False Discovery Rate Control in multiple testing problems

Ruth Heller

Tel-Aviv university
www.math.tau.ac.il/∼ruheller

Outline

- Motivation.
- The FDR criterion and the Benjamini-Hochberg procedure for FDR control.
- Estimating the proportion of null hypotheses.
- Extension 1: adjustment for discrete tests.
- Extension 2: replicability analysis.

A microarray example

An observational study by Spira et al. (2004).

The Data:

- 33 smokers, 23 non-smokers.
- Human epithelial cells from brushings of the right main bronchus proximal to the right upper lobe of the lung.
- 9968 (log) expression profiles from HG-U133A Affymetrics chip.

Gene ID	Smkr1	...	Smkr33	Non-smkr1	...	Non-smkr23
210149	6.8	...	7.0	6.5	...	6.6
...	...	...	...	...	...	...

Microarray example: the hypotheses

- The null hypothesis is that the distribution of gene expression is the same for smokers and non-smokers

$$H : F_{Smkr} = F_{Non-smkr}$$

- The alternative hypothesis is that the distributions differ, ie that smoking is associated with the gene expression.

Is gene ID 210149 associated with smoking?

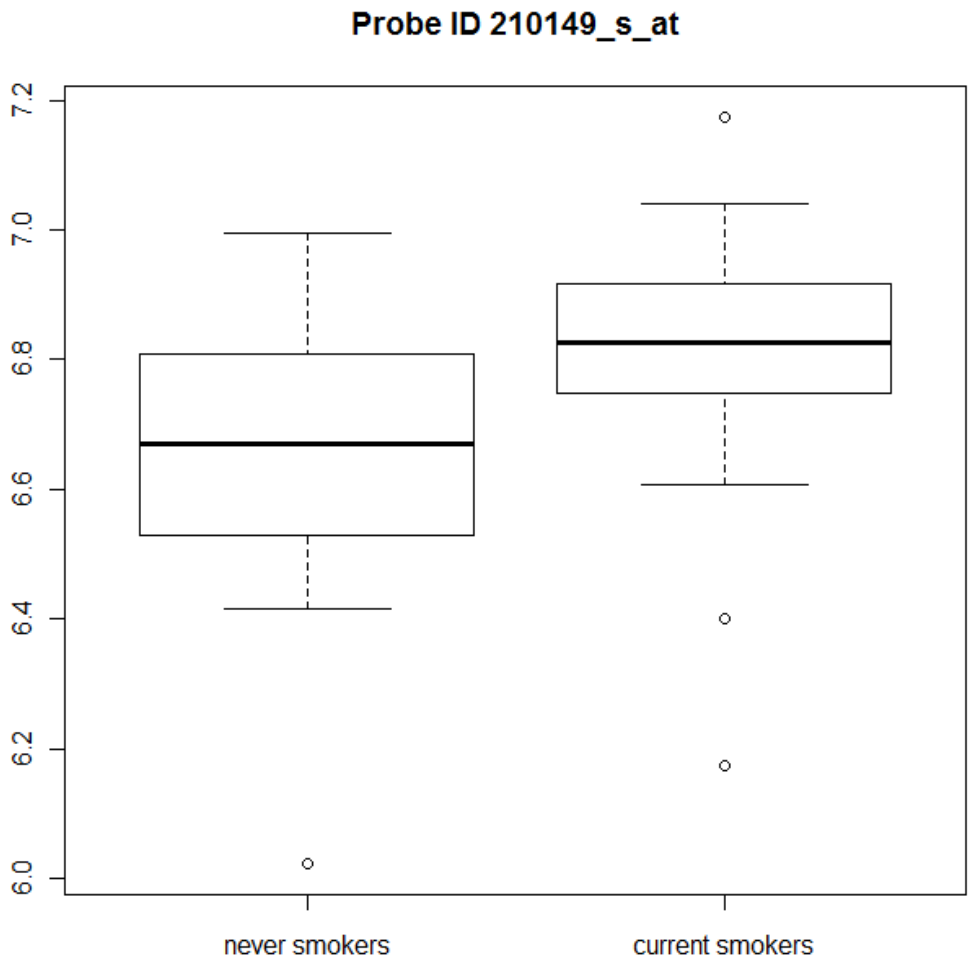


Figure: Wilcoxon rank sum test p-value 0.01.

Microarray example continued

Is gene ID 210149 interesting?

- It is below the 0.05 cut-off.
- It is just one among 9968 genes.
- We expect $0.01 \times 9968 \approx 100$ p-values to be at least as extreme as that of gene 210149 when none of the genes are associated with smoking.

What happens if we disregard multiplicity?

Suppose we test each hypothesis at level α.

- The probability of one or more false rejections rapidly increases with the number of hypotheses m.
- When the number of true nulls is large, we shall be nearly certain to reject some of them.
- Example: for m true null hypotheses tested at level $\alpha = 0.05$, the probability of rejecting at least one hypothesis is

m	1.00	2.0	10.0	100.00	1000
Pr(at leat one false rejection)	0.05	0.1	0.4	0.99	~1

Multiple Hypothesis Testing Problems in Genomics

- High-Throughput microarray gene expression analysis.
  - Test thousands of genes simultaneously for differential expression between groups.
- Genome-wide association studies.
  - Test hundreds of thousands of SNPs for association between SNP and disease.

Challenges:

- Many null hypotheses.
- Complex and unknown dependence structure among test statistics.

Fishing Expeditions

- Observational studies, as well as controlled experiments, may record many variables over time on several comparison groups. At the analysis stage, "significant" differences between the groups are sought.
- How to assess whether an extreme result is really interesting?
  - May regard the results of a fishing expedition as merely providing suggestions for further experiments.
  - May split the data randomly and use the first part for fishing and the second half for testing the promising hypothesis from the fishing expedition on the first part of the data.
  - Conduct each individual hypothesis test at a smaller significance level than the nominal level α.

