Independent filters and multiple testing

Wolfgang Huber (EMBL/EBI)
Richard Bourgon (EMBL/EBI)
Robert Gentleman (FHCRC)

Multiple testing

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

Microarray 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

Multiple testing

Classical hypothesis test:

null hypothesis H_0, alternative H_1
test statistic $X \mapsto t(X) \in \mathbb{R}$

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

$\beta = P(t(X) \notin \Gamma_{rej} \mid H_1)$ 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 α=0.05, all with H_0 true: expect 1,000 false positives

Experiment-wide type I error rates

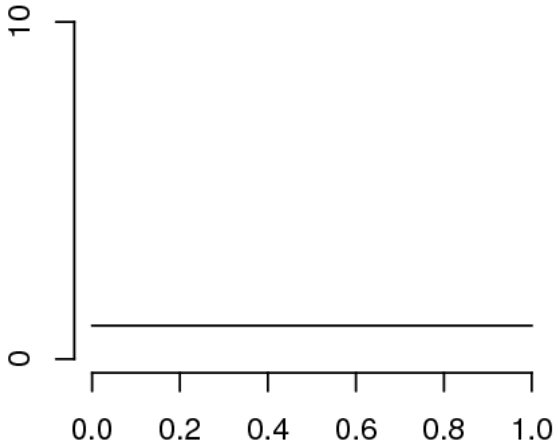
	Not rejected	Rejected	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m_1
Total	m – R	R	m

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.

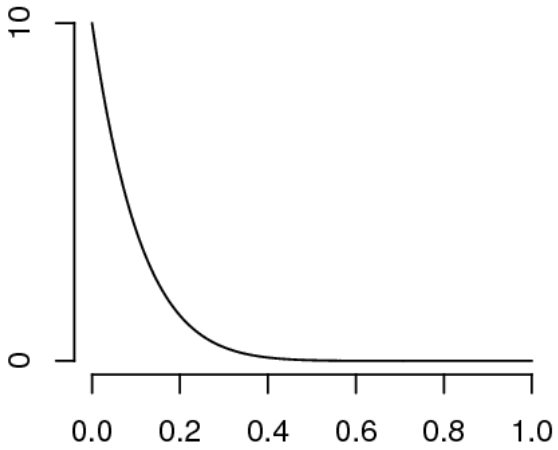
False discovery rate: E[V / max{R,1}], the expected fraction of false positives among all discoveries.

