Multiple Hypothesis Testing in Genomics III

Simes inequality and confidence for the FDP

Jelle Goeman Aldo Solari

Radboud University Medical Center

University of Milano-Bicocca

International Society for Clinical Biostatistics
23 August 2015

Post-processing

Hypothesis test results are often intermediary

P-values calculated and corrected at the probe level

Later

- Aggregated to gene or pathway level
- Selected from
- Turned into networks
- Used in integrated analysis

Question

Conclusions still OK?

Example 1: selection

Top diff. expression

Gene	p-value
XDH	5.5e-10
NEK3	6.7e-7
TAF5	7.1e-7
CYP2A7	1.6e-6
NAT2	1.8e-6
ZNF19	2.6e-6
SKP1	2.7e-6
NAT1	3.1e-6
GDF3	2.0e-5
CCDC25	2.1e-5
⋮	⋮

False discovery rate control

Expected: $< 5\%$ false positives

Practice

Genes chosen for further analysis

Question

How many false positives to expect?

Example 2: aggregation

Relationship methylation $\longleftrightarrow$ gene expression

Tested: 4,734,505 CpG—gene combinations

Found

12,159 combinations at FDR $< 5\%$.

Belonging to

6,540 CpG's and 3,521 genes

Claim: 3,521 genes have expression influenced by a CpG

How many false positives to expect?

Flexibility

Hypothesis test in bioinformatics

Often an intermediary result, not an end result

Needed: flexibility

Valid assessment of error rates after

- aggregation
- selection

Needed: allow post hoc reasoning

Decisions on aggregation/selection taken after seeing the data

Example: FDR and aggregation

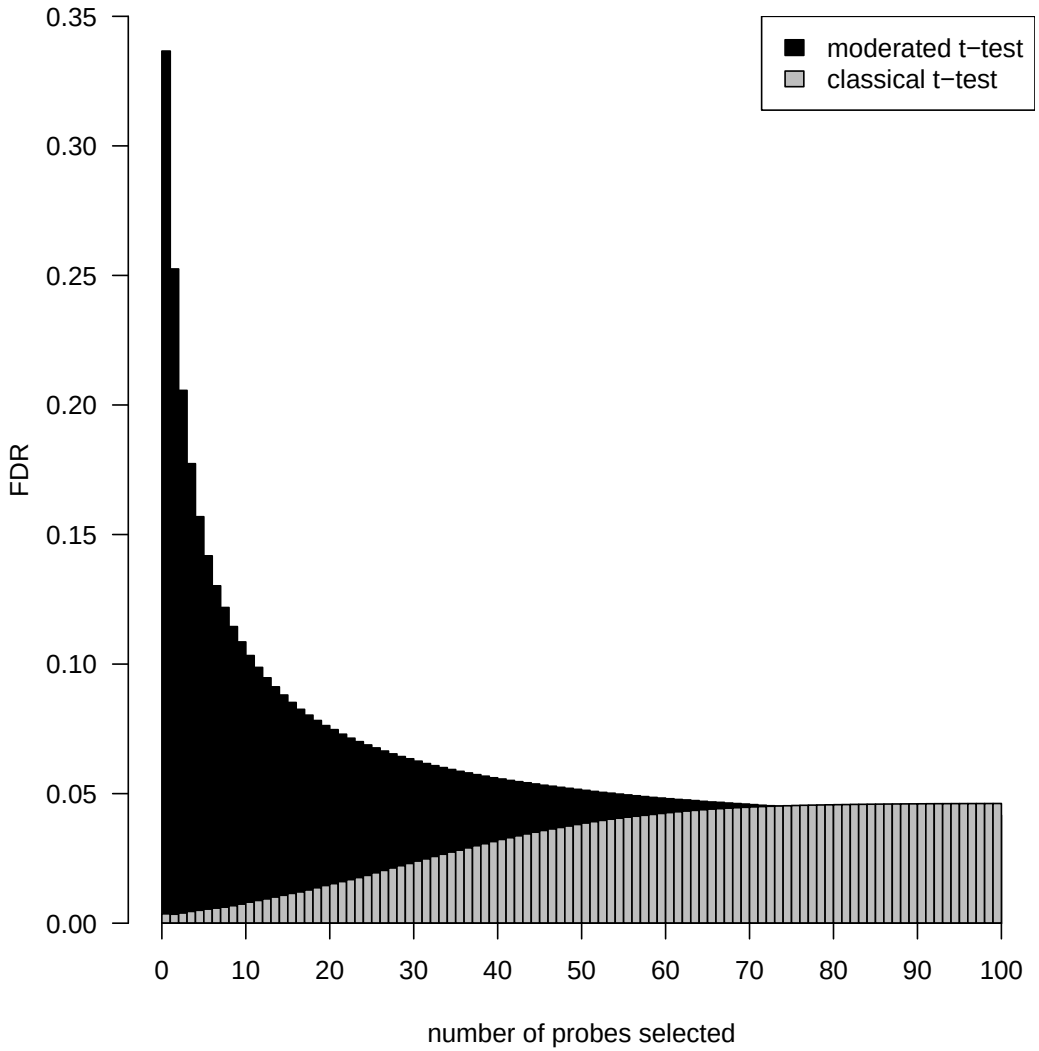
Simple simulation

- $N = 100$ genes of 10 probes each
- Two genes with strong effect in all probes
- Other 98 genes no effect
- FDR controlled at the probe level at 5%

Aggregated result

Realized FDR at the gene level 29%.

Example: FDR, limma and fold change selection



Subsetting property

Subsetting property

If an error rate is controlled on $\mathcal{R}$, also on $\mathcal{S} \subset \mathcal{R}$

