Multiple testing & Independent filtering

Wolfgang Huber, EMBL

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

(You can also think of this as an extreme form of regularisation)

Statistics 101

←bias accuracy→

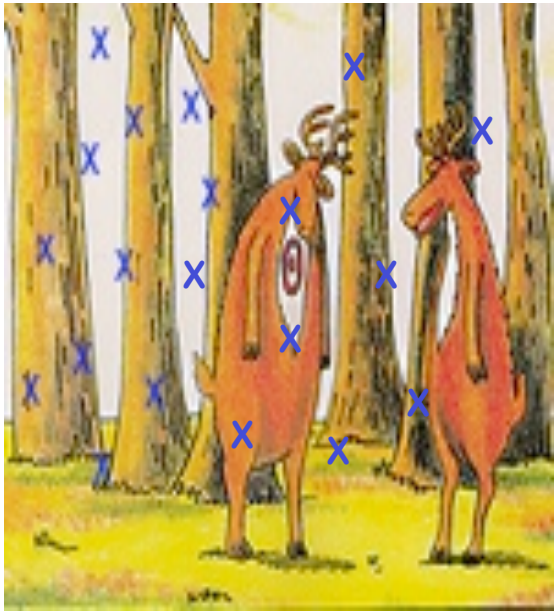
← precision dispersion→

Basic dogma of data analysis

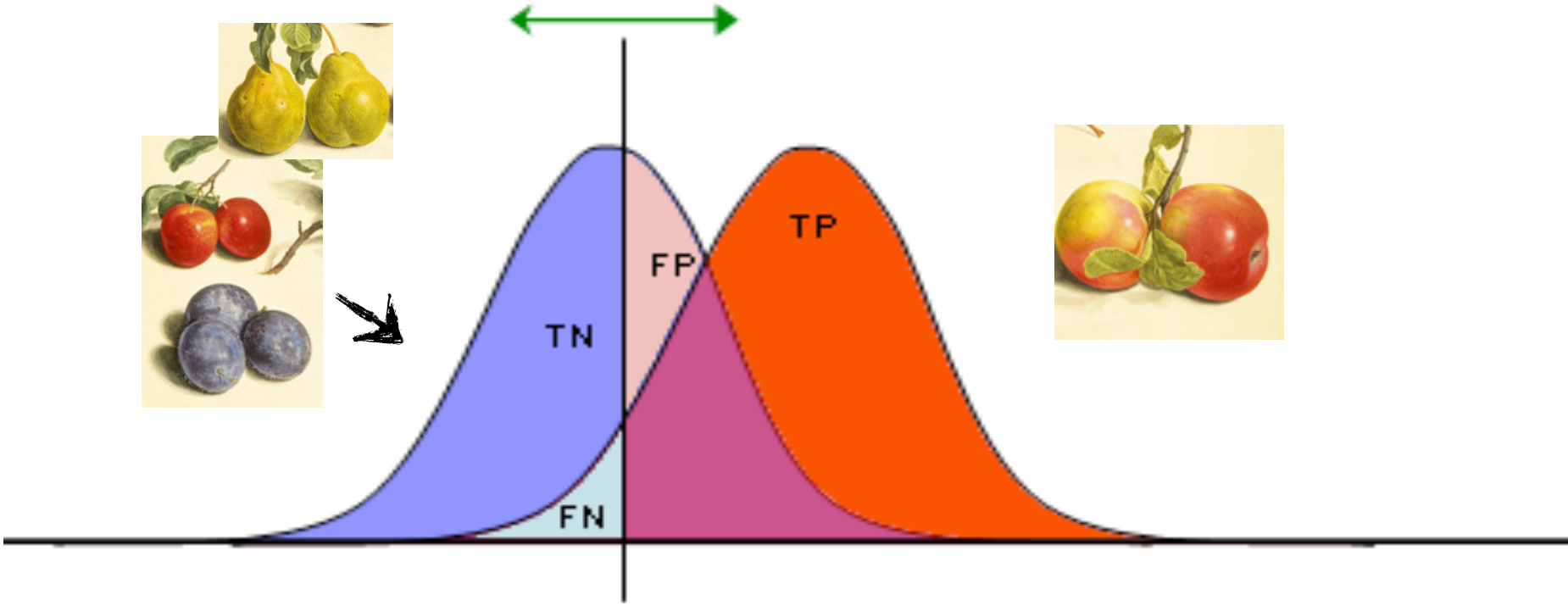
Can always increase sensitivity on the cost of specificity, or vice versa, the art is to

- optimize both

- find the best trade-off



Testing vs classification



Testing

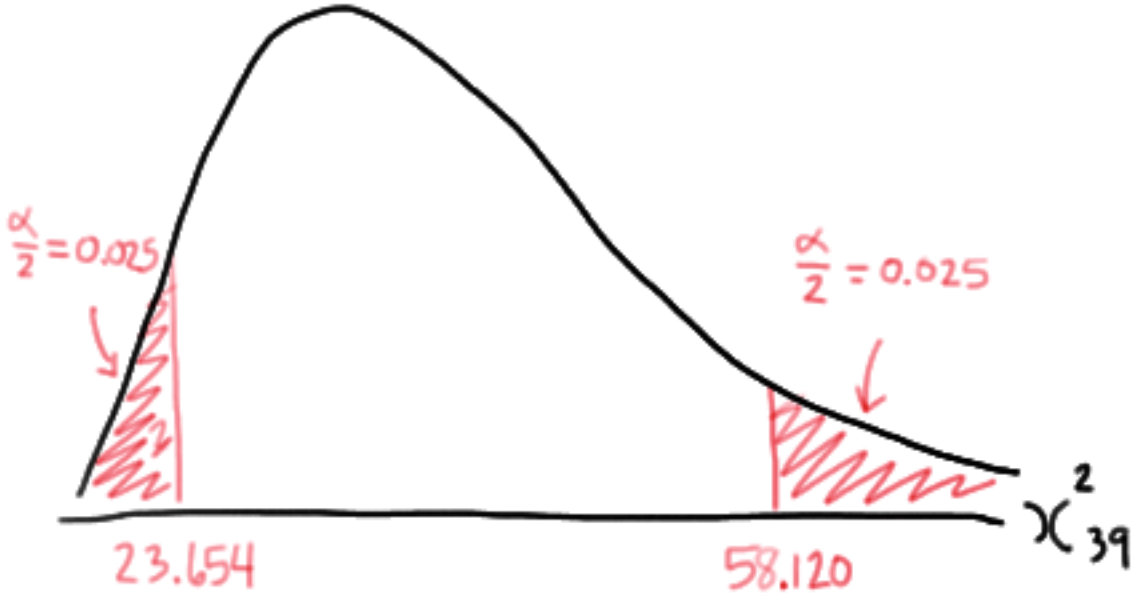
Classical hypothesis test:

null hypothesis H_0

test statistic: data $\mapsto$ real number t

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

$\beta = P(t \notin \Gamma_{rej} \mid H_0 \text{ false})$ type II error (false negative)



Avoid fallacy

The p-value is the probability of seeing a result as extreme or more extreme than the observed data, when the null hypothesis is true.