Sources of multiplicity

- Multiple outcomes.
- Multiple treatments.
- Multiple time points.
- Multiple groups.
- Multiple (different) tests of the same hypothesis.
- Variable and model selection.

The multiplicity problem increases as data becomes easier to obtain.

Families of hypotheses

- The term "family" refers to the collection of hypotheses $H_1, \ldots, H_m$ that is being considered for joint testing.
- What hypotheses are to be treated jointly as a family depends on the problem.
- For a family of hypotheses, it is meaningful to take into account some combined measure of error.

Notation

- $I_0 \subseteq \{1, \ldots, m\}$ is the index of true null hypotheses.
- $m_0 = |I_0|$ is the number of true null hypotheses.
- R is the number of hypotheses rejected.
- V is the number of hypotheses rejected in error.

sub-family	declared non-significant	declared significant	total
I_0	$m_0 - V$	V	m_0
$\{1, \ldots, m\}/I_0$	$m - R - (m_0 - V)$	$R - V$	$m - m_0$
Total	$m - R$	R	m

The family wise error rate (FWER)

- The FWER is the probability of making even one type I error in the family:

$$FWER = Prob(V \geq 1).$$

- Thus, by assuring $FWER \leq \alpha$, the probability of making even one type I error in the family is controlled at level α.
- The best known procedure for FWER control is the Bonferroni procedure, which tests each hypothesis at level $\alpha_{Bon} = \alpha/m$:

$$FWER = Pr(V \geq 1) \leq E(V) = \sum_{i \in I_0} \alpha/m \leq \alpha.$$

Microarray example continued

- With $m = 9968$, p-values must be below $0.05/m \approx 5 \times 10^{-6}$ to be significant by the Bonferroni multiple testing procedure.
- 111 genes are discovered.
- The FWER is controlled at level 0.05, but the threshold is orders of magnitude smaller than the conventional $\alpha = 0.05$ level.

The dilemma

- Not controlling for multiplicity, working at 0.05, we expect about 500 (statistical) discoveries possibly just due to noise.
- Controlling for FWER, working at 0.05, a single comparison has to be significant at 0.000005 to make it to the list of discoveries.

Outline

- Motivation.
- The FDR criterion and the Benjamini-Hochberg (BH) procedure for FDR control.
- Estimating the proportion of null hypotheses.
- Extension 1: adjustment for discrete tests.
- Extension 2: replicability analysis.

The False Discovery Rate (FDR) Criterion

[Benjamini and Hochberg, 1995]

- The proportion of the rejected null hypotheses which are erroneously rejected is

$$\frac{V}{\max(R,1)}.$$

- The FDR is the expected proportion

$$FDR = E\left(\frac{V}{\max(R,1)}\right).$$

Properties of the FDR

- If $m = m_0$, the FDR is identical to the FWER:

$$FDR = E\left(\frac{V}{\max(V,1)}\right) = EI(V > 0) = Pr(V > 0) = FWER.$$

- If $m_0 < m$, the FDR is smaller than the FWER:

$$FDR = E\left(\frac{V}{\max(R,1)}\right) \leq EI(V > 0) = FWER.$$

Therefore there is room for improving detection power with FDR control over FWER control.

The FDR is adaptive and scalable

- The FDR is adaptive:
  - 2 false among 50 discovered is bearable;
  - 2 false among 4 discovered unbearable.
- The FDR is scalable:
  - 2 false among 50 discovered is bearable;
  - 20 false among 500 discovered is bearable to the same extent.

Switching to FDR control

- FWER control is appropriate when it is important not to make any error.
- The power advantage of FDR control over FWER may be large.
- FDR is more relevant, if willing to tolerate few false positives as long as it is a small fraction of the discoveries.

Motivation for the BH step-up Procedure

- We want to find the largest cut-off T_q, so that

$$E\left(\frac{\sum_{i\in I_0} I[P_i \le T_q]}{\max(\sum_{i=1}^{m} I[P_i \le T_q], 1)}\right) \le q.$$

- Note that for large m,

$$E\left(\frac{\sum_{i\in I_0} I[P_i \le T_q]}{\sum_{i=1}^{m} I[P_i \le T(q)]}\right) \approx \frac{m_0 T_q}{\sum_{i=1}^{m} I[P_i \le T_q]}.$$

- Note that the largest cut-off is one of the ordered p-values,

$$p_{(1)} \le \ldots \le p_{(m)}.$$

- Therefore, how about finding the largest i such that

$$\frac{m_0 p(i)}{i} \le q?$$

- Since m_0 is unknown, find the largest i such that

$$\frac{m p(i)}{i} \le q.$$

The BH step-up Procedure

Let p_i be the observed p-value for test H_i, $i = 1, \ldots, m$.

- Sort the p-values:
$$p_{(1)} \leq \cdots \leq p_{(m)}.$$
- Let $R = \max\{i : p_{(i)} \leq \frac{i}{m} q\}$.
- Reject $H_{(1)}, \ldots, H_{(R)}$.