- Holds for FWER
- Does not hold for FDR

Random $\mathcal{R}$

FWER and FDR select $\mathcal{R}$ for the user

Post hoc chosen $\mathcal{R}$

User is free to choose $\mathcal{R}$ after seeing the data

Goals

Summary

- FWER: flexibility; low power
- FDR: good power; no flexibility

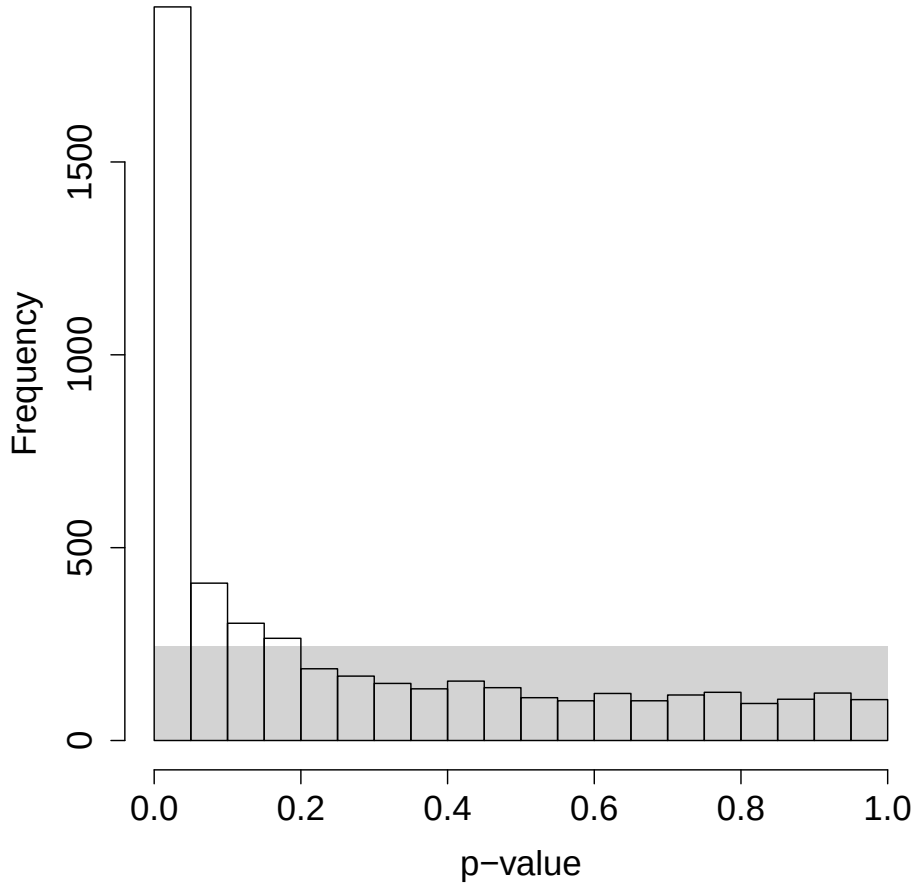
Can we have our cake and eat it too?

Flexibility and power...

Two ingredients

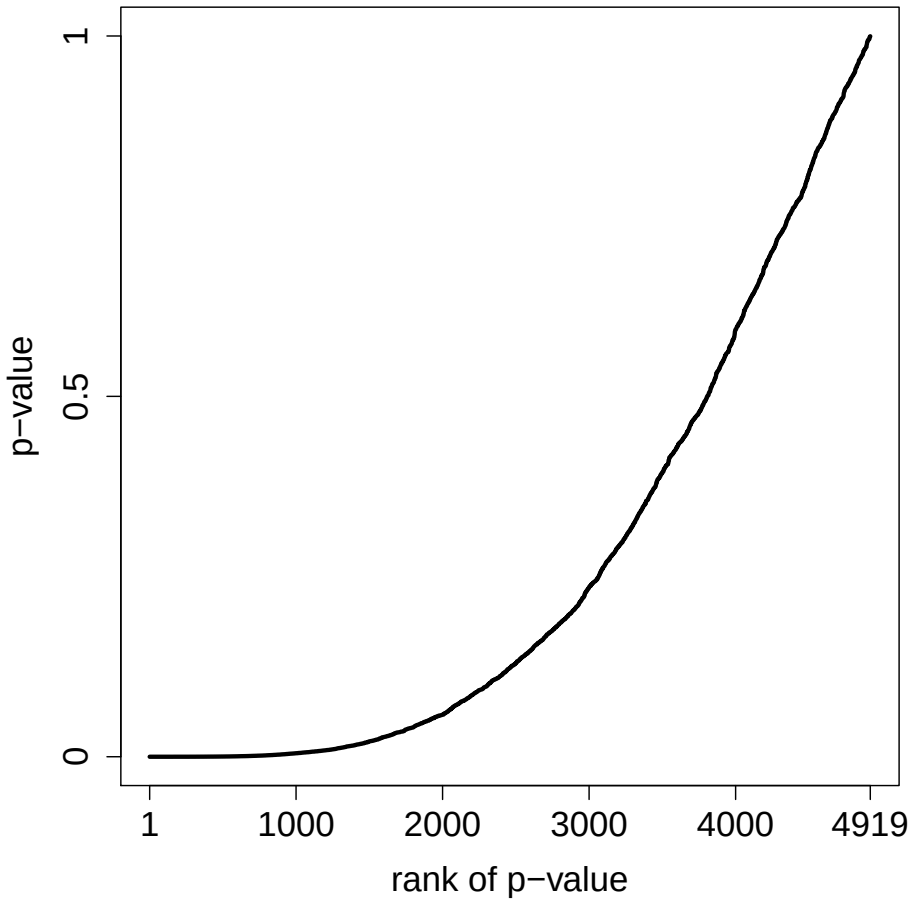
1. Simes' inequality
2. Closed testing

Histogram of p-values



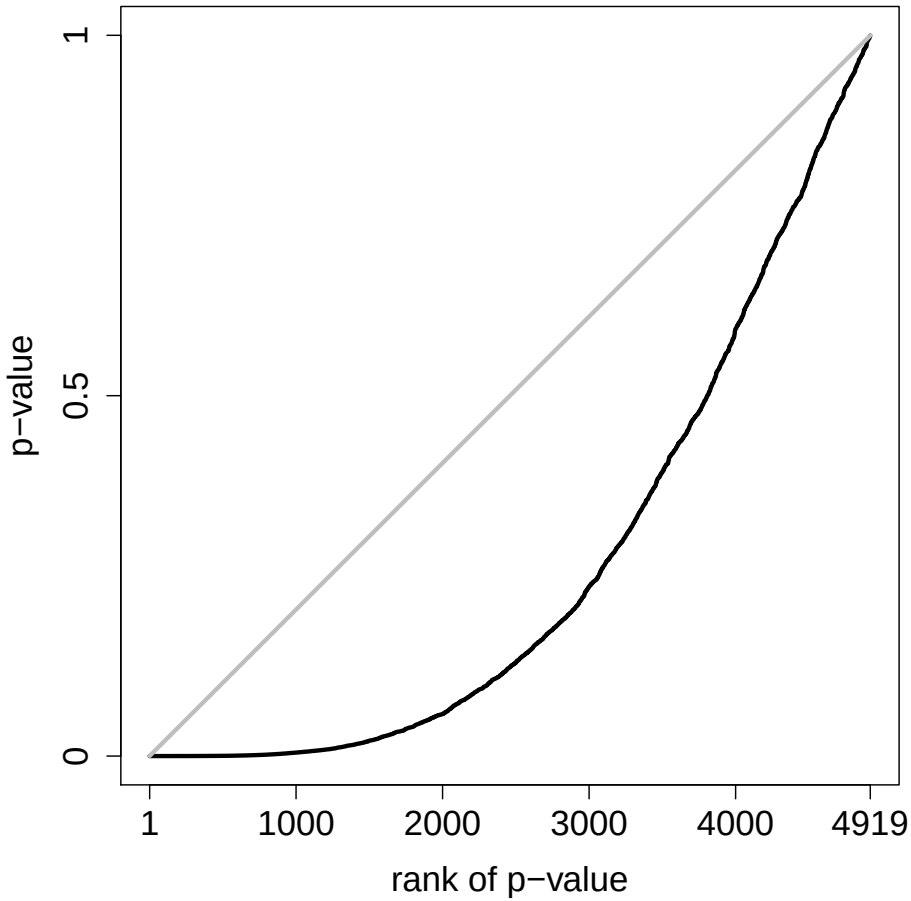
and expectation if all the hypotheses were true