p-values: a mixture

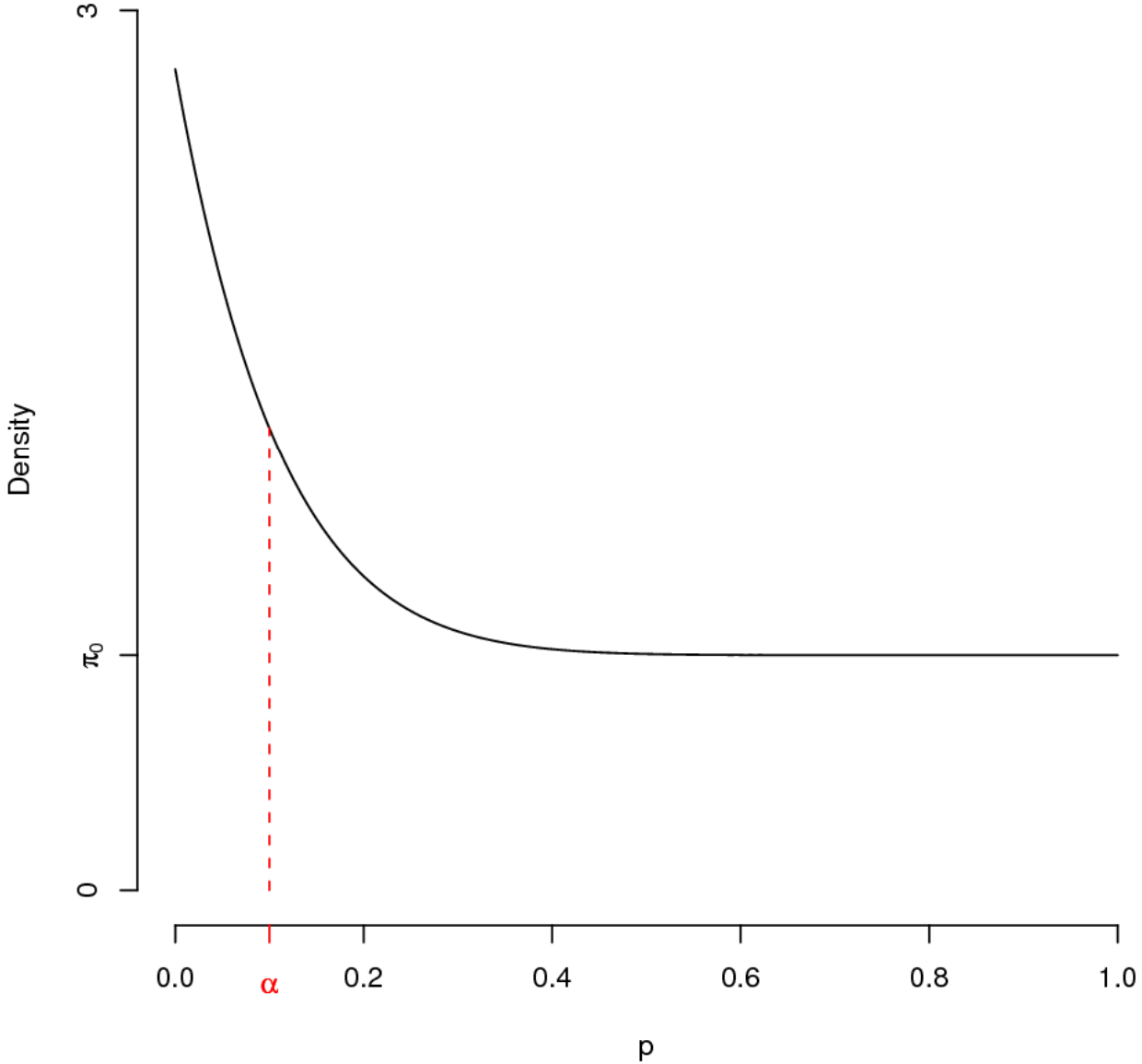
null F_0



alternative F

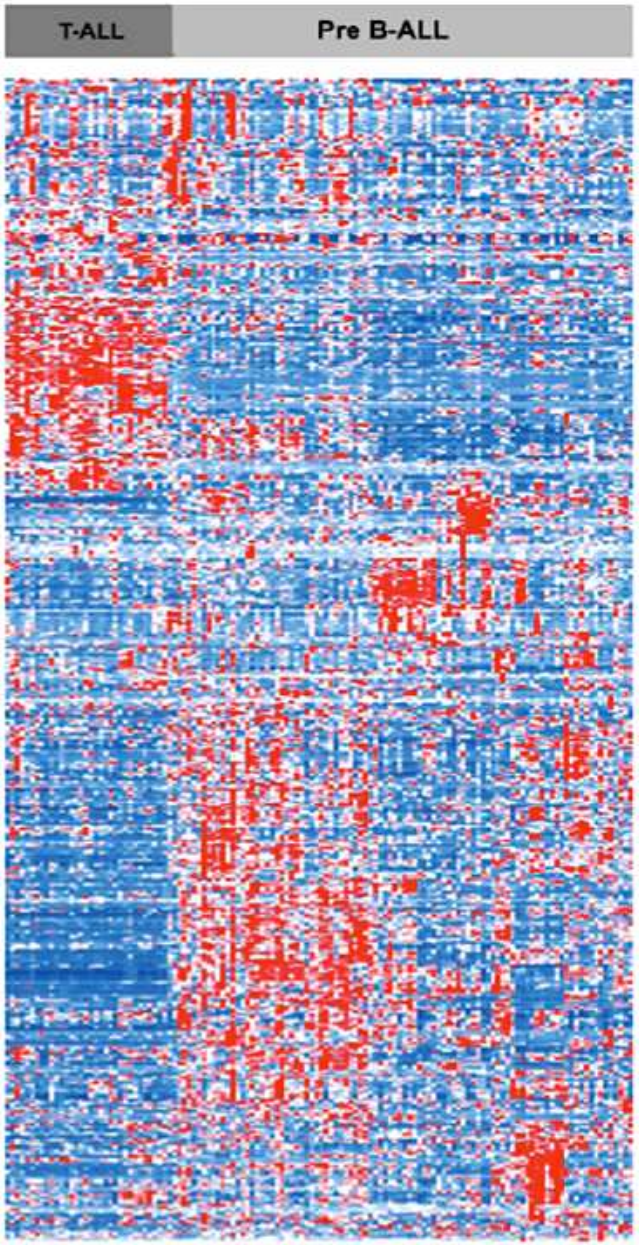


C. Mixture density $\pi_0 F_0 + \pi_1 F$ ($\pi_0 = 0.8$)



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 :

