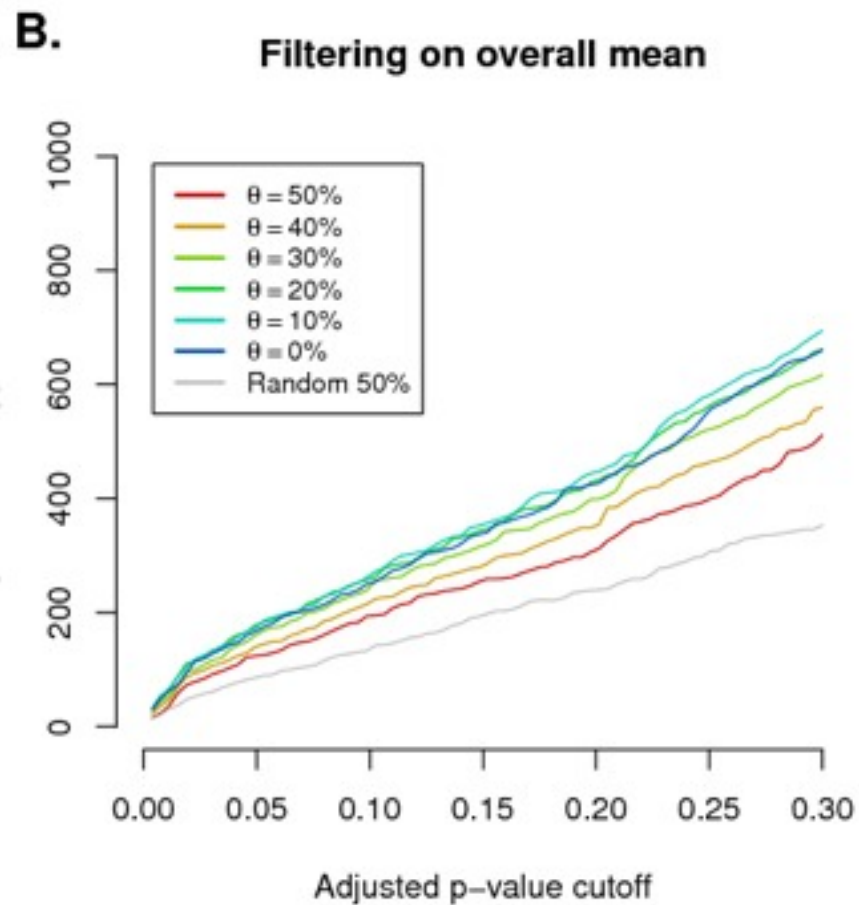
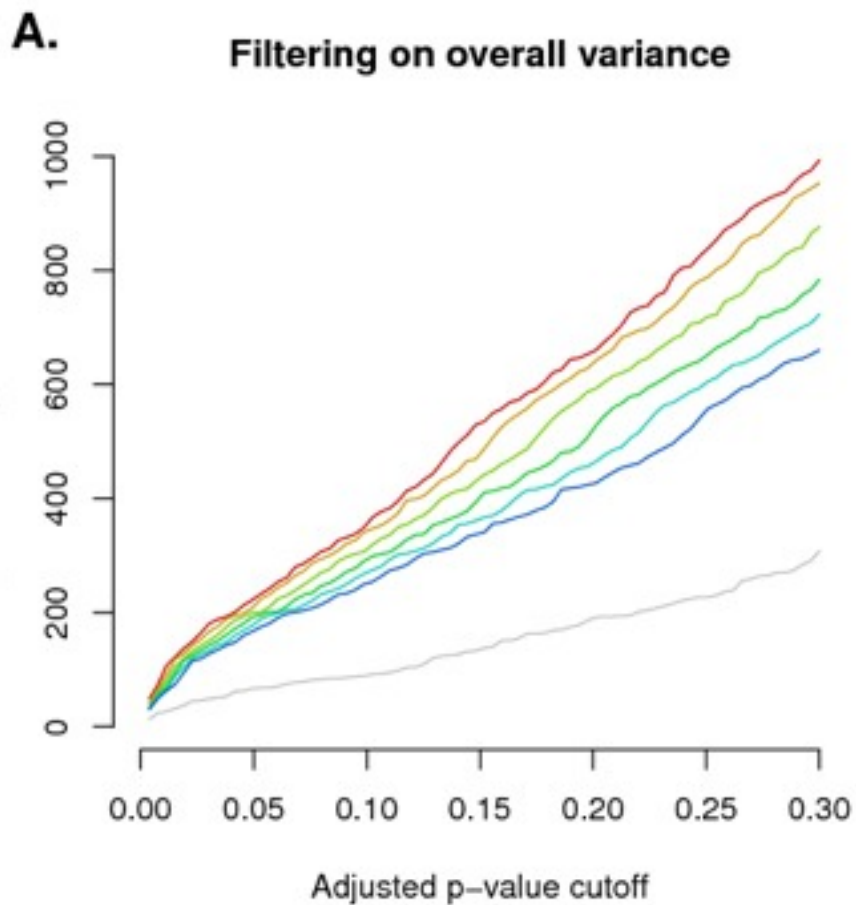
I.: one (though not the only) solution is to make sure that by filtering, the marginal null distribution is not affected - that it is the same before and after filtering