Sorted p-value curve



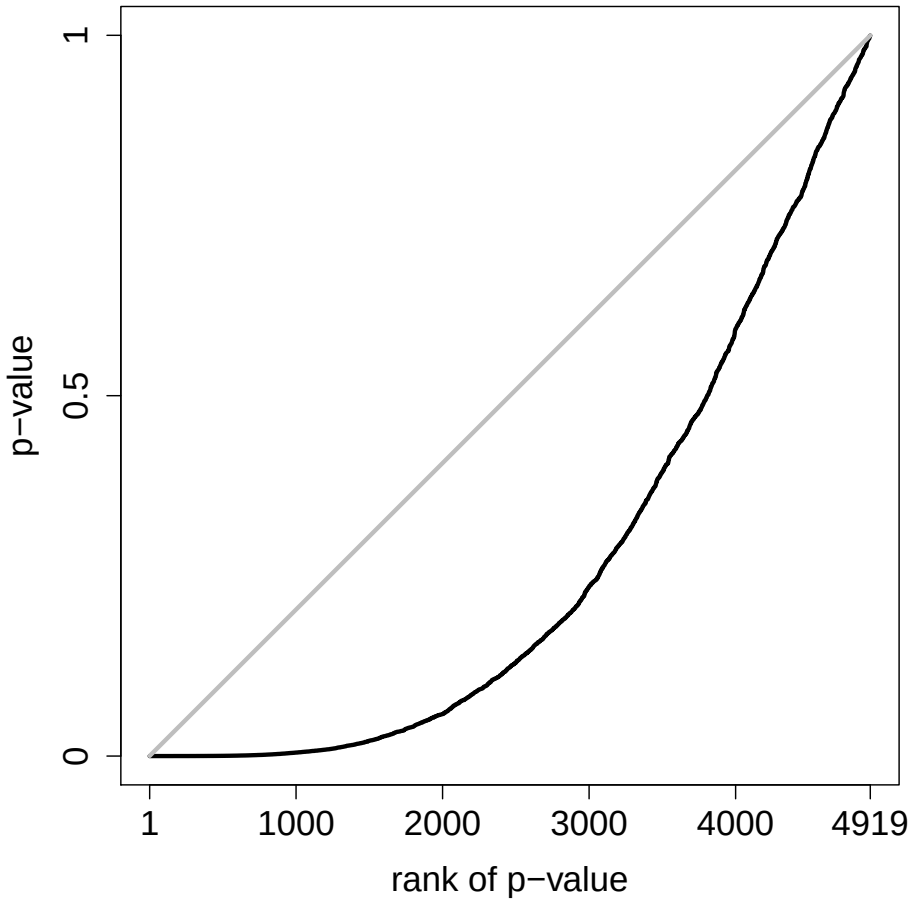
$(i, p_{(i)})$ with $p_{(i)} = i$th ordered p-value

Expected curve



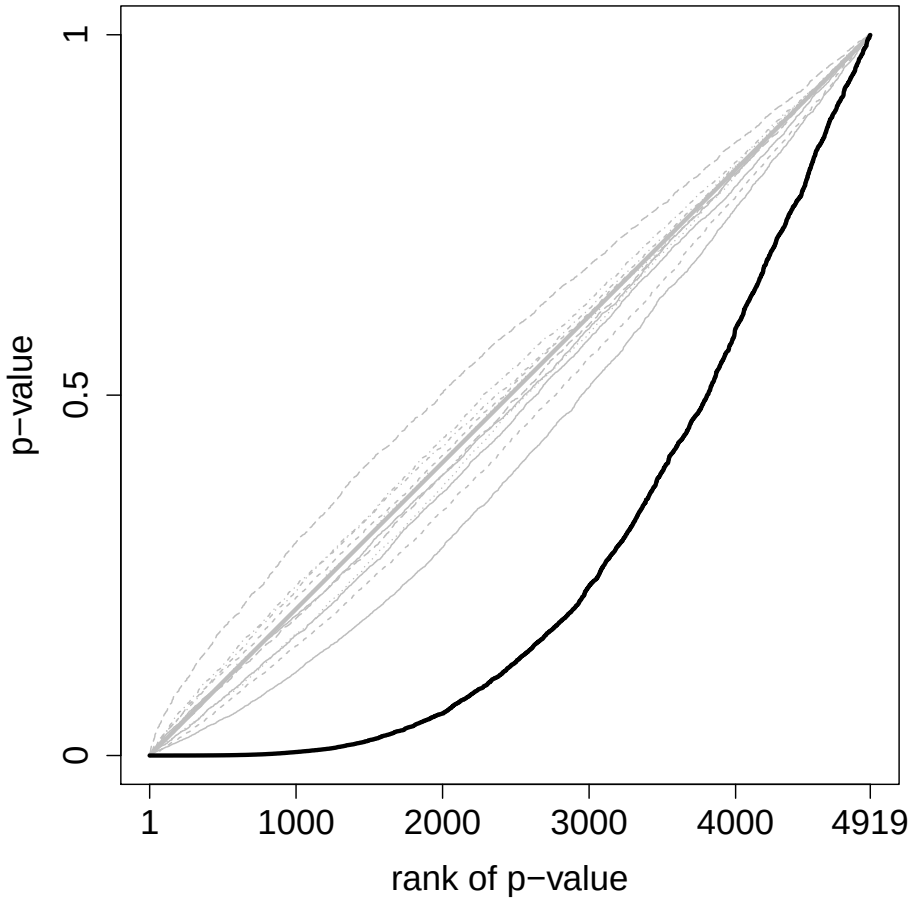
if all the hypotheses were true

Expected curve



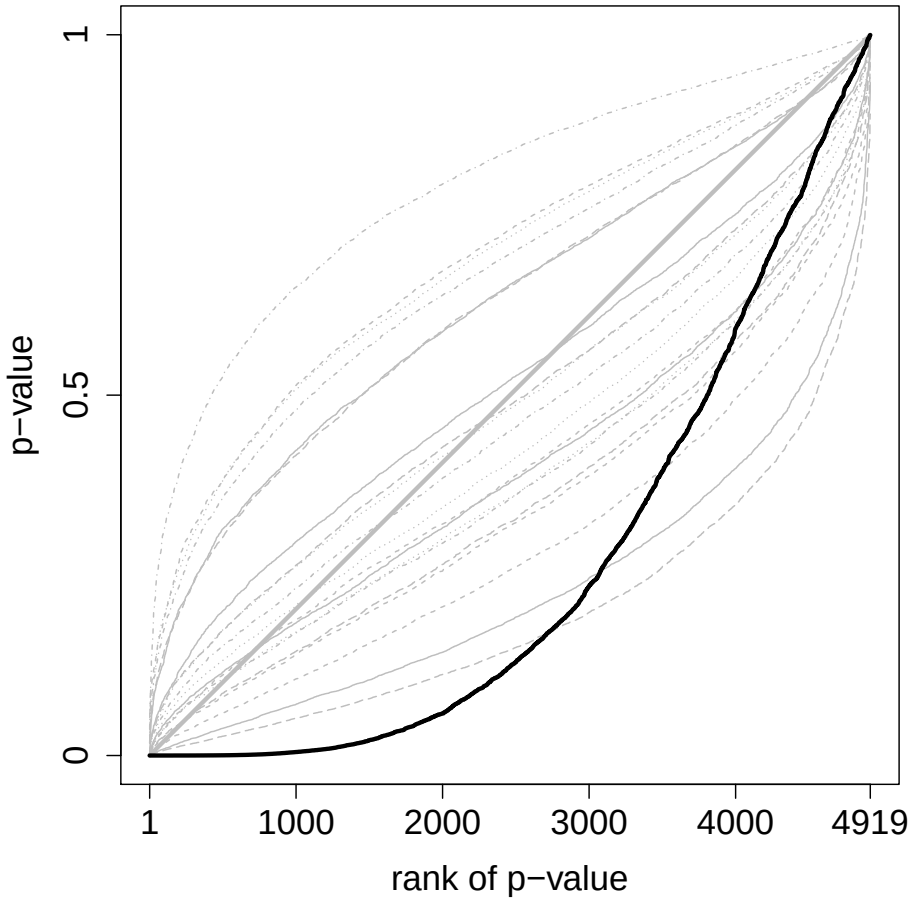
is the observed significantly smaller than expected?