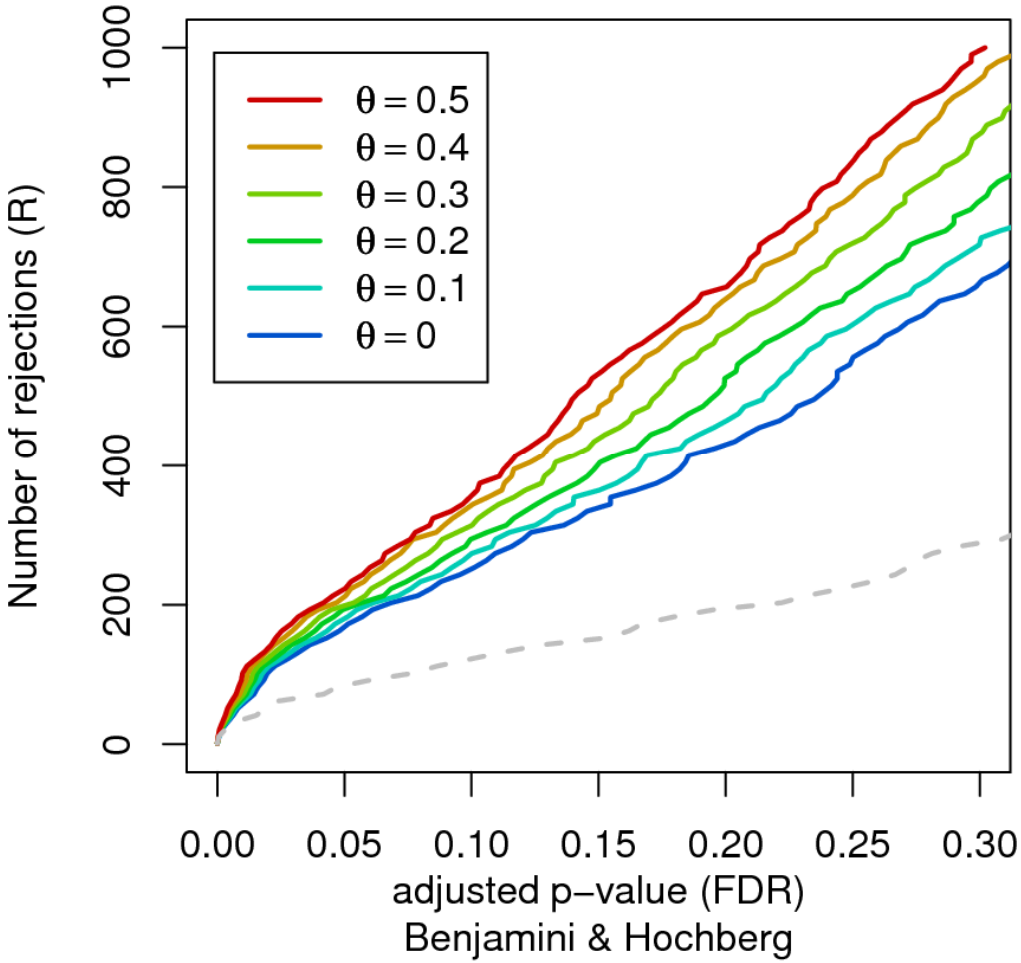
1. Target gene is absent in both samples. (Probes will still report noise and cross-hybridization.)
2. Probe set fails to detect the target.