# Multiple testing procedures and dependence

1. **Methods that work on the p-values only and allow general dependence structure: Bonferroni, Bonferroni-Holm (FWER), Benjamini-Yekutieli (FDR)**
2. **Those that work on the data matrix itself, and use permutations to estimate null distributions of relevant quantities (using the empirical correlation structure): Westfall-Young (FWER)**
3. **Those that work on the p-values only, and make dependence-related assumptions: Benjamini-Hochberg (FDR), q-value (FDR)**

# Now we are confident about type I error, but does it do any good? (power)

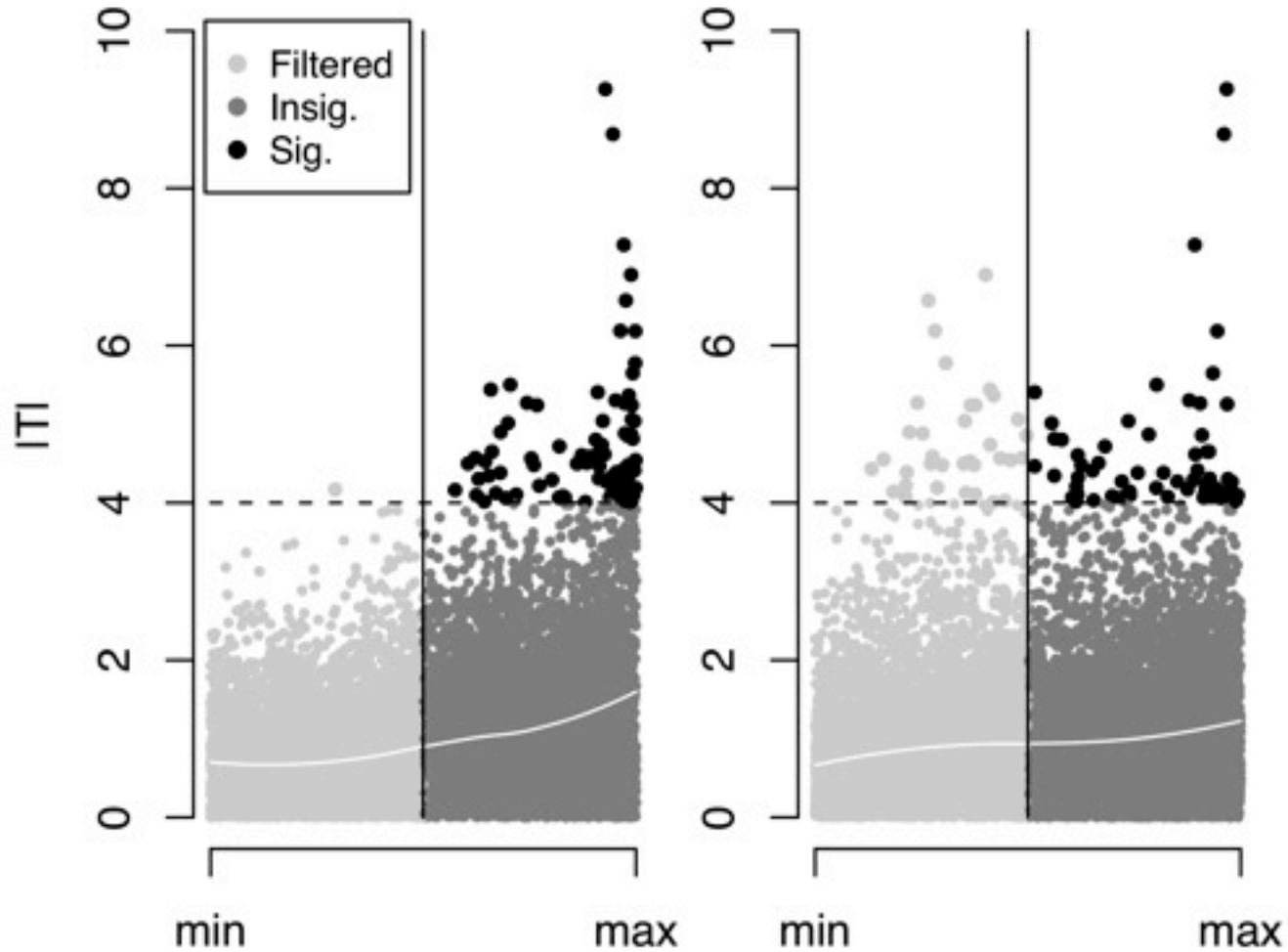


# Diagnostics

D.