Null distribution



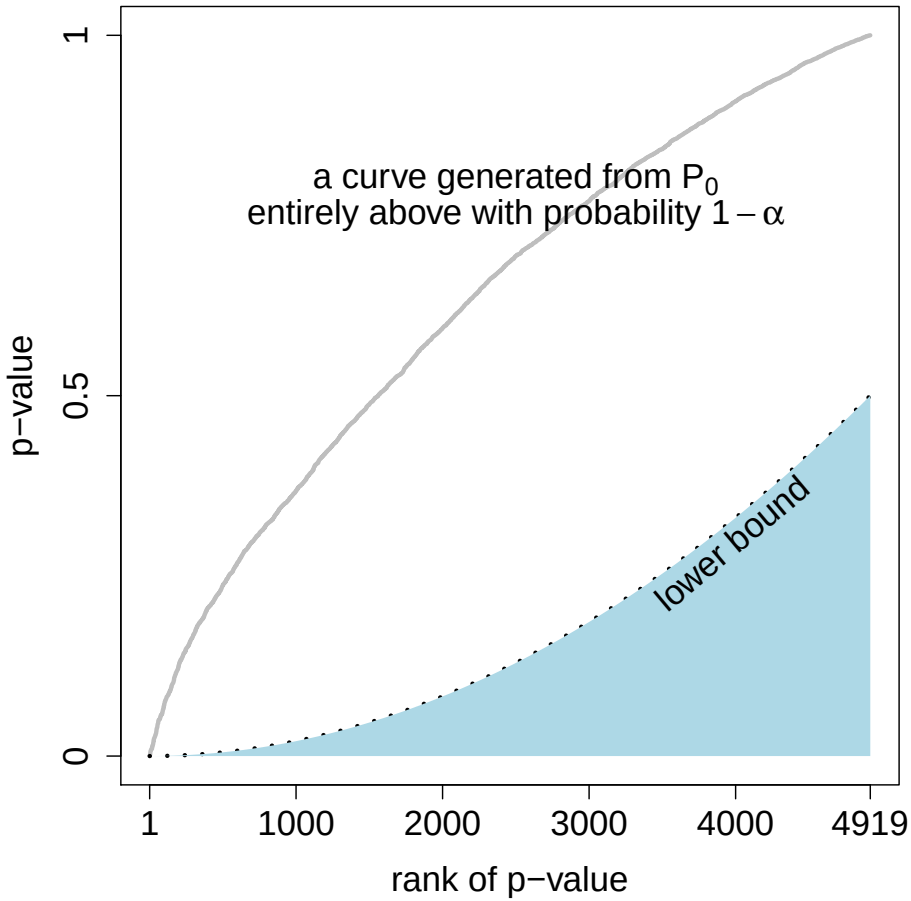
it depends on the (joint) null distribution P_0 of p-values

Null distribution



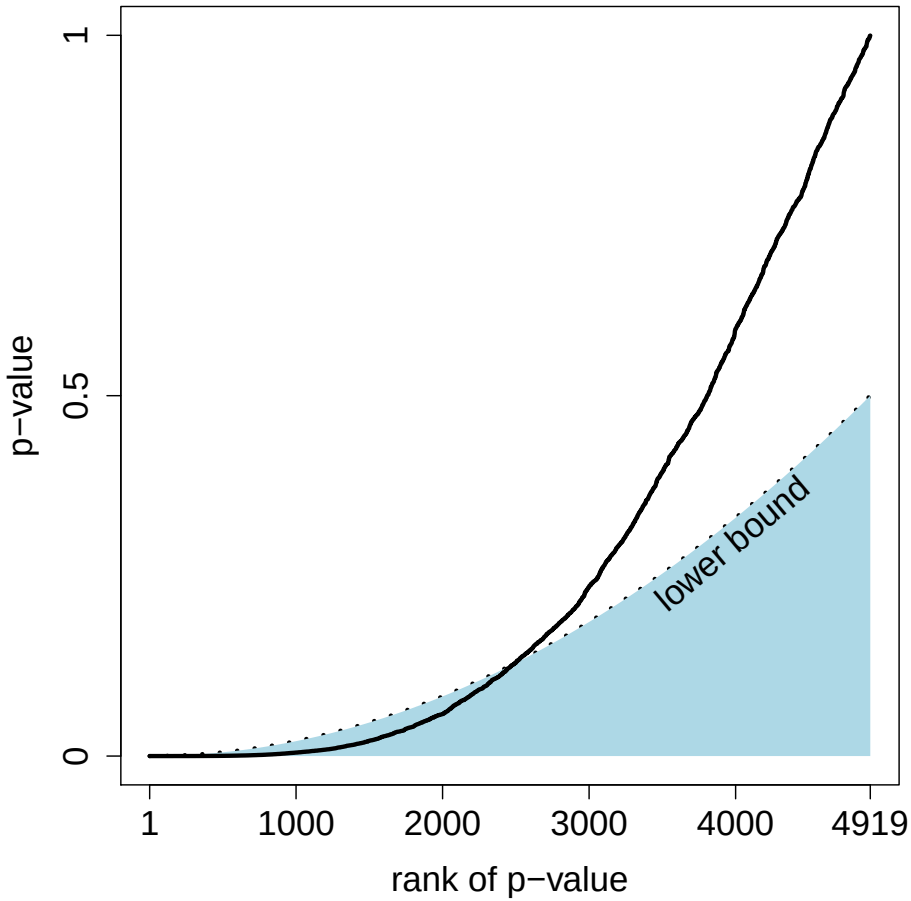
it depends on the (joint) null distribution P_0 of p-values