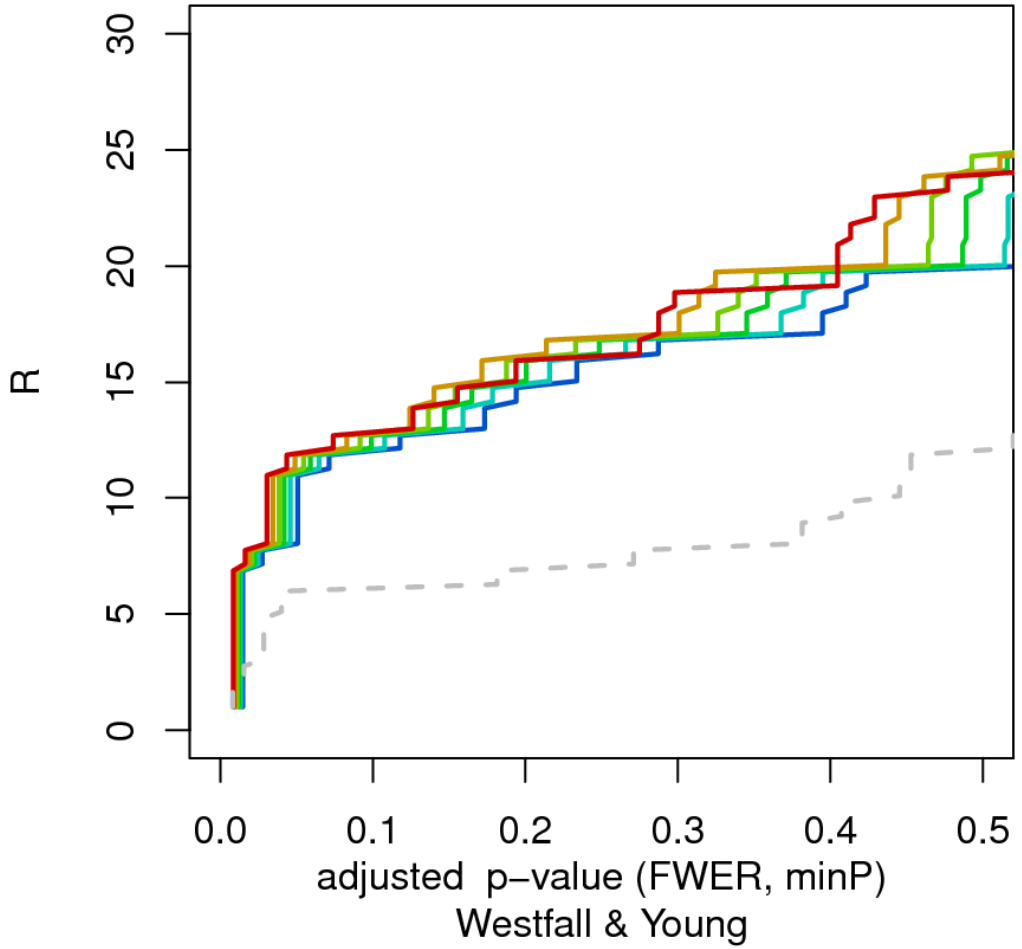
Literature: von Heydebreck et al. (2004)
McClintick and Edenberg (BMC Bioinf. 2006) and references therein
Hackstadt and Hess (BMC Bioinf. 2009)

Many others.

Increased detection rates

Stage 1 filter: compute variance, across samples, for each probeset, and remove the fraction θ that are smallest
Stage 2: standard two-sample t-test