Overall variance

Overall mean

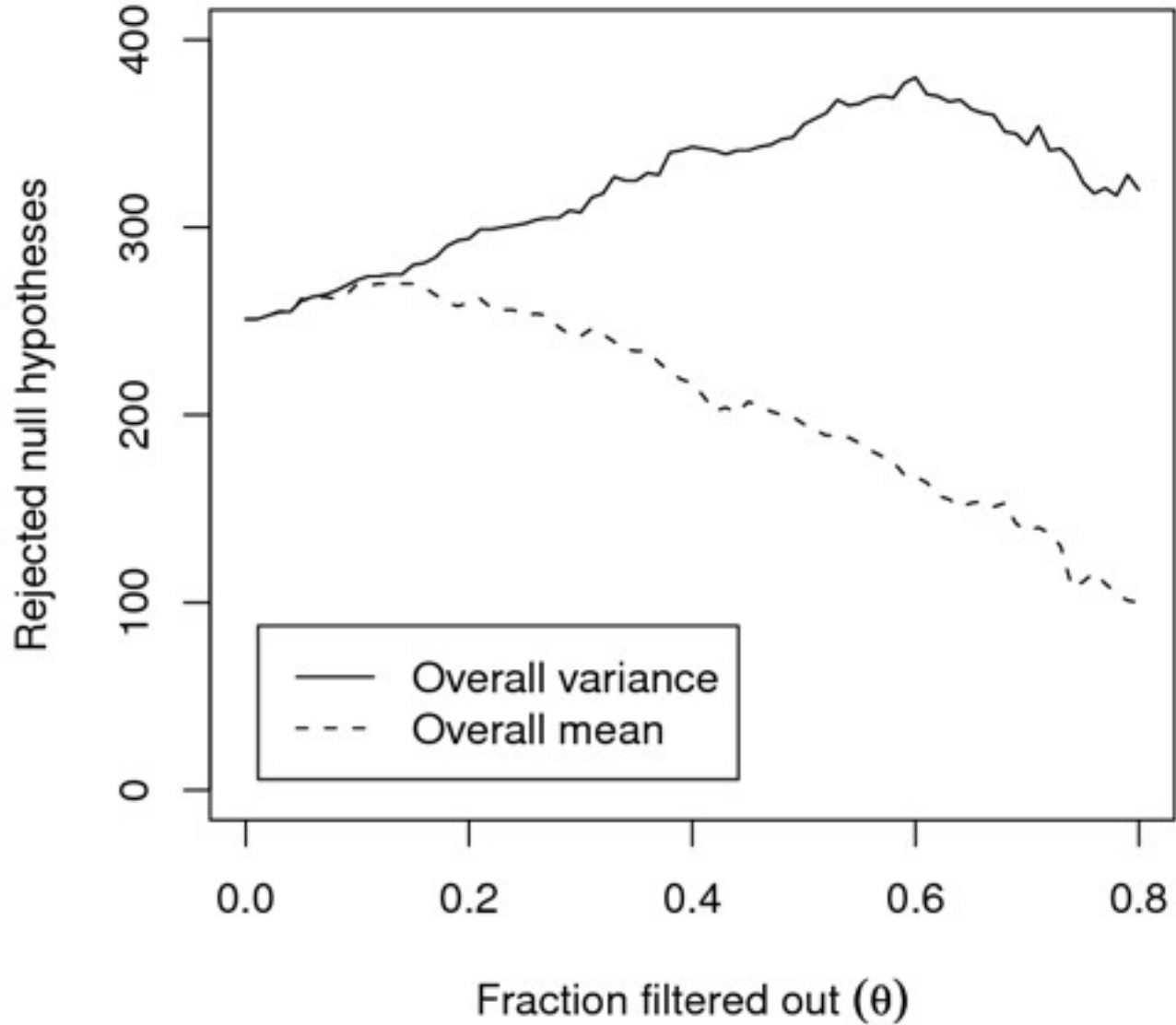


Rank of filter statistic

$\theta$

**C.**

**Rejections, for adjusted  $p < 0.10$**



# For count data (DESeq)

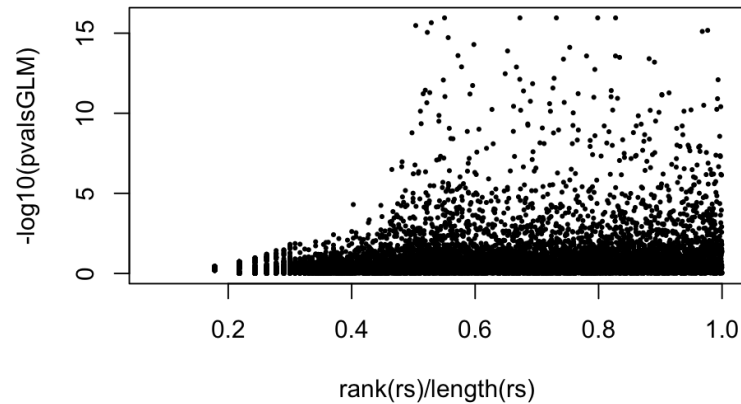


Figure 9: Scatterplot of rank of filter criterion (overall sum of counts  $rs$ ) versus the negative logarithm of the test statistic  $pvalsGLM$ .

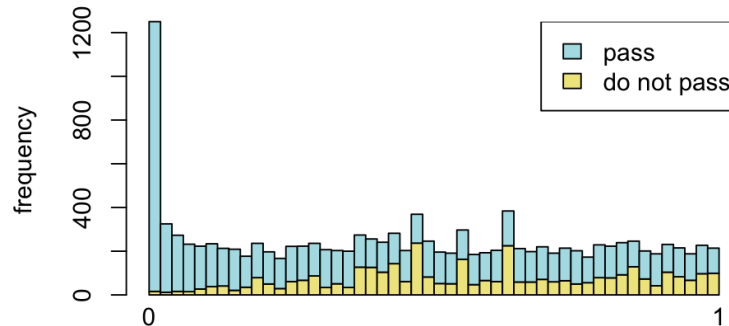


Figure 10: Histogram of  $p$  values for all tests ( $pvalsGLM$ ). The area shaded in blue indicates the subset of those that pass the filtering, the area in khaki those that do not pass.

# Results summary

If done improperly, "filtering" invalidates type-I error control.

One way to do it properly is to make sure that stage-one (filter) and stage-two (differential expression) statistics are **marginally independent**:

1. (Normal distributed data): overall variance or mean, followed by t-test
2. Any permutation invariant statistic, followed by Wilcoxon rank sum test

Marginal independence is sufficient to maintain control of FWER at nominal level.

Control of FDR is usually also maintained.

(It could in principle be affected by filter-induced changes to correlation structure of the data. Check your data for indications of that. We have never seen it to be a problem in practice.)



# Conclusion

Correct use of this two-stage approach can substantially increase power at same type I error.

# Conclusion

Correct use of this two-stage approach can substantially increase power at same type I error.



# References

**Bourgon R., Gentleman R. and Huber W. Independent filtering increases detection power for high-throughput experiments, PNAS (2010)**