Simultaneous lower bound



if joint distribution P_0 known

Simultaneous lower bound



observed curve crosses: evidence of at least one false hypothesis

Simes inequality

Simes: general lower bound

If all hypotheses true, with probability at least $1 - \alpha$,

$$p_{(i)} > \frac{i\alpha}{n} \quad \text{for all } i = 1, \ldots, n$$

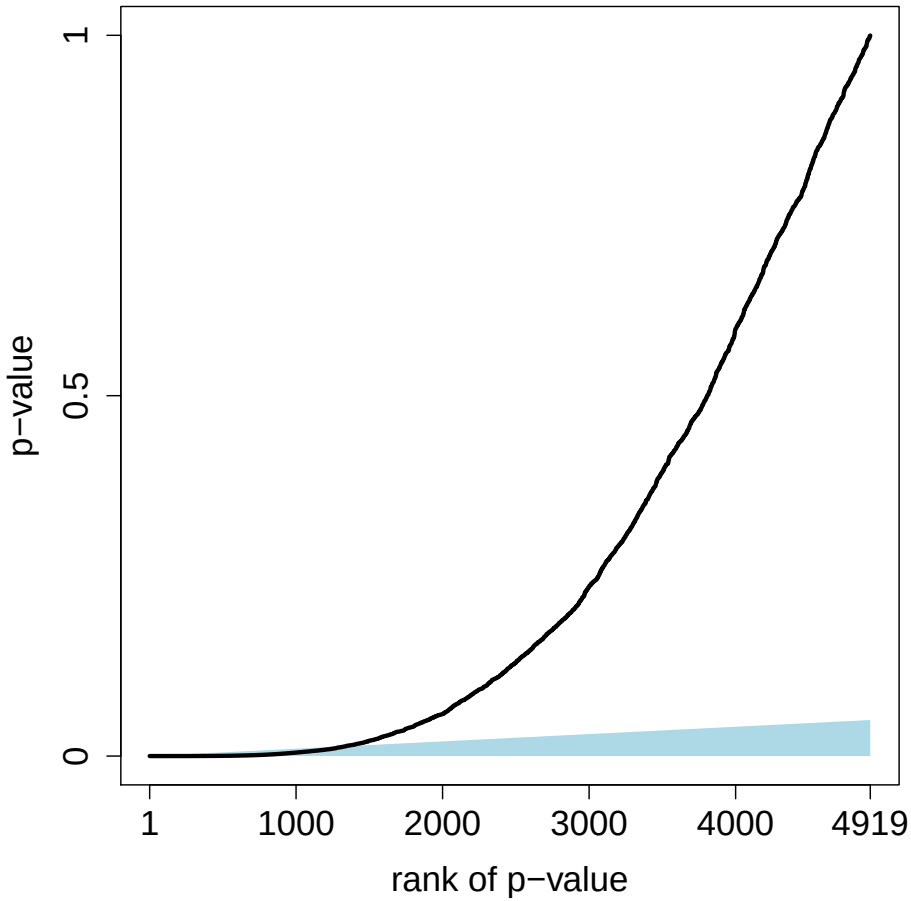
Assumptions

- Research ongoing (Sarkar and others)
- Sufficient: two-sided joint normal test statistics
- Same assumptions as Benjamini & Hochberg FDR

Use: global test

If Simes violated: at least one null hypothesis false

Simes' test



Simes test: reject if observed curve crosses Simes curve

Validity of Simes

Sufficient: PDS condition

For the $p$values $q_1, \ldots, q_{m_0}$ of true hypotheses

$$\mathrm{E}[f(q_1, \ldots, q_{m_0}) \mid q_i = u]$$

is non-decreasing in u for every i and for every coordinate-wise non-decreasing function f.

Rodland

Distributions violating Simes' inequality are quite exotic

Laüter

Simes holds for two-sided p-values from normally distributed statistics

Hommel's inequality

Version of Simes that is always valid

$$q_{(i)} > \frac{i\alpha}{m_0 \sum_{j=1}^{m_0} 1/j} \qquad \text{for all } i = 1, \ldots, m_0,$$

Advantages

No worries on PDS conditions

Can always be substituted for Simes in methods

Drawbacks

Caters for a very exotic worst case distribution

Quite conservative for most other distributions

Intersection hypothesis

Intersection hypothesis

$H = A \cap B \cap C$ is true if and only if A, B and C are all true

Example

A: gene A is not differentially expressed

B: gene B is not differentially expressed

C: gene C is not differentially expressed

Then H: A, B and C are not differentially expressed

Compare

Gene set tests for pathway testing

Set-up

Hypotheses

$H_1, \ldots, H_m$

True hypotheses

$\mathcal{T} \subseteq \{1, \ldots, m\}$ indices of true hypotheses

Intersection hypothesis

$H_C = \bigcap_{i \in C} H_i$, for $C \subseteq \{1, \ldots, m\}$

Closure

Collection of all $2^m - 1$ intersection hypotheses

Closed Testing: ingredients

Marcus, Peritz and Gabriel (1976)

Fundamental principle of FWER control

Local test

Valid α-level test for every intersection hypothesis

Example of local test

- global test
- Bonferroni local test
- Simes test

Closed testing: procedure

Raw rejections

Hypotheses rejected by the local test (with $p < \alpha$)