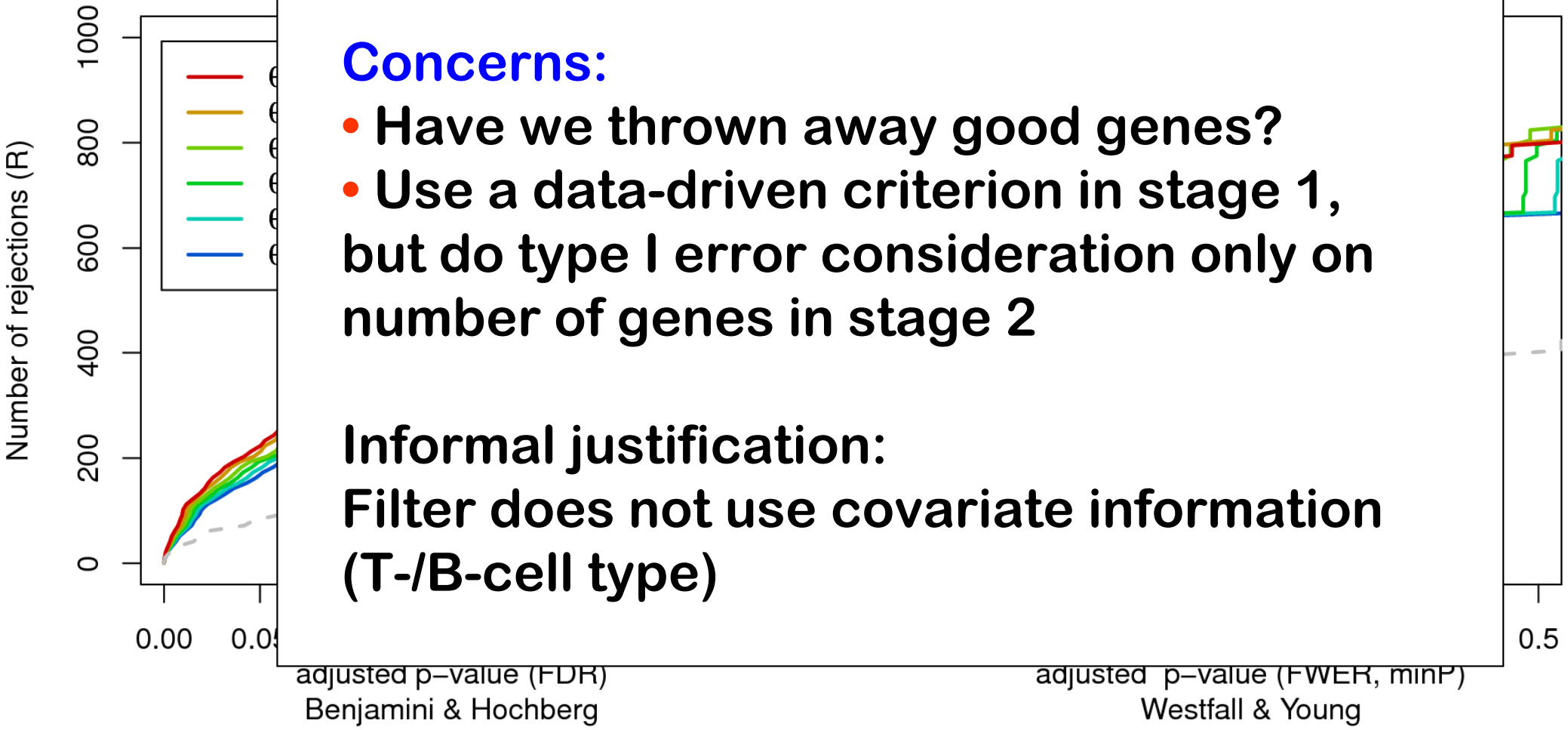




ALL data

Increased power?

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



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 (T-/B-cell type)

ALL data

Non-specific filtering?

An informal explanation has been that the filtering "does not use any information from the class labels". However, this is not enough, as these examples show:

1. Unsupervised clustering of the samples into two groups, then filter by t-statistic for these groups. Asymptotically, and for certain data,

$$\text{stage 1 statistic} \equiv \text{stage 2 statistic}$$

2. Certain null distributions of the data across samples that are not rotation symmetric (but iid).Then,

$$\mathcal{L}_t \neq \mathcal{L}_t \mid \sigma$$

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

For genes for which the null hypothesis is true (X_1 ,..., 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 $\mathcal{P}$, T ancillary $\Rightarrow$ T, V independent)