**Bioconductor package `genefilter` vignette**

**DESeq vignette**

**On pseudo-ROC: Richard Bourgon's PhD thesis**

**Simon Anders**  
**Richard Bourgon**  
**Bernd Fischer**  
**Gregoire Pau**

**Robert Gentleman**, F. Hahne, M.  
Morgan (FHCRC)

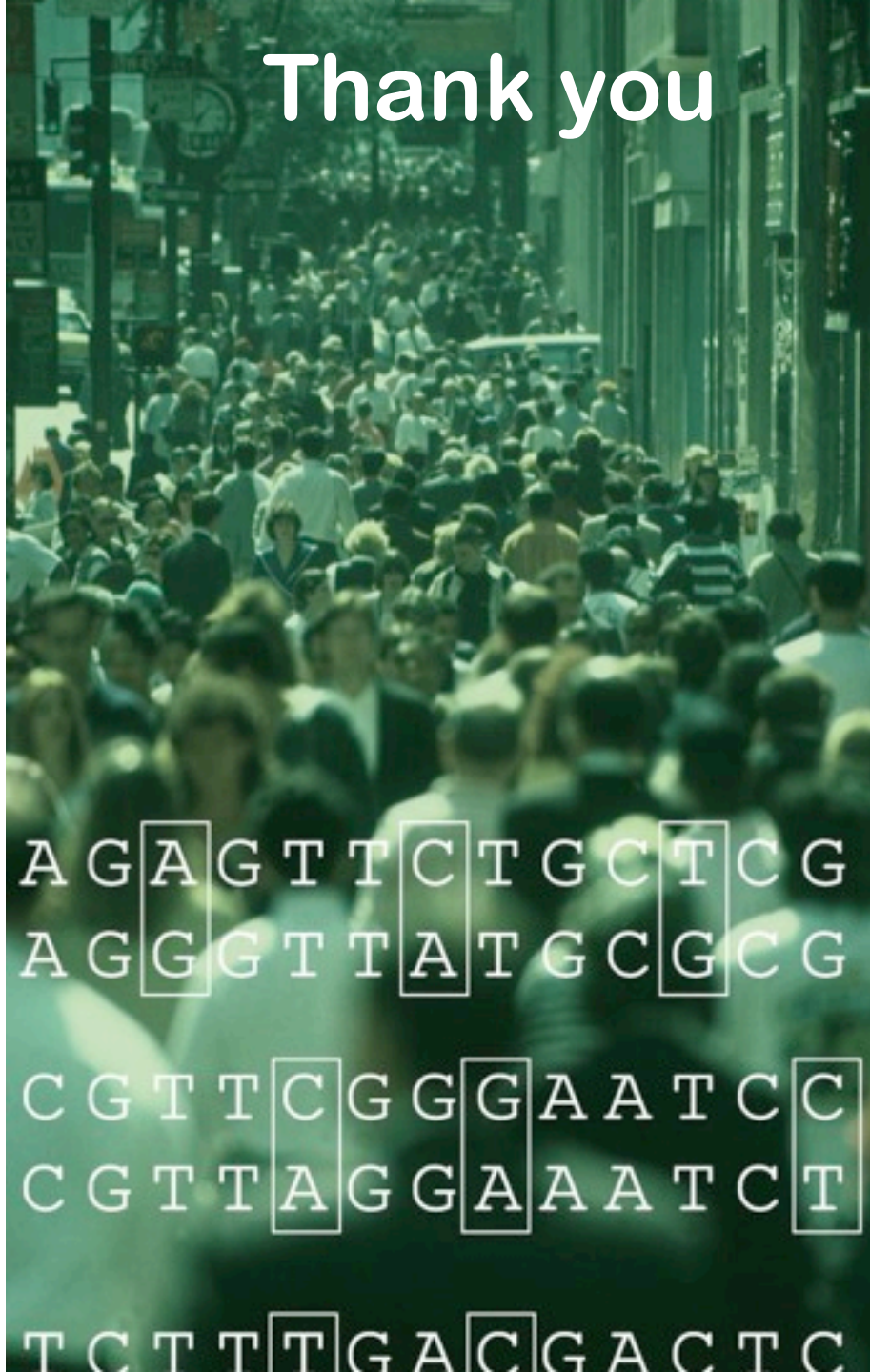
Lars Steinmetz, J. Gagneur, Z. Xu, W.  
Wei (EMBL)

Michael Boutros, F. Fuchs, D.  
Ingelfinger, T. Horn, T. Sandmann  
(DKFZ)

Steffen Durinck (Illumina)

All contributors to the R and  
Bioconductor projects

**Thank you**



A G A G T T C T G C T C G  
A G G G T T A T G C G C G  
C G T T C G G G A A T C C  
C G T T A G G A A A T C T  
T C T T T G A C G A C T C

# Derivation (non-parametric case)

$$P(f \in A, g \in B)$$

A, B: measurable sets  
f: stage 1, g: stage 2

$$= \int_{i^n} \delta_A(f(X)) \delta_B(g(X)) dP_X$$

exchangeability

$$= \frac{1}{n!} \sum_{\pi \in \Pi_n} \int_{i^n} \delta_A(f \circ \pi(X)) \delta_B(g \circ \pi(X)) dP_X$$

f's permutation invariance

$$= \int_{i^n} \delta_A(f(X)) \left( \frac{1}{n!} \sum_{\pi \in \Pi_n} \delta_B(g \circ \pi(X)) \right) dP_X$$

distribution of g generated  
by permutations

$$= \int_{i^n} \delta_A(f(X)) P(g \in B) dP_X$$

$$= P(f \in A) \cdot P(g \in B) \quad \#$$

# Positive Regression Dependency

On the subset of true null hypotheses:

If the test statistics are  $X = (X_1, X_2, \dots, X_m)$ :

For any increasing set  $D$  (the product of rays, each infinite on the right), and  $H_{0i}$  true, require that

$\text{Prob}(X \text{ in } D \mid X_i = s)$  is increasing in  $s$ , for all  $i$ .

**Important Examples**

**Multivariate Normal with positive correlation**

**Absolute Studentized independent normal**