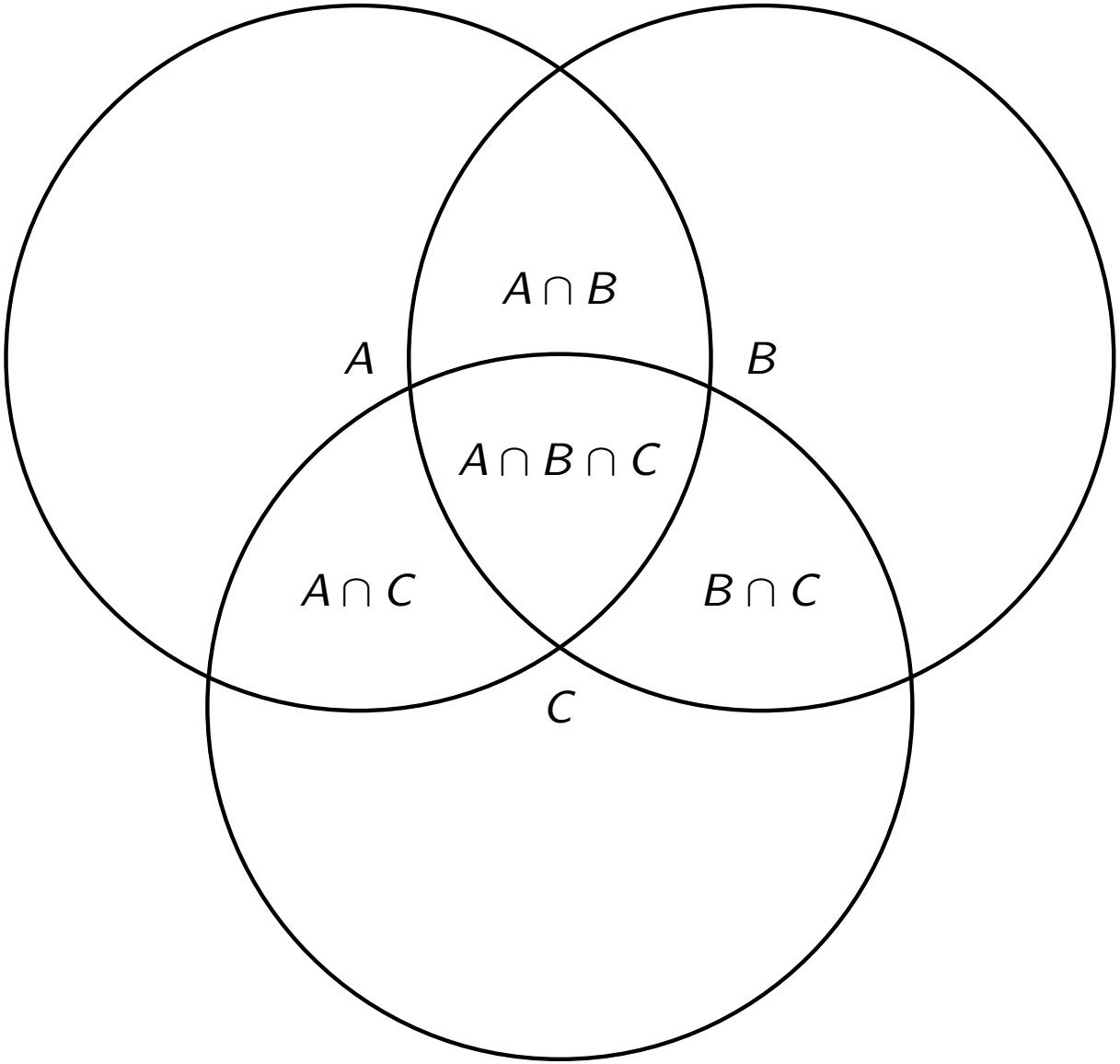
Multiplicity-corrected rejections

Reject hypotheses H_C, if H_J rejected for every $J \supseteq H$

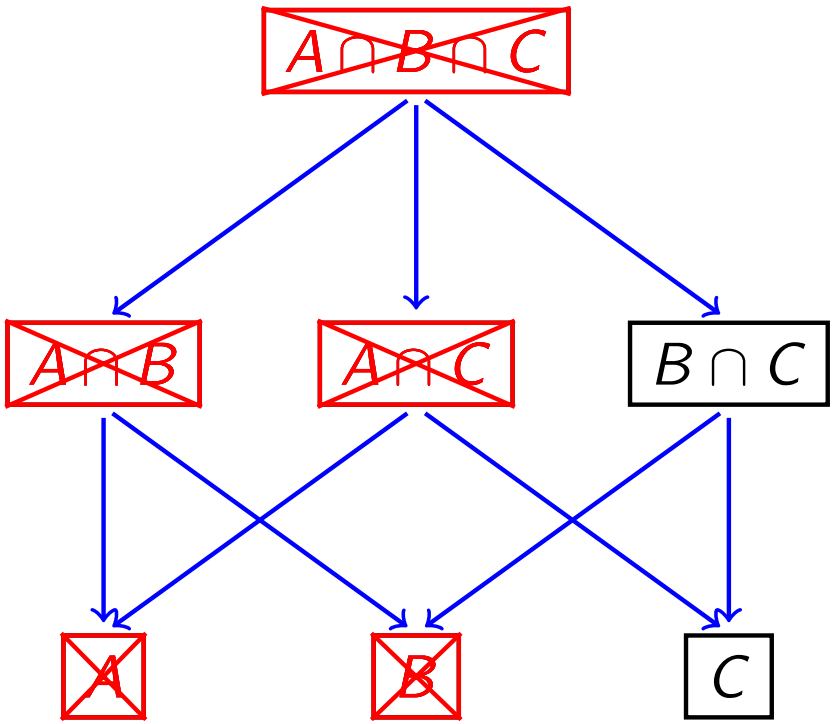
Statement

With probability at least $1 - \alpha$, no true (intersection) hypothesis rejected

Closed testing (graphically)



Closed testing: graphically



Multiplicity-corrected rejections

Closed testing on gene sets

Huge number of tests performed

All $2^m - 1$ possible gene sets

No Bonferroni factors

All tests performed at level α

Replaced by constraint

Only reject gene set R if all gene sets containing R are rejected

Still familywise error control

Probability $(1 - \alpha)$: no false rejection

Shortcut: fast algorithms

Start with m hypothesis

Require $2^m - 1$ tests in closed testing

Use Simes as a local test

Obtain adjusted test results in $m \log(m)$ time

Reason

Simple structure of test $\rightarrow$ implications between test results

FWER results

Hommel's method

- Use closed testing plus Simes
- Report all rejected elementary hypotheses
- Use shortcut to avoid exponential calculations

Hochberg's method

- Also uses closed testing plus Simes
- Even faster simpler shortcut
- Sacrifices power for quick calculations

Adjusted p-values for Hochberg's procedure

Start with p-values for m hypotheses

1. Sort the p-values $p_{(1)}, \ldots, p_{(m)}$.
2. Multiply each $p_{(i)}$ by its adjustment factor $a_i = m - i + 1$, $i = 1, \ldots, m$
3. If the multiplication in step 2 violates the original ordering, repair this: decrease the highest p-value in all violating pairs:

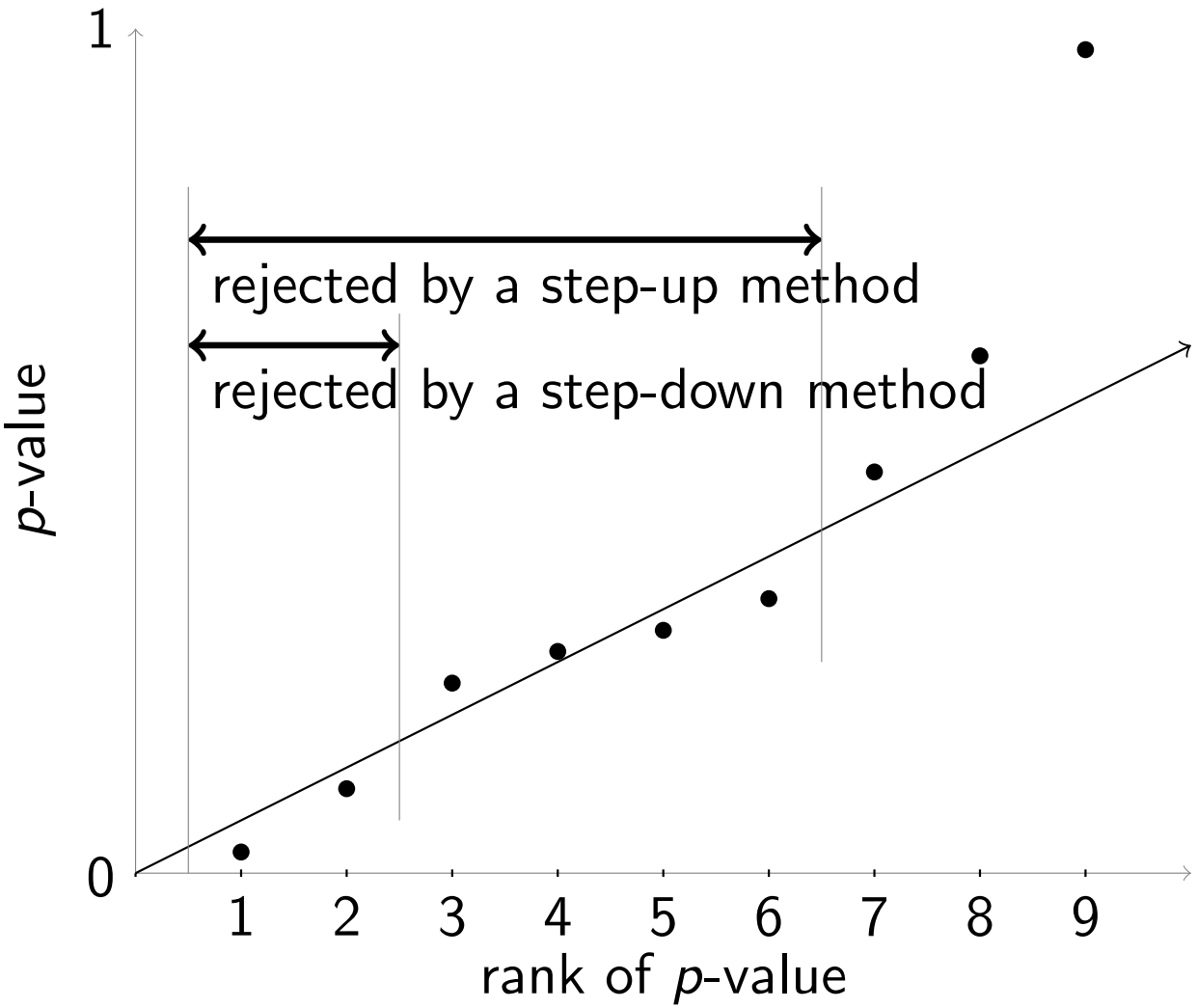
$$\tilde{p}_{(i)} = \min_{j=i,\ldots,m} a_j p_{(j)}$$

4. Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i.