Type I error control requires

1. Correct specification of the the *marginal* distribution of the test statistic for the true nulls. ✓

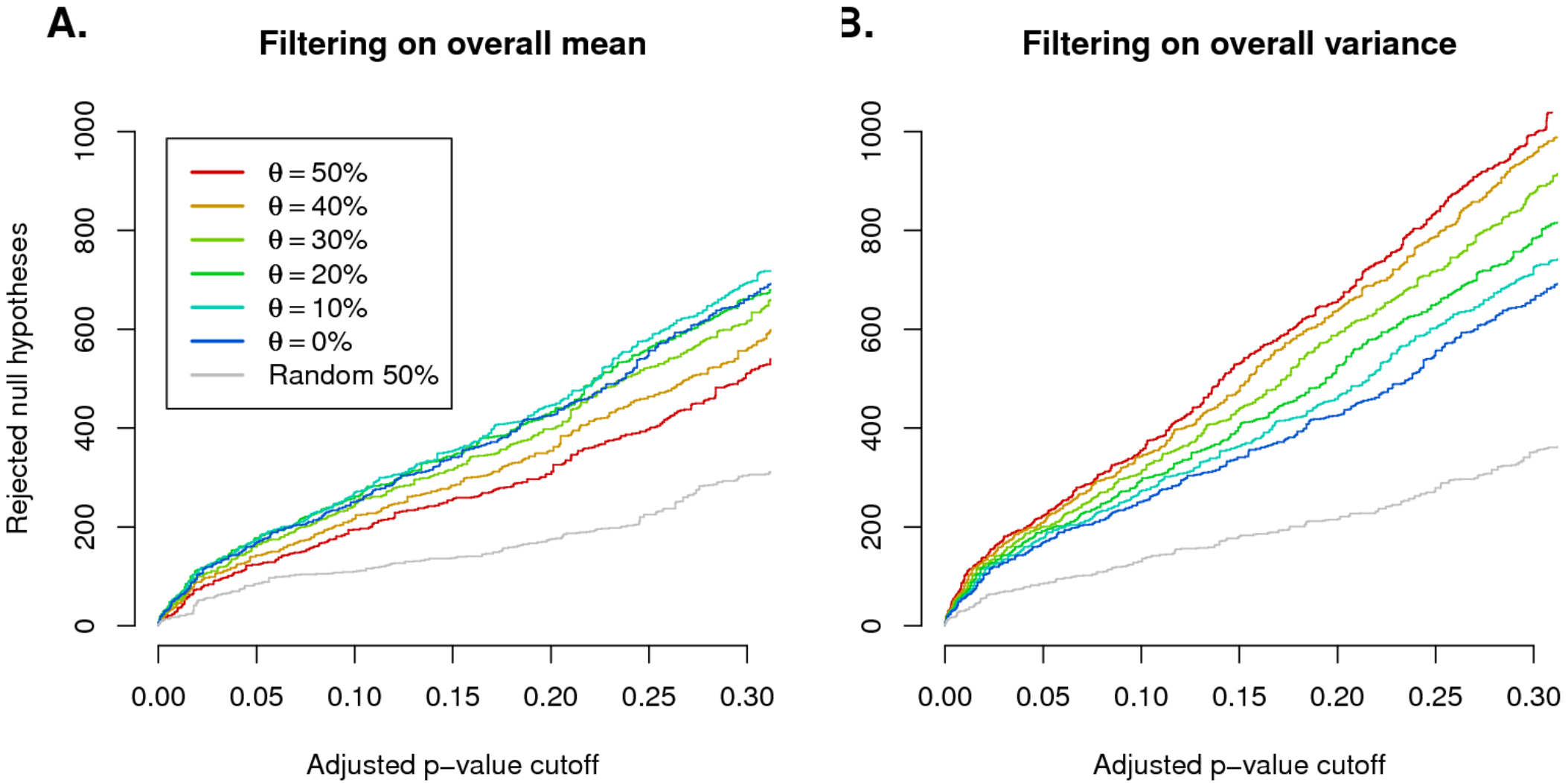
2. A dependence structure which is appropriate for the method being used.