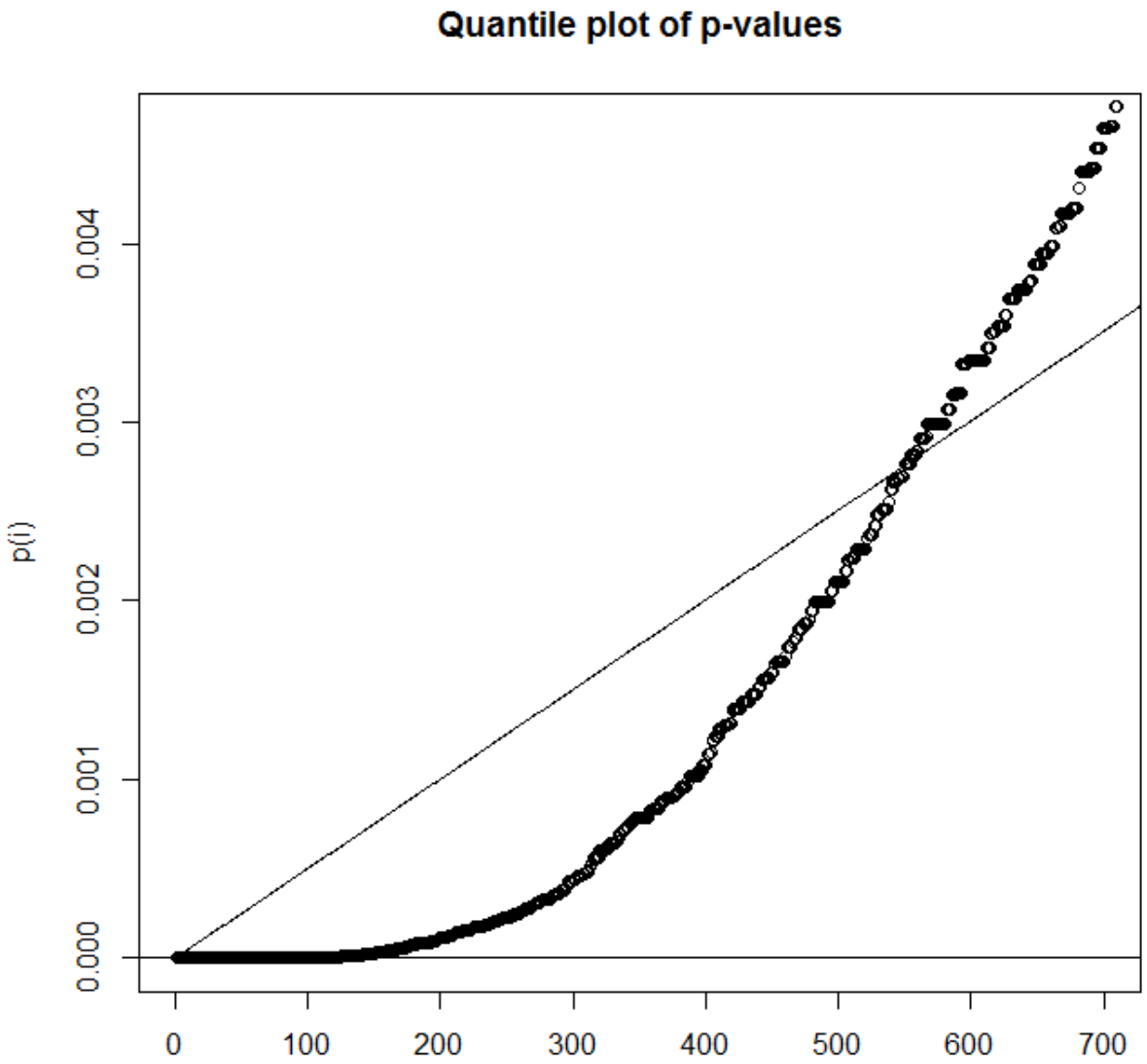
A numerical example

$m = 4, q = 0.05$:

Ordered p-value	BH critical value $\frac{i}{m}q$	Bonferroni's critical value $\frac{1}{m}q$
0.0130	0.0125	0.0125
0.0142	$\frac{2}{4} \times 0.05 = 0.025$	0.0125
0.0191	$\frac{3}{4} \times 0.05 = 0.0375$	0.0125
0.1986	$\frac{4}{4} \times 0.05 = 0.05$	0.0125

How many hypotheses are rejected by the BH procedure at level q=0.05?

The graphical way to look at the BH procedure in the microarray example



Adjusted p-values for a multiple testing procedure (MTP)

Definition: an *MTP* adjusted p-value for a hypothesis H_i is the smallest nominal level of the *MTP* at which H_i would be rejected, given the value of all test statistics involved.

- The Bonferroni-adjusted p-value for hypothesis H_i is

$$p_i^{Bonfadj} = mp_i.$$

- The BH-adjusted p-value for hypothesis H_i is

$$p_{(i)}^{BHadj} = \min_{j \geq i} \{ \frac{mp_{(j)}}{j} \}.$$

The BH-adjusted p-values are smaller (or equal to) the Bonferroni-adjusted p-values.

The numerical example

$m = 4, q = 0.05$:

$p_{(i)}$	$\frac{mp_{(i)}}{i}$	BH adjusted p-value	Bonferroni adjusted p-value
0.0130	0.0520	0.0255	0.0520
0.0142	0.0284	0.0255	0.0568
0.0191	0.0255	0.0255	0.0764
0.1986	0.1986	0.1986	0.7944