It it not the probability that the null hypothesis is true.

Absence of evidence ≠ evidence of absence

Multiple Testing

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 α=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 $g = 1, \ldots, m$, producing

an observed test statistic: T_g

an unadjusted p–value: p_g.

Bonferroni adjusted p–values:

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

Selecting all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α, 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: $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m}$.

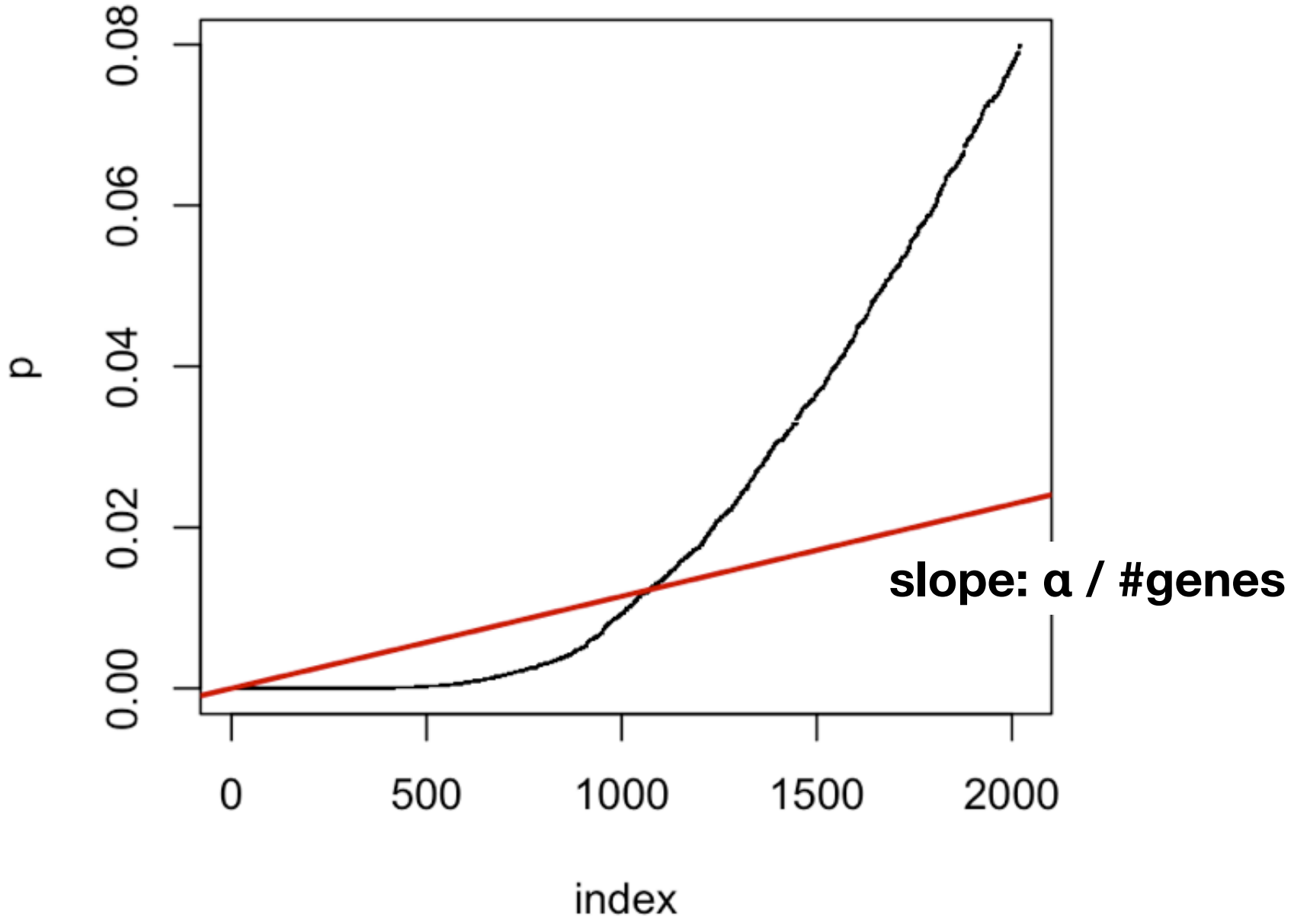
- To control $FDR = E(V/R)$ at level α, let

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

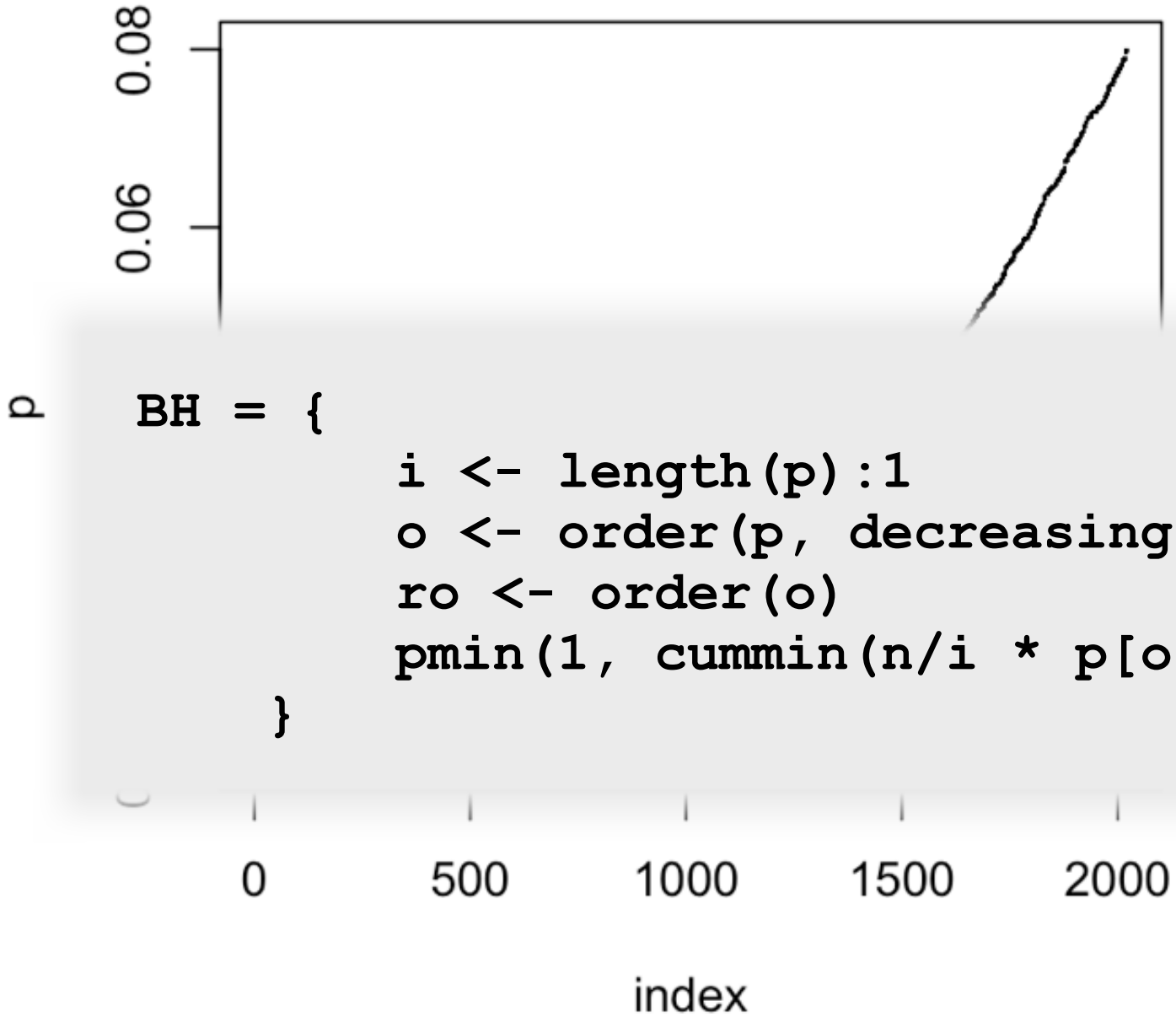
Reject the hypotheses H_{r_j} for $j = 1, \ldots, j^\star$.