more subtle

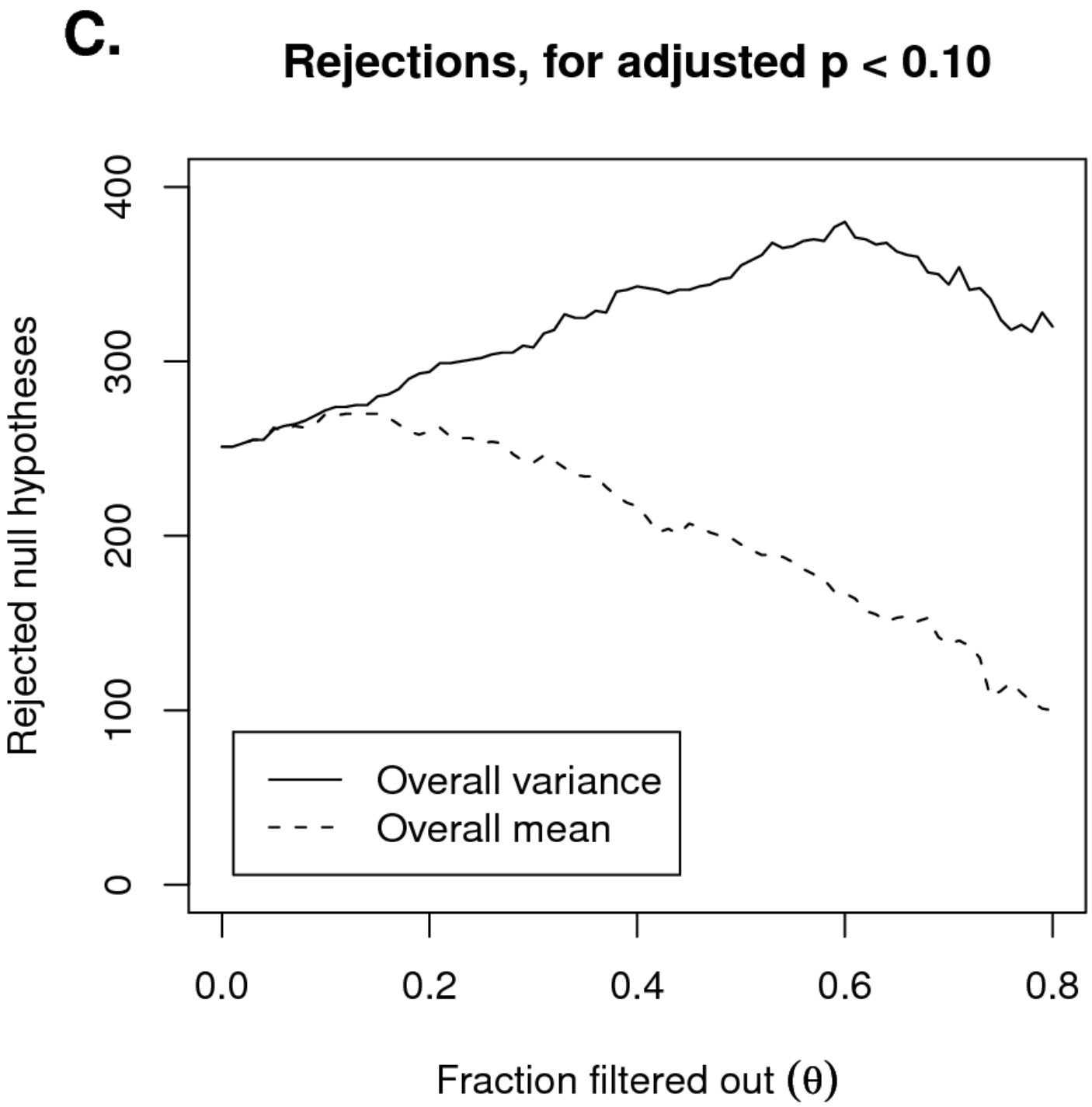
How multiple testing procedures deal with dependence

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 require dependence-related assumptions: Benjamini-Hochberg (FDR), q-value (FDR)