R code for computing adjusted p-values

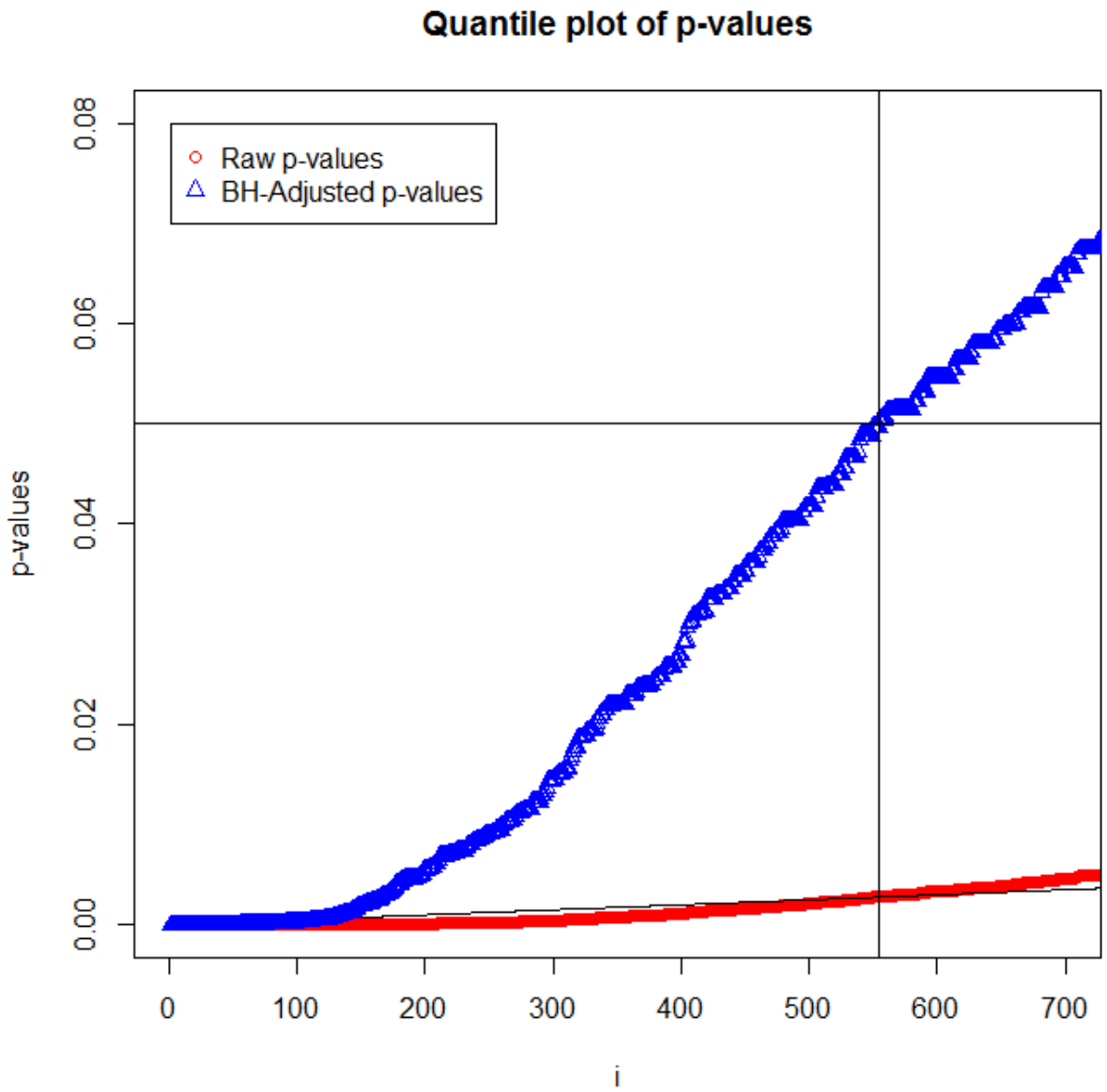
```
padjBonf = p.adjust(pv,method="bonferroni")

padjBH=p.adjust(pv,method="BH")

sum(padjBonf<=0.05) #111

sum(padjBH<=0.05) #554
```

The graphical way to look at the BH-adjusted *p*-values



FDR control of the BH procedure

For the BH procedure, if the p-values are:

- independent and uniformly distributed under the null, then [Benjamini and Hochberg, 1995] prove that

$$FDR = \frac{m_0}{m} q.$$

- independent, or positive regression dependent on the subset of null hypotheses (PRDS), then [Benjamini and Yekutieli, 2001] prove that

$$FDR \leq \frac{m_0}{m} q.$$

- dependent, then [Benjamini and Yekutieli, 2001] prove that

$$FDR \leq \frac{m_0}{m} q(1 + 1/2 + 1/3 + \ldots + 1/m) \approx \frac{m_0}{m} q \log(m).$$

Property PRDS

Property PRDS. For any increasing set D, and for each $i \in I_0$, $Pr(\vec{P} \in D | P_i = x)$ is non-decreasing in x.

- The conditioning is one variable at a time.
- The conditioning is required to hold only for the subset of the null hypotheses I_0.

An example of PRDS dependency: multivariate normal test statistics with positive correlation

$$X \sim N(\mu, \Sigma) \quad , \mu_i = 0 \quad i \in I_0, \mu_i > 0 \quad i \notin I_0, \quad \Sigma_{ij} \geq 0.$$

BH procedure on non-PRDS dependent p-values

- [Benjamini and Yekutieli, 2001] prove that the FDR is controlled for the following dependency which is not PRDS: one-sided correlated t-tests: $\vec{t} = X/S$, where

$$X \sim N(\mu, \Sigma) \quad , \mu_i = 0 \quad i \in I_0, \mu_i > 0 \quad i \notin I_0, \quad \Sigma_{ij} \geq 0,$$

  and S is the pooled standard deviation.
- In many simulations examined the FDR was shown to be conservative. For example,
  1. Absolute values of correlated normal test statistics [Reiner, 2007].
  2. Pairwise comparisons [Yekutieli, 2008b].

The BH procedure on p-values appears to control the FDR in most circumstances that are not highly artificial [Yekutieli, 2008a].