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

Benjamini Hochberg multiple testing adjustment

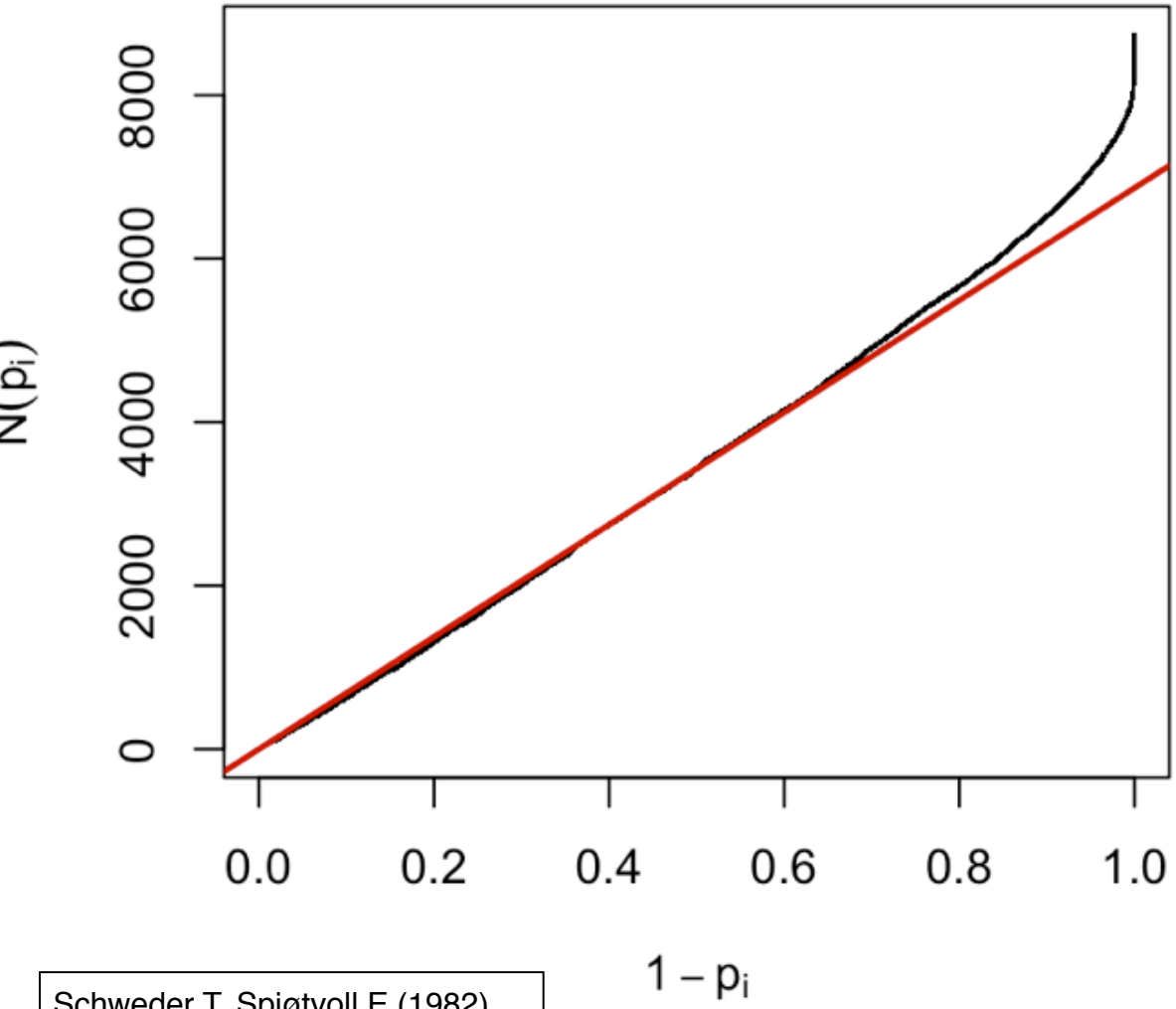


Benjamini Hochberg multiple testing adjustment




```
BH = {
        i <- length(p):1
        o <- order(p, decreasing = TRUE)
        ro <- order(o)
        pmin(1, cummin(n/i * p[o]))[ro]
     }
```