Compare with Holm

Same critical values, but step-up rather than step-down

Step-down and step-up



Back to flexibility

Selection and aggregation

Probability statements must be post hoc valid

Desired statements

- Is there any signal in this set?
- What is the false discovery proportion (FDP) of this set?
- What is the FDP of this collection of sets?

Single set statements

Closed testing

All possible sets tested; FWER control

Is there signal in this set?

Automatically answered for all sets

Post hoc validity (flexibility)

Automatic from FWER control

False discovery proportion

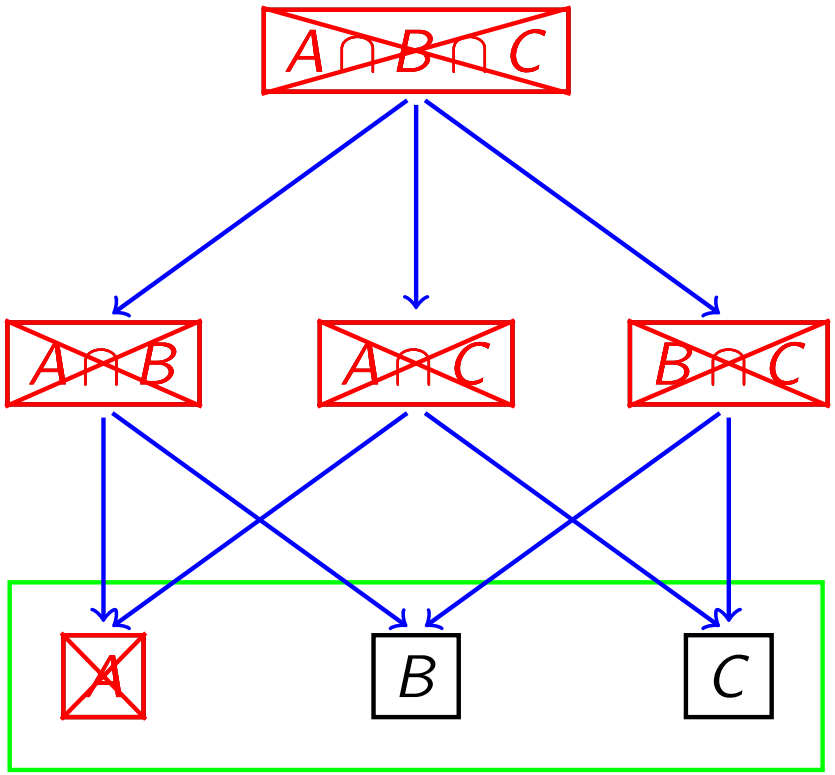
Set R of interest

- Chosen post hoc
- Starting point for further analysis

Questions

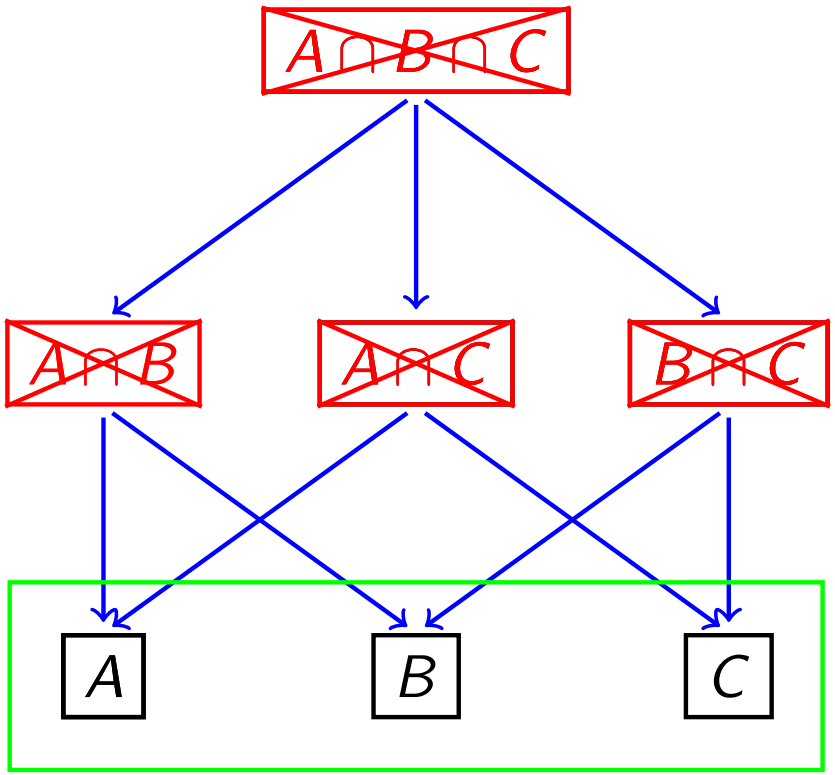
- Number of true/false hypotheses in R
- Proportion of true/false hypotheses in R

Inference on false discovery proportion



Many weak effects

Bounding the false discovery proportion



With probability $1 - \alpha$: ≤ 1 true null hypothesis among A, B, C

Results

Closed testing gives flexibility

- Confidence statements for FDP for all sets
- Also: point estimate for FDP (use $\alpha = 0.50$)

Simultaneous coverage for free

Probability $1 - \alpha$: closed testing makes no error:
all confidence statements simultaneously correct

Simultaneous control over all R

Consequence: coverage robust against post hoc selection of R

Simultaneous confidence intervals

Regular confidence interval

- Each individually covers true parameter with probability $1 - \alpha$
- Some confidence intervals cover, some don't
- Some non-covering intervals are present in every experiment
- Selected (interesting) confidence interval is likely non-covering

Simultaneous confidence intervals

- With probability $1 - \alpha$, all intervals cover simultaneously
- With probability $1 - \alpha$ no non-covering intervals present
- With probability $1 - \alpha$ the selected interval covers
- Simultaneous intervals are robust against selection

Confidence statements for all sets

R	confidence set for $\tau(R)$	confidence set for $\phi(R)$
$\{A\}$	$\{0,1\}$	$\{0,1\}$
$\{B\}$	$\{0,1\}$	$\{0,1\}$
$\{C\}$	$\{0,1\}$	$\{0,1\}$
$\{A,B\}$	$\{0,1\}$	$\{1,2\}$
$\{A,C\}$	$\{0,1\}$	$\{1,2\}$
$\{B,C\}$	$\{0,1\}$	$\{1,2\}$
$\{A,B,C\}$	$\{0,1\}$	$\{2,3\}$