Outline

- Motivation.
- The FDR criterion and the Benjamini-Hochberg (BH) procedure for FDR control.
- Estimating the proportion of null hypotheses.
- Extension 1: adjustment for discrete tests.
- Extension 2: replicability analysis.

Estimating the proportion of null hypotheses

- The BH procedure is conservative by the factor $\pi_0 = \frac{m_0}{m}$.
- If π_0 were known, the BH procedure could be applied at level q/π_0 for FDR control at level q.
- Many methods have been developed to estimate π_0.

The estimator of [Storey and Tibshirani, 2003]

- Select a large enough tuning parameter $\lambda \in (0,1)$, say $\lambda = 0.5$.
- Consider $U = \sum_{i \in I_0} I[P_i > \lambda]$, the number of p-values above λ.
- The expectation of U is $E(U) = m_0(1-\lambda)$.
- As long as each test has reasonable power, most p-values above λ should be null. The estimator

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I[P_i > \lambda]}{m(1-\lambda)}$$

results from

$$\hat{E}(U) = \hat{m}_0(1-\lambda) = \sum_{i=1}^{m} I[P_i > \lambda].$$

This is perhaps the most widely used estimator of π_0, and it is implemented in the R packages *SAM* and *locfdr*.

The estimator of [Storey et al., 2004]

- The modified estimator

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I[P_i > \lambda] + 1}{m(1-\lambda)}$$

  is greater than zero for $\lambda > 0$.
- Considering the discoveries from a BH procedure at level $q/\hat{\pi}_0$, the FDR was proven to be controlled for independent p-values at the nominal level q ([Storey et al., 2004], [Benjamini et al., 2006]).
- The modified BH procedure is not robust to deviations from independence. Specifically, the FDR may not be controlled for p-values with property PRDS ([Benjamini et al., 2006]).

The estimator of [Benjamini et al., 2006]

- Use the BH procedure at level q. Denote by R_1 the number of discoveries.
- Estimate π_0 conservatively by $\hat{\pi}_0 = \frac{\sum_{i=1}^{m} I[P_i > R_1 q/m]}{m(1-q)}$.

Considering the discoveries from a BH procedure at level $q/\hat{\pi}_0$, the FDR was proven to be controlled for independent p-values at the nominal level q.
Moreover, it is conjectured to control the FDR at the nominal level q for p-values with property PRDS.

Power of a multiple testing procedure

The two most relevant measures of power with FDR control are:

- The average power

$$\frac{E(R-V)}{m_1},$$

  that we want to maximize.

- The false non-discovery rate (Genovese and Wasserman 2002)

$$FNR = E(\frac{m - m_0 - (R-V)}{\max(m-R,1)}),$$

  that we want to minimize.

Simulation results in [Benjamini et al., 2006] for independent p-values

Table: Power comparison for $m = 256$ hypotheses, the power relative to the oracle procedure that uses BH at level $0.05/\pi_0$. The values of μ_i where zero for null hypotheses, and $\{1, 2, 3, 4\}$ for non-null hypotheses.

Method	$\pi_0 = 0.75$	$\pi_0 = 0.50$
BH	0.941	0.874
BH with estimator of [Benjamini et al., 2006]	0.959	0.926
BH with estimator of [Storey et al., 2004]	0.993	0.982

Simulation results in [Benjamini et al., 2006]