Schweder and Spjøtvoll p-value plot



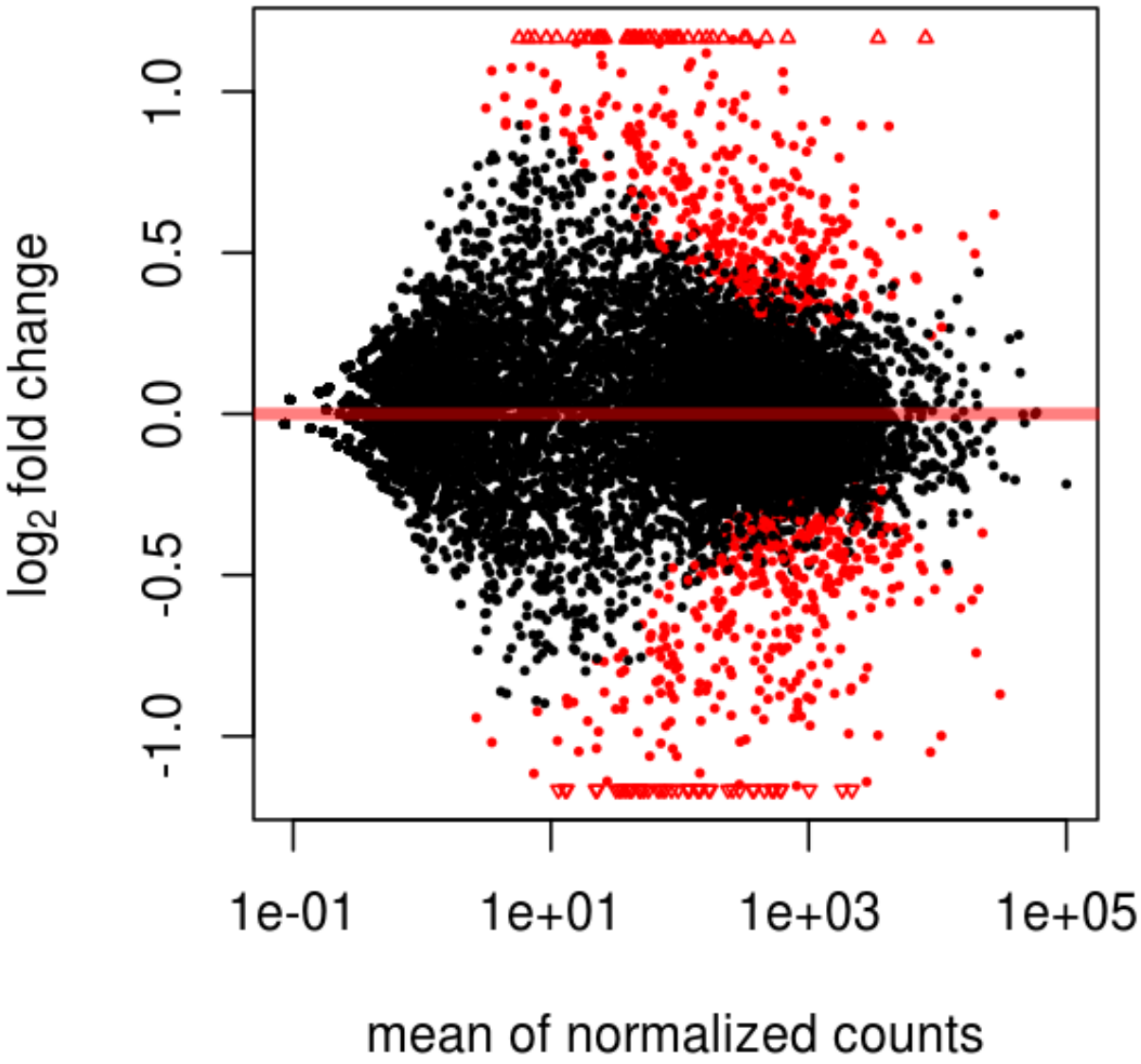
For a series of hypothesis tests $H_1 \ldots H_m$ with p-values p_i, plot

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

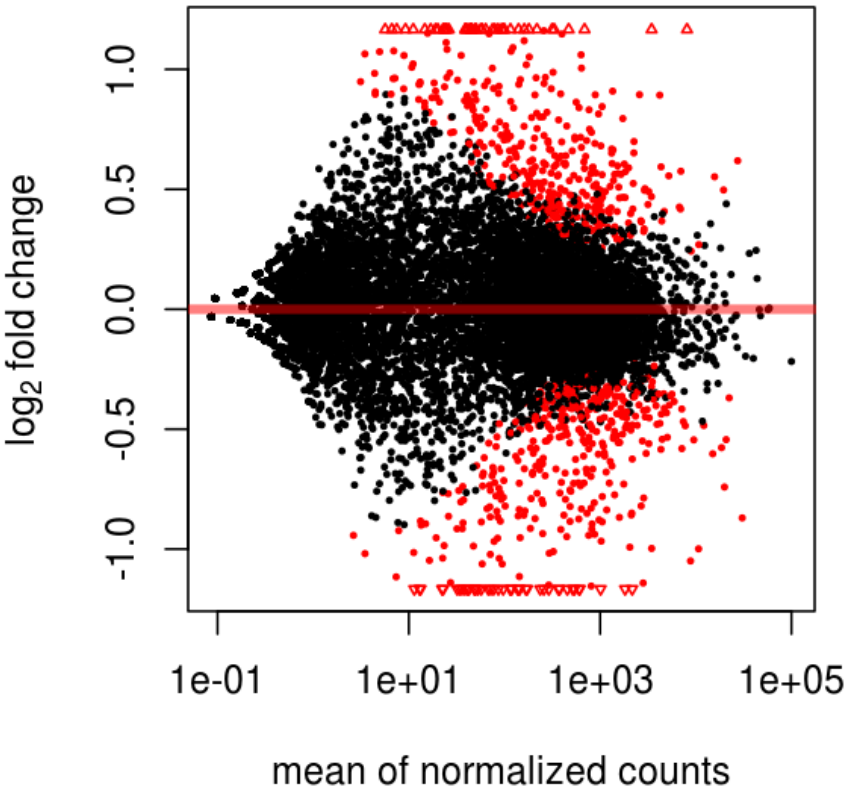
where N(p) is the number of p-values greater than p.

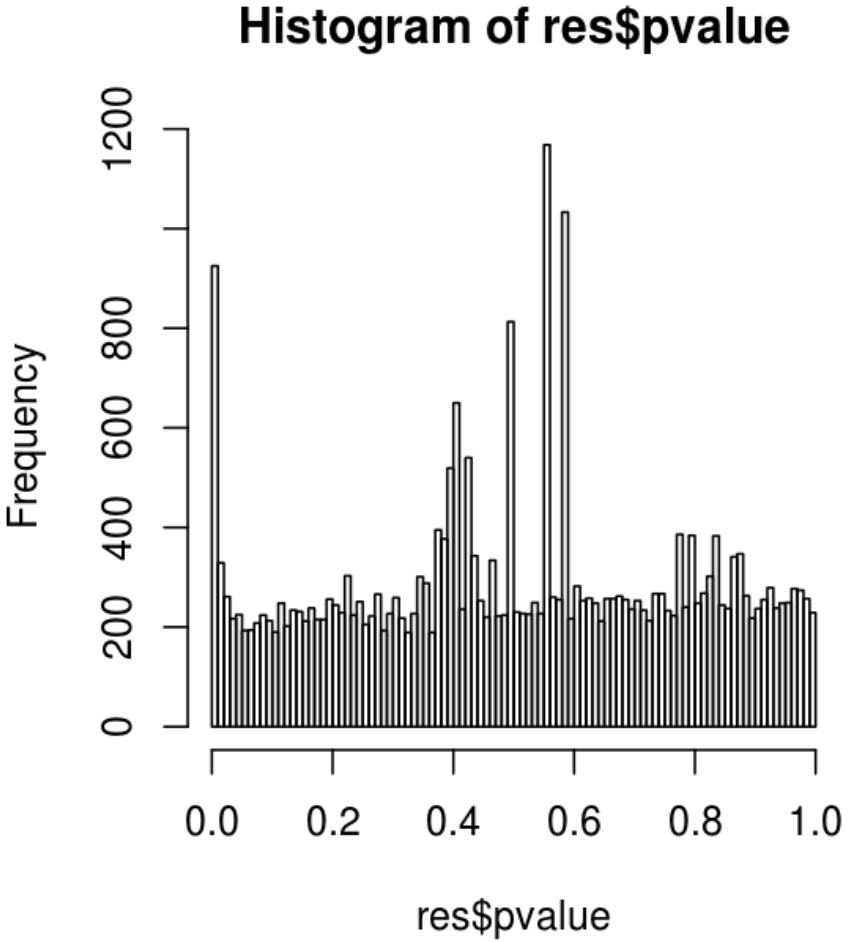
Schweder T, Spjøtvoll E (1982) Plots of P-values to evaluate many tests simultaneously. Biometrika 69:493–502.