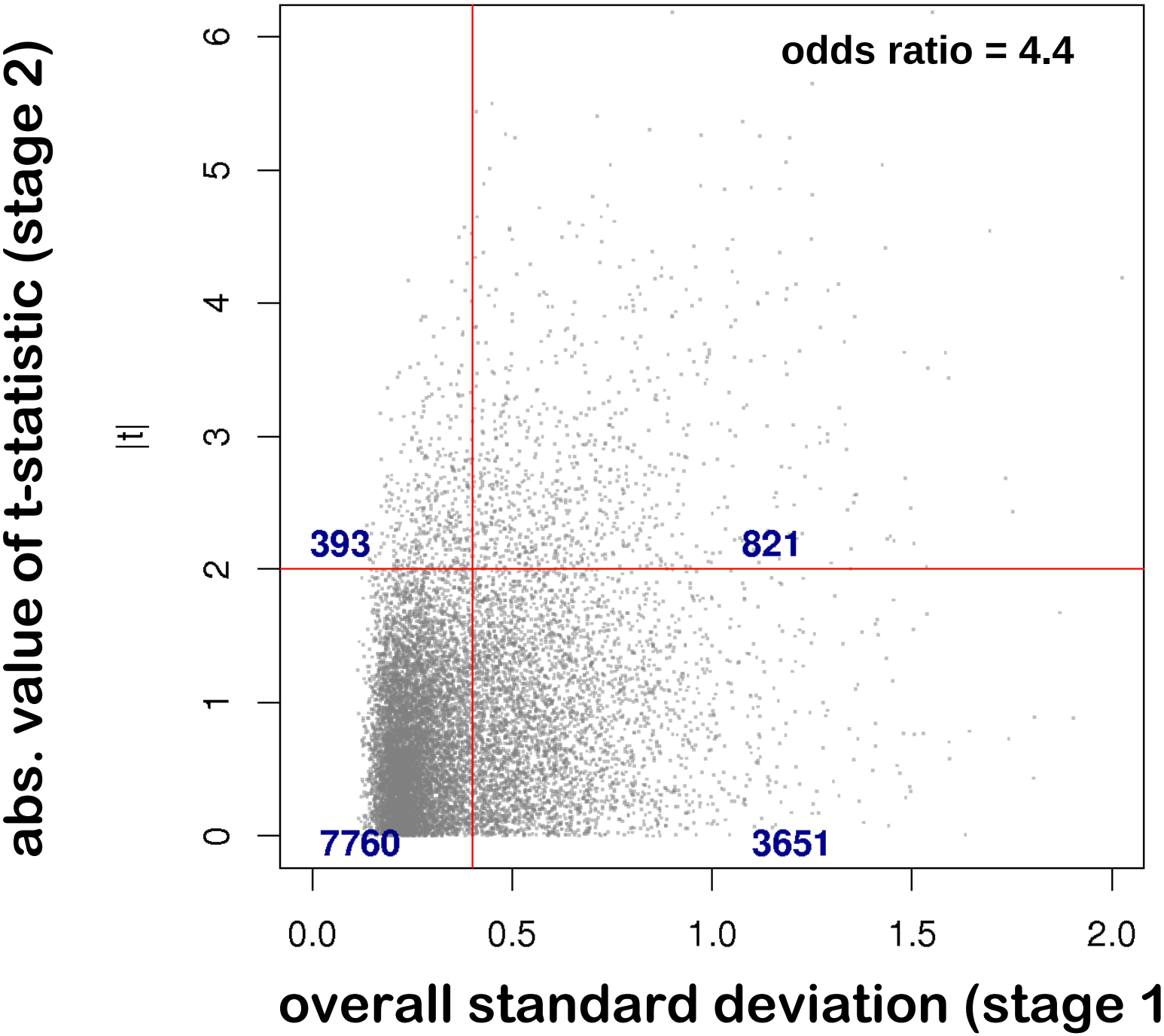
Now we are confident about type I error, but does it do any good? (power)



θ


C.
Rejections, for adjusted p < 0.10
Rejected null hypotheses
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
Fraction filtered out (θ)
Overall variance
Overall mean

Diagnostics



ALL data

Results summary

There are cases in which "filtering" leads to incorrect type-I error control.

In other cases, the 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.

Marginal independence does not preclude changes to correlation structure in filtered data: control of FDR not guaranteed; this is not likely a problem in practice.

Conclusion

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

Why does it work?

The filtering step is an (informal) way to bring in additional knowledge about the data (a „model refinement“)

EMBL

Premier lab for biological research in Europe, with five sites, in Heidelberg, Cambridge (UK), Grenoble (F), Rome and Hamburg.

Cell biology, Biophysics, Developmental Biology, Structural Biology, Genome Biology, Computational Biology

EMBL

Projects

Genetics of complex phenotypes

Understanding complex (multi-variate) genotype-phenotype relationships from high-resolution model organism (yeast, fly) data

Next Gen Sequencing (Genotype, RNA-profiling, ChIP), microscopy

Based at EMBL Heidelberg‘s Genome Biology Unit

Solexa – better base-calling and image processing

Based at EBI Cambridge in Nick Goldman‘s group

limma-t (moderated t, empirical Bayes)

- Is not independent of *variance* stage 1 filter, but simulations and theoretical arguments suggest then effect will be anti-conservative
- Is independent of *mean* stage 1 filter – but at least for Affymetrix type data, that will not do much good.

Derivation (non-parametric case)

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

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

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

Parametric case: see Richard Bourgon's poster

$= \frac{1}{n!} \sum_{\pi \in \Pi_n} \int_{\mathbb{R}^n} \delta_A(f \circ \pi(X$

f's permutation invariance

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

distribution of g generated by permutations

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

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