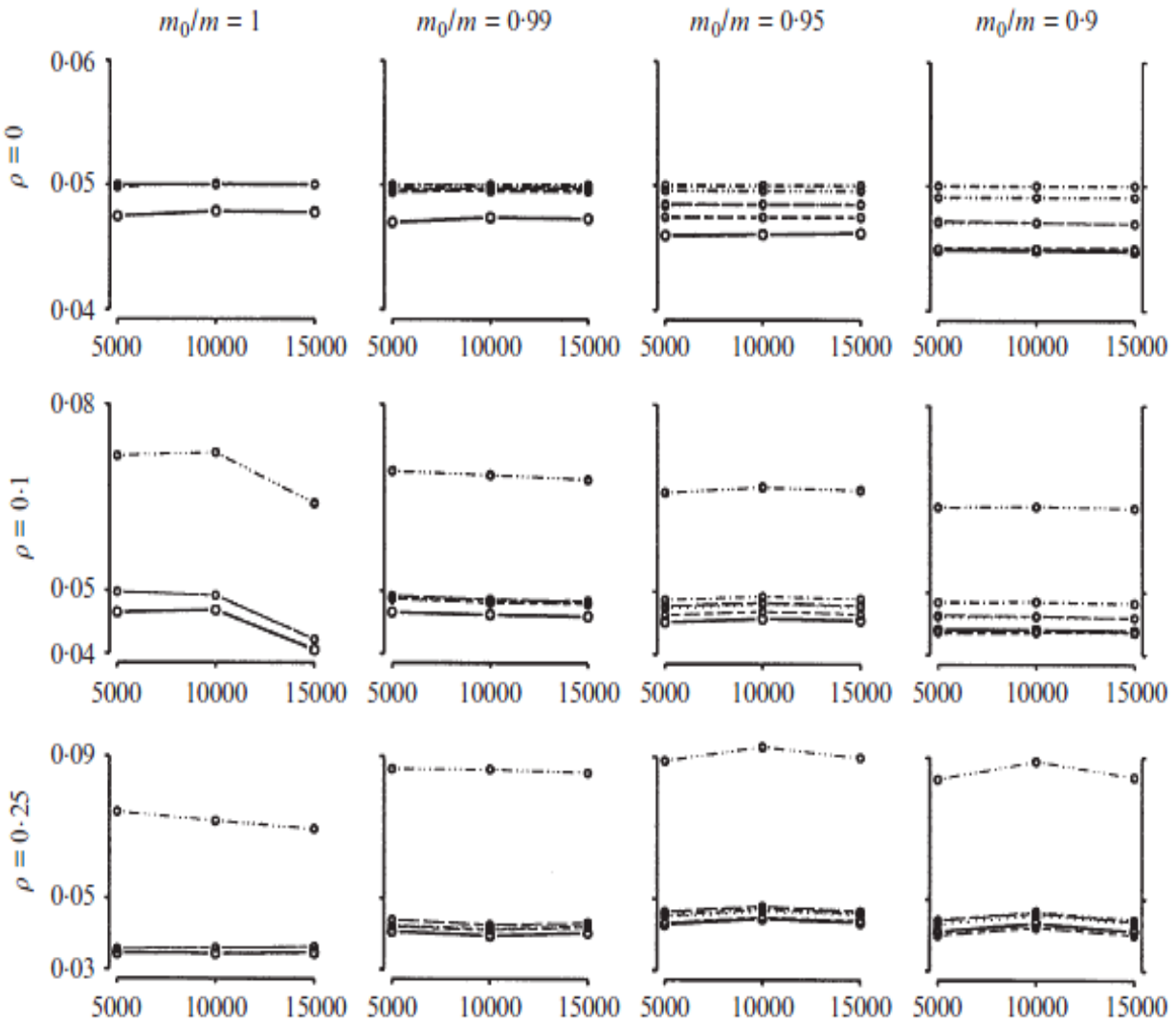


Fig. 2: Simulation study. Estimated FDR values for $m = 5000, 10\,000$ and $15\,000$ and $\rho = 0, 0{\cdot}1$ and $0{\cdot}25$. results for procedure TST, solid line; MST, dotted line; ORC, dotted-dashed; ABH, dashed; M-S-HLF, dashed triple-dotted; LSU, short-dash long-dash.

Simulation results in [Benjamini et al., 2006]

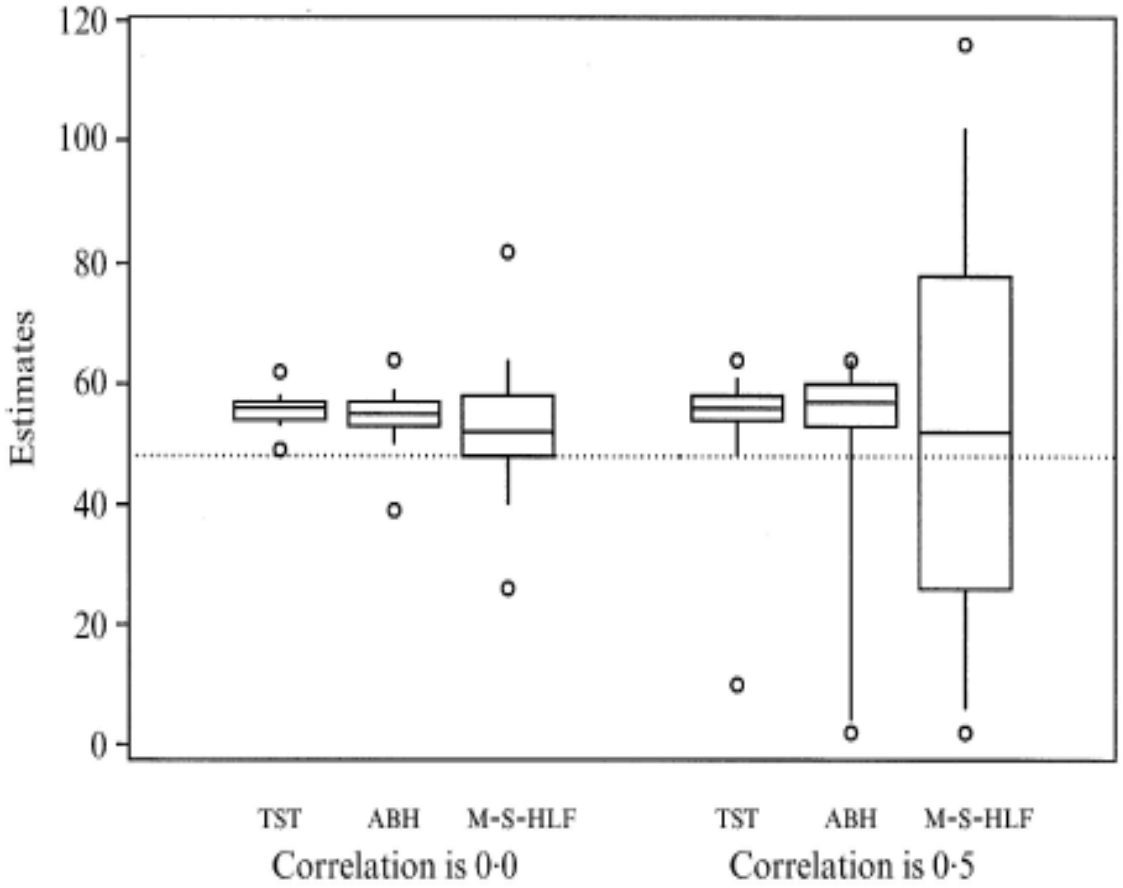


Fig. 3: The simulated distribution of the estimators $\hat{m}_0$ used in the TST, ABH and M-S-HLF adaptive procedures for estimating the number of true hypotheses with independent and positively correlated statistics for the case of $m_0 = 48$ and $m = 64$. Each box displays the median and quartiles as usual. The whiskers extend to the 5% and the 95% quantiles. The circles are located at the extremes, i.e. the 0·01% and 99·99% percentiles.