DESeq2 lab - parathyroid dataset

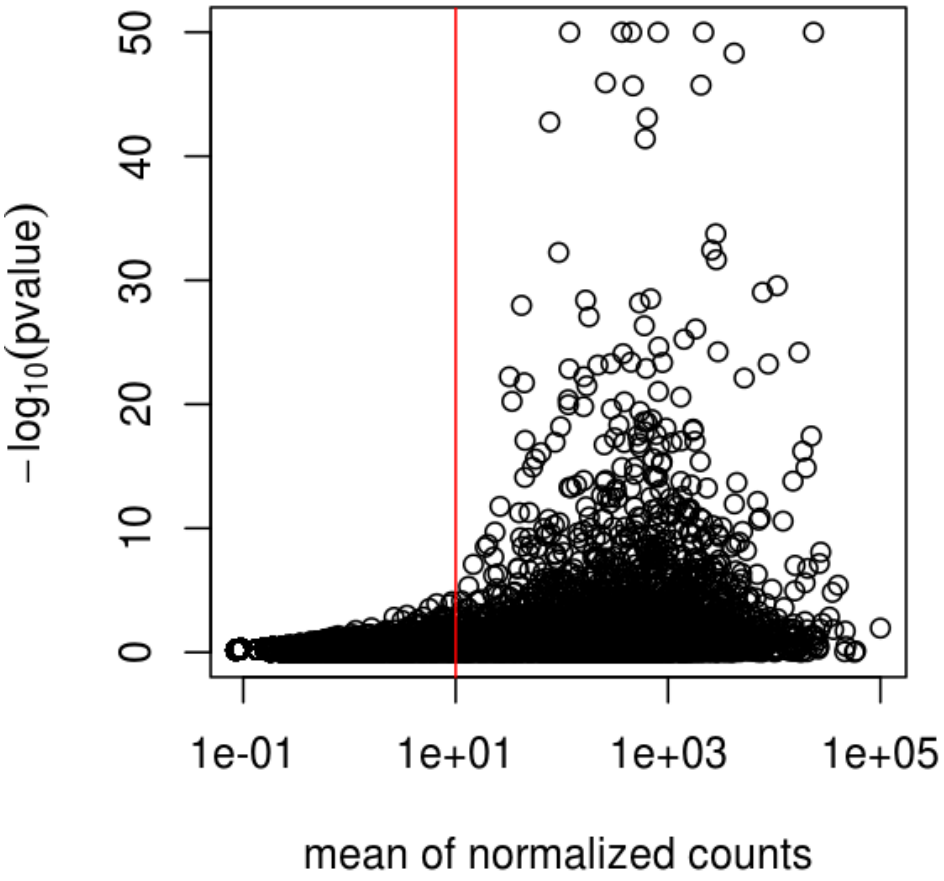


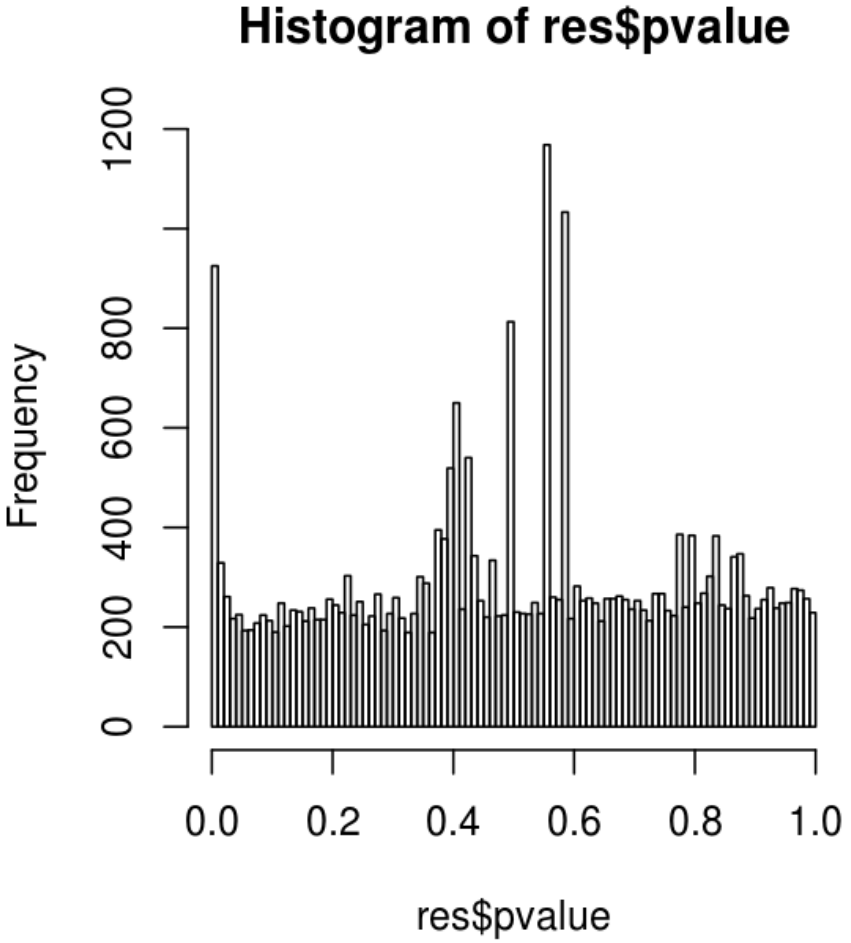
DESeq2 lab - parathyroid dataset





DESeq2 lab - parathyroid dataset





Independent filtering

From the set of all rows in the table,

first filter out those that seem to report negligible signal,

then formally test for differential expression on the rest.

Literature:

von Heydebreck, Huber, Gentleman (2004)

Chiaretti et al., Clinical Cancer Research (2005)