Links with other procedures

Link with Benjamini & Hochberg FDR control

- Same assumptions, same weak FWER control
- FDR rejected set R: 95% conf. FDP $\leq (\#R - 1)/\#R$
- FDR rejected set R always has estimated FDP $\leq 10\%$

Link with Hommel/Bonferroni FWER procedures

- Hommel-rejected set always has 95% conf. FDP $= 0$
- Hommel is a more powerful variant of Bonferroni
- But obtain additional FDP statements for other sets

Data analysis in genomics

Top diff. expression

Gen	p-waarde
XDH	5.5e-10
NEK3	6.7e-7
TAF5	7.1e-7
CYP2A7	1.6e-6
NAT2	1.8e-6
ZNF19	2.6e-6
SKP1	2.7e-6
NAT1	3.1e-6
GDF3	2.0e-5
CCDC25	2.1e-5
⋮	⋮

False discovery rate control

Expected: $< 5\%$ false positives

Practice

Genes chosen for further analysis

How many false positives to expect?

95% conf.: max. 1 false positive

Point estimate

No false positives

Example: Rosenwald DLBCL data

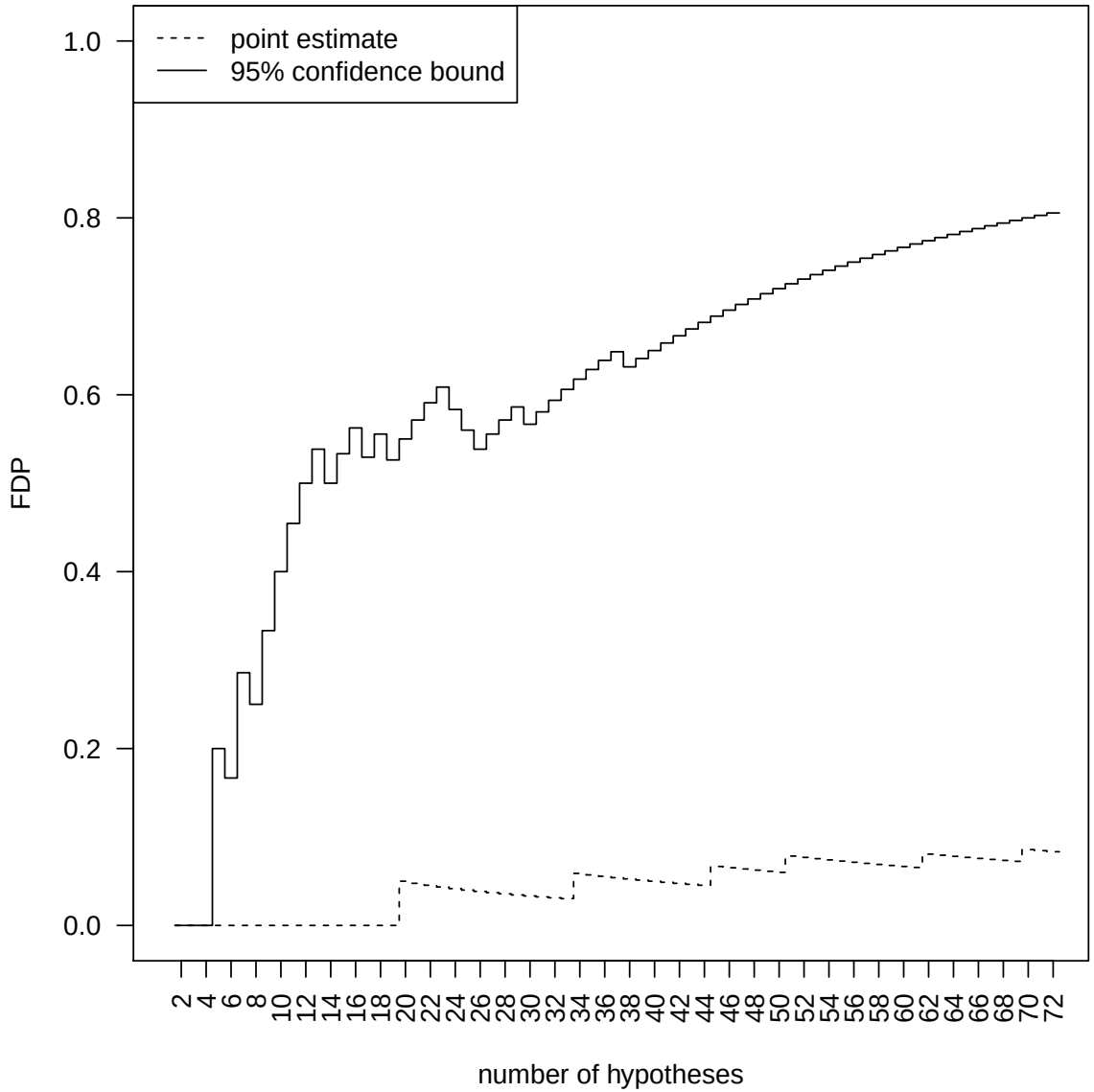
Data

240 diffuse large B-cell lymphoma patients; 7399 hypotheses

Classical results

- Bonferroni, Holm, Hocherg, Hommel: 4 hypotheses
- Benjamini and Hochberg: 72 hypotheses

FDP estimates and bounds: top k p-values



In R

```
# start with a named vector of sorted p-values
ps

# prepare the closed testing procedure
hom <- hommelFast(ps)

# how many false hypotheses in the top 10?
pickSimes(hom, 1:10)
10 hypotheses selected. At confidence level 0.95:
False null-hypotheses >= 6; True null-hypotheses <= 4.

# or in some other group
pickSimes(hom, c('RAF', 'ERK1', 'ERK2', 'MEK1', 'MEK2'))
5 hypotheses selected. At confidence level 0.95:
False null-hypotheses >= 2; True null-hypotheses <= 3.
```

In R

```
# estimate
pickSimes(hom, 1:10, alpha=0.5)
10 hypotheses selected. At confidence level 0.5:
False null-hypotheses >= 10; True null-hypotheses <= 0.

# make a curve like in the plot
false <- curveSimes(hom, 1:72)
fdp <- 1-false/1:72

# curve for the fdp
plot(fdp, type='S')
```

Subsetting property

FDP type statements

Never have the subsetting property

However

Simultaneous FDP statements over all subsets

Advantages

- FDP statement for the set of real interest
- FDP statements for subsets of that set
- All simultaneous $\rightarrow$ confidence guaranteed

Multiple testing in genomics

Flexibility is important

Users will always aggregate or select

Flexibility is limited

- FWER control allows it
- FDR control does not

FDP confidence methods

Combine flexibility and control

FDP confidence based on Simes

Closed testing

Tests all intersection hypotheses ('gene set tests')

Simes' inequality

Makes this computationally feasible

Result

- Great flexibility
- Point estimate and confidence bound of FDP for any set
- Simultaneous $\rightarrow$ post hoc valid
- Little loss of power relative to Benjamini & Hochberg