Outline

- Motivation.
- The FDR criterion and the Benjamini-Hochberg (BH) procedure for FDR control.
- Estimating the proportion of null hypotheses.
- Extension 1: adjustment for discrete tests.
- Extension 2: replicability analysis.

A Small Example

Table: Table relating treatment to adverse event, for 10 hospitals. The p-value was computed from a one-sided Fisher's exact test for 2×2 tables.

	X_{11}	X_{12}	X_{21}	X_{22}	p value	$midP$ value
1	1.000	15.000	13.000	3.000	0.000	0.000
2	2.000	36.000	12.000	20.000	0.001	0.000
3	1.000	14.000	7.000	6.000	0.009	0.005
4	10.000	30.000	12.000	8.000	0.009	0.006
5	0.000	20.000	5.000	18.000	0.035	0.017
6	2.000	5.000	7.000	2.000	0.072	0.039
7	8.000	16.000	15.000	12.000	0.095	0.062
8	3.000	11.000	7.000	15.000	0.389	0.267
9	5.000	12.000	5.000	10.000	0.555	0.411
10	7.000	14.000	5.000	20.000	0.914	0.834

The conservatism of discrete tests

- The p-value distribution under the null is stochastically larger than the uniform, resulting in a conservative test

$$Pr_{H_0}(P \leq \alpha) \leq \alpha.$$

- Lancaster (1961) suggested to use the $midP$-value:

$$midP = Pr_{H_0}(P < p) + \frac{1}{2} \times Pr_{H_0}(P = p)$$

- The actual level of the test is closer to the nominal level if the decision is based on the $midP$-value and not on the p-value.

The conservatism of the BH procedure for discrete tests

For independent p-values (Benjamini and Hochberg, 1995):

$$FDR \leq \frac{m_0}{m} q.$$

- If the null distribution of the p-values is uniform, then equality holds because key terms in the sum of the FDR expression are

$$Pr_{H_i}(P_i \leq \frac{k}{m} q) = \frac{k}{m} q.$$

- For discrete test statistics,

$$Pr_{H_i}(P_i \leq \frac{k}{m} q) \leq \frac{k}{m} q.$$

  The greater the gaps between LHS and RHS, the smaller the true FDR level of the BH procedure.
- Suggestion: use the BH procedure on *midP*-values instead of on p-values.

The Small Example with BH adjusted p and $midP$ values

	X_{11}	X_{12}	X_{21}	X_{22}	p value	$midP$ value	BH adjusted	$midP$+BH adjusted
1	1.000	15.000	13.000	3.000	0.000	0.000	0.000	0.000
2	2.000	36.000	12.000	20.000	0.001	0.000	0.004	0.002
3	1.000	14.000	7.000	6.000	0.009	0.005	0.023	0.014
4	10.000	30.000	12.000	8.000	0.009	0.006	0.023	0.014
5	0.000	20.000	5.000	18.000	0.035	0.017	0.070	0.035
6	2.000	5.000	7.000	2.000	0.072	0.039	0.119	0.064
7	8.000	16.000	15.000	12.000	0.095	0.062	0.135	0.089
8	3.000	11.000	7.000	15.000	0.389	0.267	0.486	0.333
9	5.000	12.000	5.000	10.000	0.555	0.411	0.617	0.457
10	7.000	14.000	5.000	20.000	0.914	0.834	0.914	0.834

Further adjustment for discreteness

Apply a preliminary step of removing from consideration null hypotheses for which rejection is not possible (Tarone, 1990, Gilbert, 2005, `http://www.math.tau.ac.il/~ruheller/Papers/1112.4627v2.pdf`).

Definition

The Tarone+*midP* BH procedure at level q is the following two-step procedure:

1. Remove all hypotheses for which the most extreme possible p-value is above the threshold $\alpha*$.
2. Apply the BH procedure on the *midP*-values of the reduced family of hypotheses.

R package *discreteMTP* available from CRAN.

Outline

- Motivation.
- The FDR criterion and the Benjamini-Hochberg (BH) procedure for FDR control.
- Estimating the proportion of null hypotheses.
- Extension 1: adjustment for discrete tests.
- Extension 2: replicability analysis.

A motivating example

In genome-wide association studies (GWAS), it is customary that a primary study is followed by an independent study.

- In the primary study, hundreds of thousands of single nucleotide polymorphisms (SNPs) are tested for association with the disease.
- In the follow-up study, a handful of promising SNPs are tested for association with the disease.
- If the p-value is fairly small in the follow-up study it is informally regarded as a replicated finding.

Table 1 of [Bis, J. et al., 2012]

Locus	SNP	Gene	Primary	Follow-up	Combined
2q24	rs6741949	DPP4	5.2×10^{-8}	0.7	2.9×10^{-7}
9q33	rs7852872	ASTN2	1.0×10^{-7}	0.2	1.0×10^{-7}
12q14	rs17178006	MSRB3	5.5×10^{-9}	0.002	5.3×10^{-11}
	rs6581612	WIF1	2.2×10^{-8}	0.0007	7.1×10^{-11}
12q24	rs7294919	HRK	4.8×10^{-8}	5.8×10^{-5}	2.9×10^{-11}

Can we order the SNPs by evidence towards replicability?
Which SNPs are significant when testing for replicability?

The BH procedure for replicability analysis

- The approach in [Benjamini and Heller, 2008]
  - Take the maximum p-values across studies.
  - Next, apply the BH procedure at level q on these p-values, correcting for multiplicity of the number of hypotheses in the primary study.
- A new approach in [Bogomolov and Heller, 2012] for two studies:
  - A two-dimensional variant of the BH procedure.