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

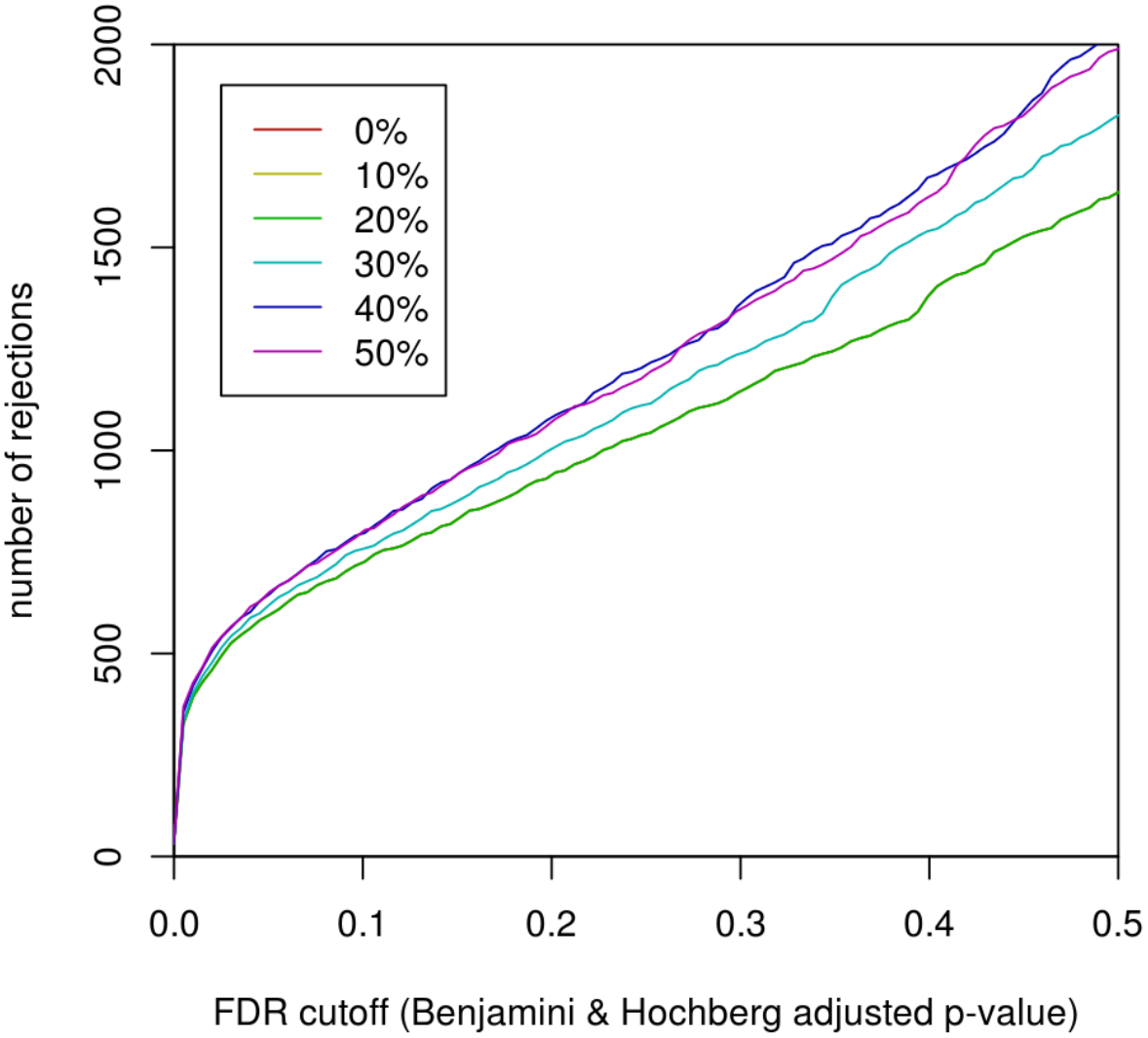
Hackstadt and Hess (BMC Bioinf. 2009)

Bourgon et al. (PNAS 2010)

Many others.

Increased detection rates

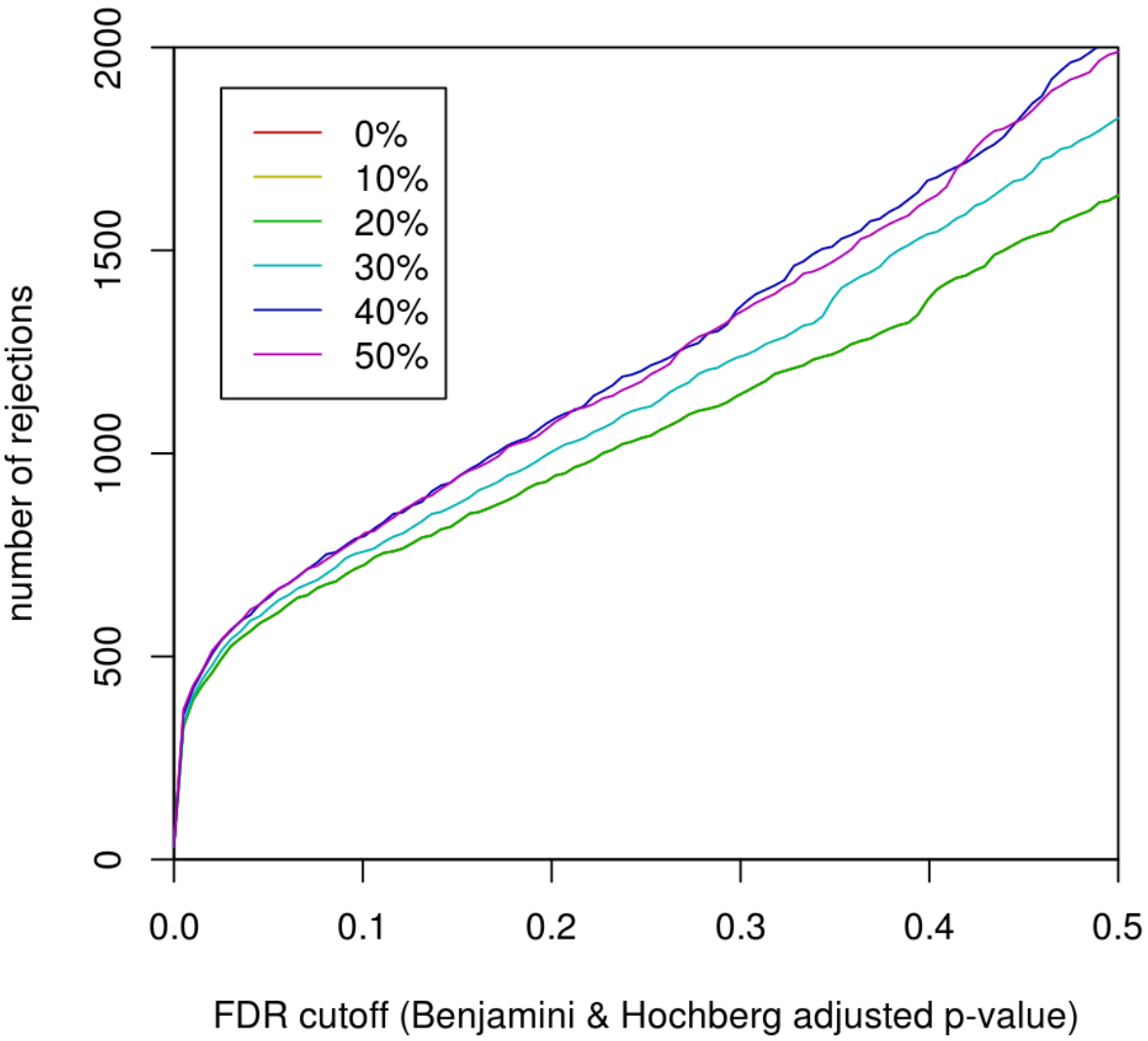
Stage 1 filter: sum of counts, across samples, for each row, and remove the fraction θ that are smallest
Stage 2: standard NB-GLM test



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?