A new procedure for replicability analysis with FDR control

- The BH procedure can be expressed as follows:
  1. $R = \max\{r : \sum_{j=1}^{m} I[p_j \leq rq/m] = r\}$.
  2. Reject the R hypotheses with smallest p-values.
- Replicability analysis procedure with parameters (q_1, q), where $0 < q_1 < q < 1$:
  1. $\mathcal{R}_1 =$ Selected set for follow-up based on data in primary study,

$$R_1 = |\mathcal{R}_1|.$$

  2. $R_2 \triangleq \max\left\{r : \sum_{j\in\mathcal{R}_1} \mathbf{I}[(p_{1j}, p_{2j}) \leq \left(\frac{rq_1}{m}, \frac{r(q-q_1)}{R_1}\right)] = r\right\}$. Then the indices of replicated findings are

$$\mathcal{R}_2 = \{j : (p_{1j}, p_{2j}) \leq \left(\frac{R_2 q_1}{m}, \frac{R_2(q-q_1)}{R_1}\right), j \in \mathcal{R}_1\}.$$

The adjusted p-values

The results of the FDR-Replicability Procedure can be reported in terms of FDR-replicability adjusted p-values. Let $c = q_1/q$,

$$Z_j = \max\left(\frac{mP_{1j}}{c}, \frac{R_1 P_{2j}}{1-c}\right), j \in \mathcal{R}_1, \tag{1}$$

and let $Z_{(1)} \leq \ldots \leq Z_{(R_1)}$ be the sorted Z-values. Then the ith largest FDR-replicability adjusted p-value is

$$p_{(i)}^{REPadj} = \min_{j \geq i} \frac{Z_{(j)}}{j}. \tag{2}$$

The FDR-Replicability Procedure with parameters $(q_1, q) = (cq, q)$ is equivalent to rejecting all no replicability hypotheses with FDR-replicability adjusted p-values below q.

The example

Table: The FDR-replicability adjusted p-values for $c = q_1/q$

Locus	SNP	Gene	$c = 0.2$	$c = 0.5$	$c = 0.8$
2q24	rs6741949	DPP4	0.8750	1.000	1.0000
9q33	rs7852872	ASTN2	0.3125	0.5000	1.0000
12q14	rs17178006	MSRB3	0.0688	0.0275	0.03438
	rs6581612	WIF1	0.1375	0.0550	0.03438
12q24	rs7294919	HRK	0.2000	0.0800	0.05000

- At an FDR level of 0.05, the number of replicated discoveries increases with c from 0 to 1 to 3.
- c has to be fixed in advance, and $c = 0.8$ is the most reasonable of the three choices for GWAS, since the power to reach genome-wide significance in the primary study (i.e. p-values of order $1/m$) is typically much smaller than the power to reach significance in the follow up study (i.e. p-values of order $1/R_1$).

The example

Gene	Primary	Follow-up	Replicability adjusted p-value
DPP4	5.2×10^{-8}	0.7	1.0000
ASTN2	1.0×10^{-7}	0.2	1.0000
MSRB3	5.5×10^{-9}	0.002	0.03438
WIF1	2.2×10^{-8}	0.0007	0.03438
HRK	4.8×10^{-8}	5.8×10^{-5}	0.05000

- Typical meta-analysis is not replicability analysis.
- Reporting the replicability-adjusted p-values for the most interesting SNPs complements the table of discoveries, and gives credibility to replicability claims.
- Software available at `http://www.math.tau.ac.il/~ruheller/App.html`

Summary

1. Adjusting for multiplicity is necessary for control of false positives.
2. The FDR criterion is often used in modern application, where many hypotheses are simultaneously tested and few false positives may be tolerated as long as they are a small fraction of the discoveries.
3. The BH procedure is robust to many types of dependencies.
4. Estimating the proportion of null hypotheses may increase the power slightly, at the expense of possibly being anti-conservative for dependent p-values.
5. For discrete tests, the BH procedure should be applied on *midP* values.
6. A two dimensional variant of the BH procedure can be used to establish replicability of discoveries, when a follow-up study examines few promising hypotheses from the primary study.


Benjamini, Y. and Heller, R. (2008).
Screening for partial conjunction hypotheses.
Biometrics, 64:1215–1222.


Benjamini, Y. and Hochberg, Y. (1995).
Controlling the false discovery rate - a practical and powerful approach to multiple testing.
J. Roy. Stat. Soc. B Met., 57 (1):289–300.


Benjamini, Y., Krieger, M., and Yekutieli, D. (2006).
Adaptive linear step-up false discovery rate controlling procedures.
Biometrika, 93 (3):491–507.


Benjamini, Y. and Yekutieli, D. (2001).
The control of the false discovery rate in multiple testing under dependency.
The Annals of Statistics, 29 (4):1165–1188.


Bis, J. et al. (2012).
Common variants at 12q14 and 12q24 are associated with hippocampal volume.

Nature genetics, page doi:10.1038/ng.2237.


Bogomolov, M. and Heller, R. (2012).
Discovering findings that replicate from a primary study of high dimension to a follow-up study.
arXiv:1207.0187v1.


Reiner, A. (2007).
Fdr control by the bh procedure for two-sided correlated tests with implications to gene expression data analysis.
Biometrical Journal, 49(1):107–126.


Storey, J., Taylor, J., and Siegmund, D. (2004).
Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach.
Journal of the Royal Statistical Society, Series B, 66:187–205.


Storey, J. and Tibshirani, R. (2003).
Statistical significance for genomewide studies.
Proceedings of the National Academy of Sciences, 100 (16):9440–9445.


Yekutieli, D. (2008a).
Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling.
Test, 17 (3):458–460.


Yekutieli, D. (2008b).
False discovery rate control for non-positively regression dependent test statistics.
Journal of Statistical Planning and Inference, 138 (2):405–415.