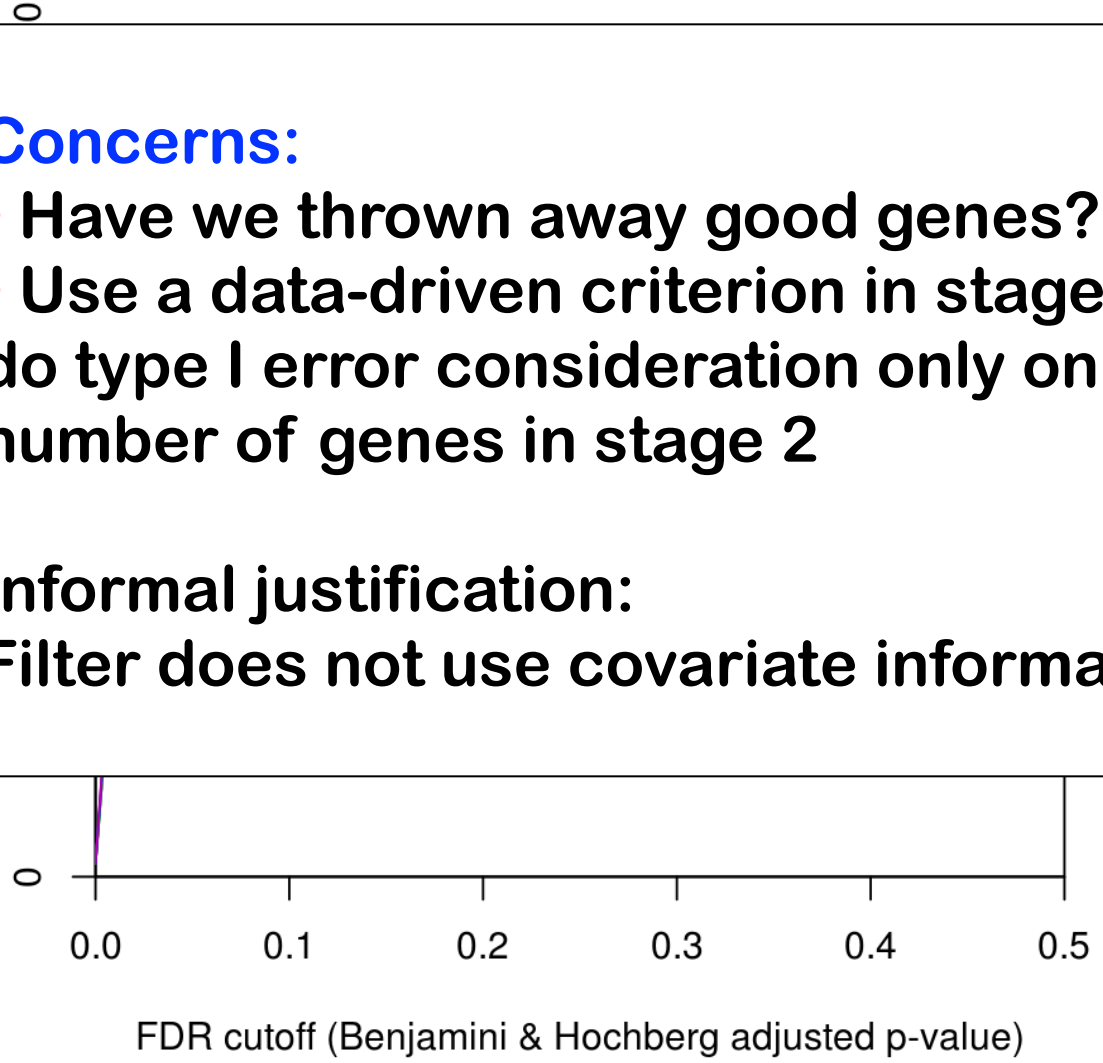
Increased detection rate implies increased power

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 filter and test statistics under the null hypothesis

For genes for which the null hypothesis is true (X_1 ,...., X_n exchangeable),

f (filter) and g (test) are statistically independent in all of the following cases:

- NB-test (DESeq(2)):

  f: overall count sum (or mean)

- Normally distributed data (e.g. microarray data after `rma` or `vsn`):

  f: overall variance, overall mean

  g: 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.

Also in the multi-class context: ANOVA, 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 multiple testing procedures relies on certain independence properties of the joint distribution

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

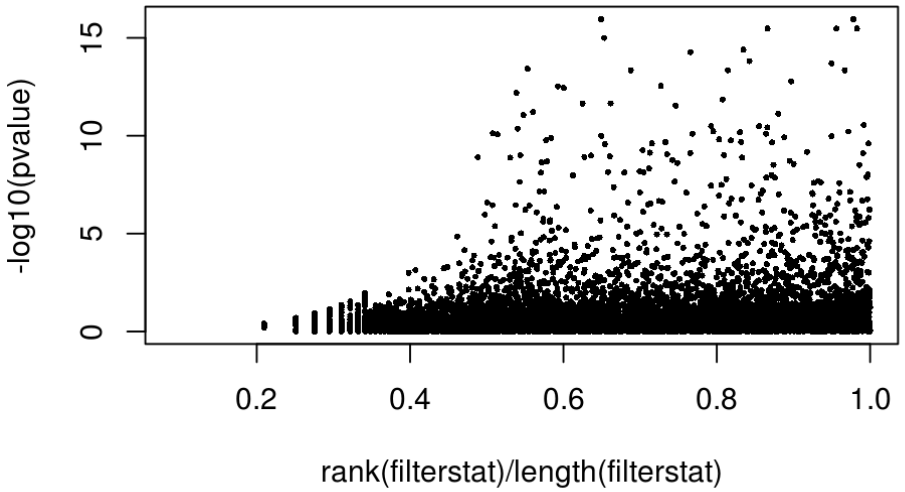
Multiple testing procedures and dependence

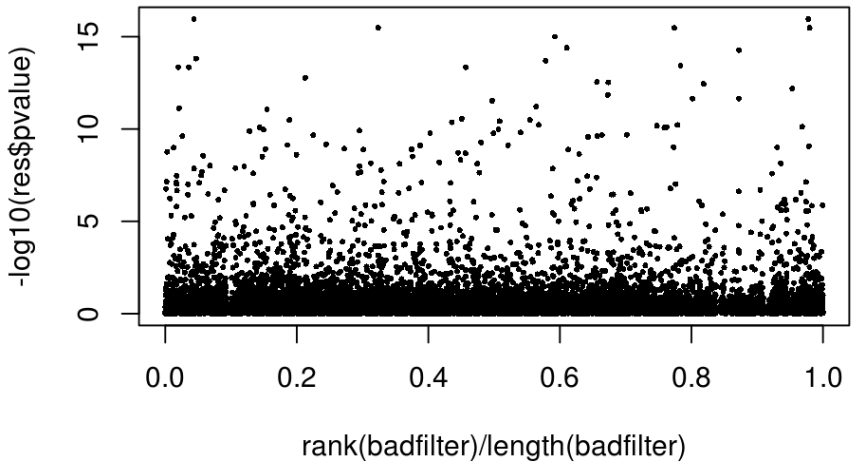
1. Methods that work on the p-values only and allow general dependence structure: Bonferroni, Bonferroni-Holm (FWER), Benjamini-Yekutieli (FDR)

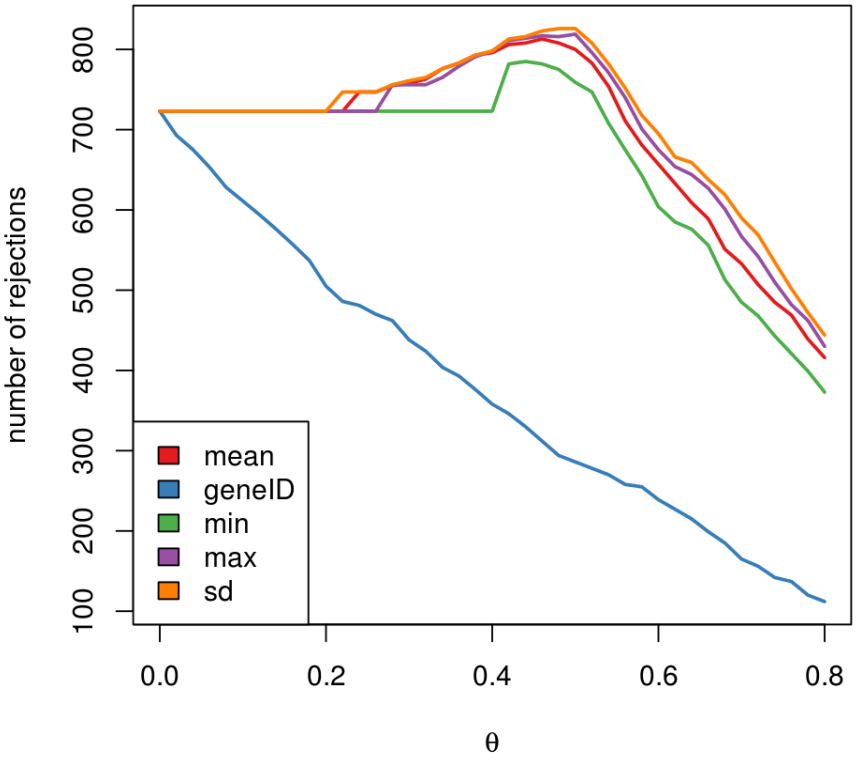
2. Those that work on the data matrix itself, and use permutations to estimate null distributions of relevant quantities (using the empirical correlation structure): Westfall-Young (FWER)

3. Those that work on the p-values only, and make dependence-related assumptions: Benjamini-Hochberg (FDR), q-value (FDR)

Diagnostics







Conclusion

Independent filtering can substantially increase your power at same type I error.

Conclusion

Independent filtering can substantially increase your power at same type I error.

References

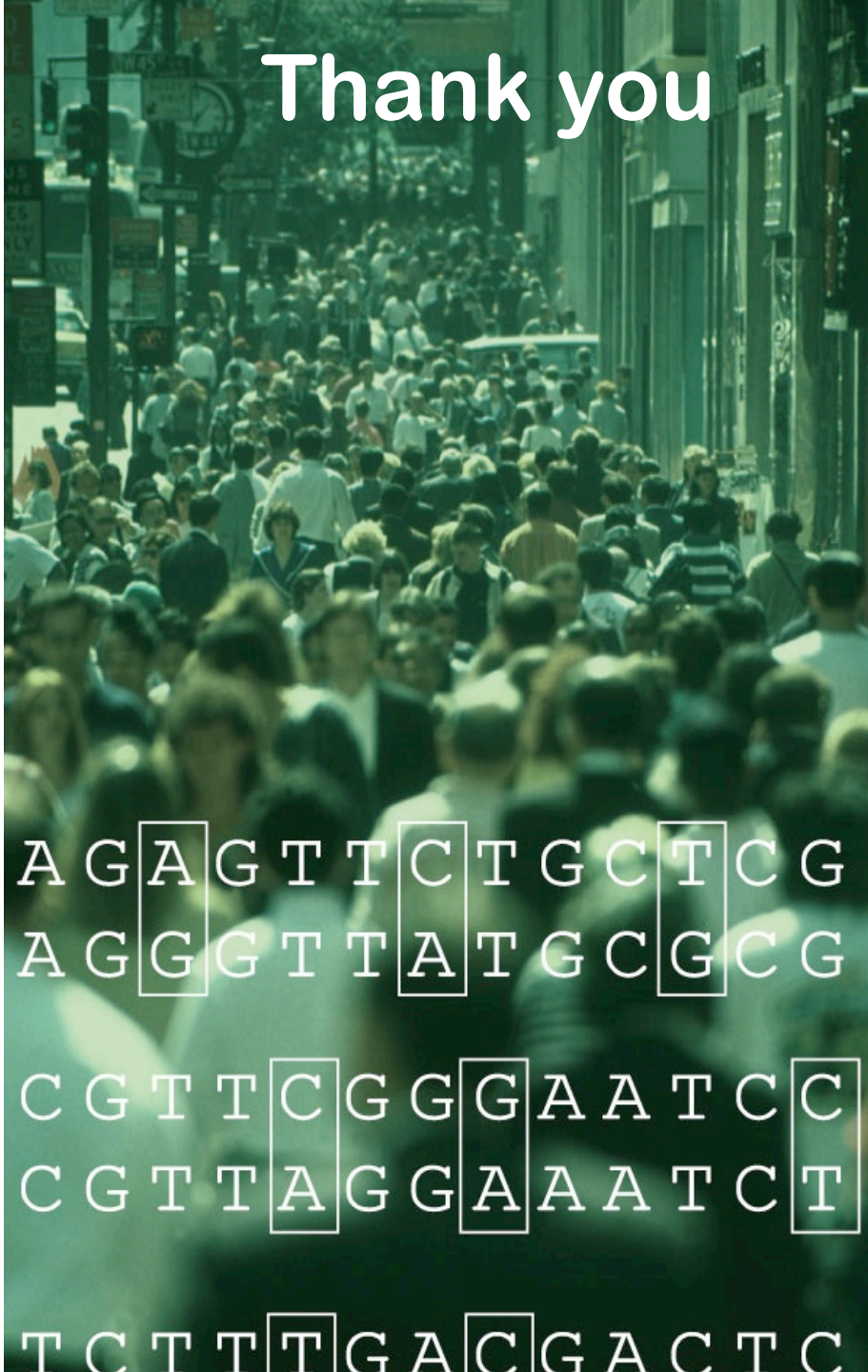
Bourgon R., Gentleman R. and Huber W. Independent filtering increases detection power for high-throughput experiments, PNAS (2010)

Bioconductor package `genefilter` vignette: Diagnostics for independent filtering

`DESeq2` vignette

Richard Bourgon

Robert Gentleman



Derivation (non-parametric case)

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

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

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

exchangeability

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

f's permutation invariance

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

distribution of g generated by permutations

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

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

Positive Regression Dependency

On the subset of true null hypotheses:

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

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

Prob(X in D | X_i = s) is increasing in s, for all i.

Important Examples

Multivariate Normal with positive correlation

Absolute Studentized independent normal