Resampling-Based Empirical Bayes Multiple Testing Procedure for Controlling the False Discovery Rate with Applications to Genomics

H.N. Gilbert S. Dudoit M.J. van der Laan[†]

[†]Division of Biostatistics
University of California, Berkeley

5th International Conference on
Multiple Comparison Procedures
Vienna, July 2007

Outline

The False Discovery Rate (FDR)

The false discovery rate (FDR) is the expected value of the proportion of false positives (V_n) among the rejections made (R_n), with the convention that $V_n/R_n = 0$ if $R_n = 0$.

$$FDR \equiv E\left[\frac{V_n}{R_n}\right] = E\left[\frac{V_n}{R_n}\middle| R_n > 0\right] \Pr(R_n > 0). \quad (1)$$

Most common procedures for controlling the FDR:

1. Focus on the marginal distribution of test statistics.
2. Rely on assumptions about the distribution of test statistics.

OUR GOAL Present a powerful procedure which accurately controls Type I error by taking into account the *joint distribution of test statistics* while placing as *as few assumptions as possible* on the data generating or test statistics distributions.

FDR-Controlling Procedures

Adjusted p-value representations

Let $P_{0n}(O_n(1)) \leq \ldots \leq P_{0n}(O_n(M))$ denote the marginal ordered unadjusted p-values corresponding to M hypothesis tests. The adjusted p-value for the mth ordered test statistic is given by:

- Benjamini-Hochberg (1995), marginal, step-up procedure:

$$\widetilde{P}_{0n}(O_n(m)) = \min_{h=m,\ldots,M} \left\{ \min \left\{ \frac{M}{h} P_{0n}(O_n(h)), 1 \right\} \right\}. \quad (2)$$

- Benjamini-Yekutieli (2001), marginal, step-up, "distribution-free" procedure:

$$\widetilde{P}_{0n}(O_n(m)) = \min_{h=m,\ldots,M} \left\{ \min \left\{ C(M) \frac{M}{h} P_{0n}(O_n(h)), 1 \right\} \right\}, \quad (3)$$

where $C(M) = \sum_{m=1}^{M} \frac{1}{m} \approx \log(M)$.

Other FDR-Controlling Procedures

- Benjamini-Hochberg (2000), adaptive procedure
- Yekutieli-Benjamini (1999), *p*-value resampling-based procedure
- Abramovich et al. (2000), for unknown sparsity in regression
- Sarkar (2002), finite-sample stepwise procedure results
- Bayesian pFDR "fixed rejection region" procedure, i.e., "*q*-value" (cf. Efron, Tibshirani, Storey 2001-2004)

q-values

We assume that the test statistics share a *common marginal non-parametric mixture distribution*, i.e.,

$$T_n(m) \sim f = \pi_0 f_0 + (1 - \pi_0) f_1, \quad m = 1, ..., M, \tag{4}$$

- $\pi_0 = \Pr(H_0(m) = 1)$ is the prior probability of belonging to the set of true nulls, $\mathcal{H}_0$,
- f_0 represents the *marginal null density*, $T_n(m)|\{H_0(m) = 1\} \sim f_0$,
- f_1 represents the *marginal alternative density*, $T_n(m)|\{H_0(m) = 0\} \sim f_1$.

"Local *q*-values": *Posterior probabilities of belonging to the set of true null hypotheses given the value of a test statistic:*

$$\pi_0(t) \equiv \Pr(H_0(m) = 1 | T_n(m) = t) = \frac{\pi_0 f_0(t)}{f(t)}, \quad m = 1, \ldots, M. \tag{5}$$

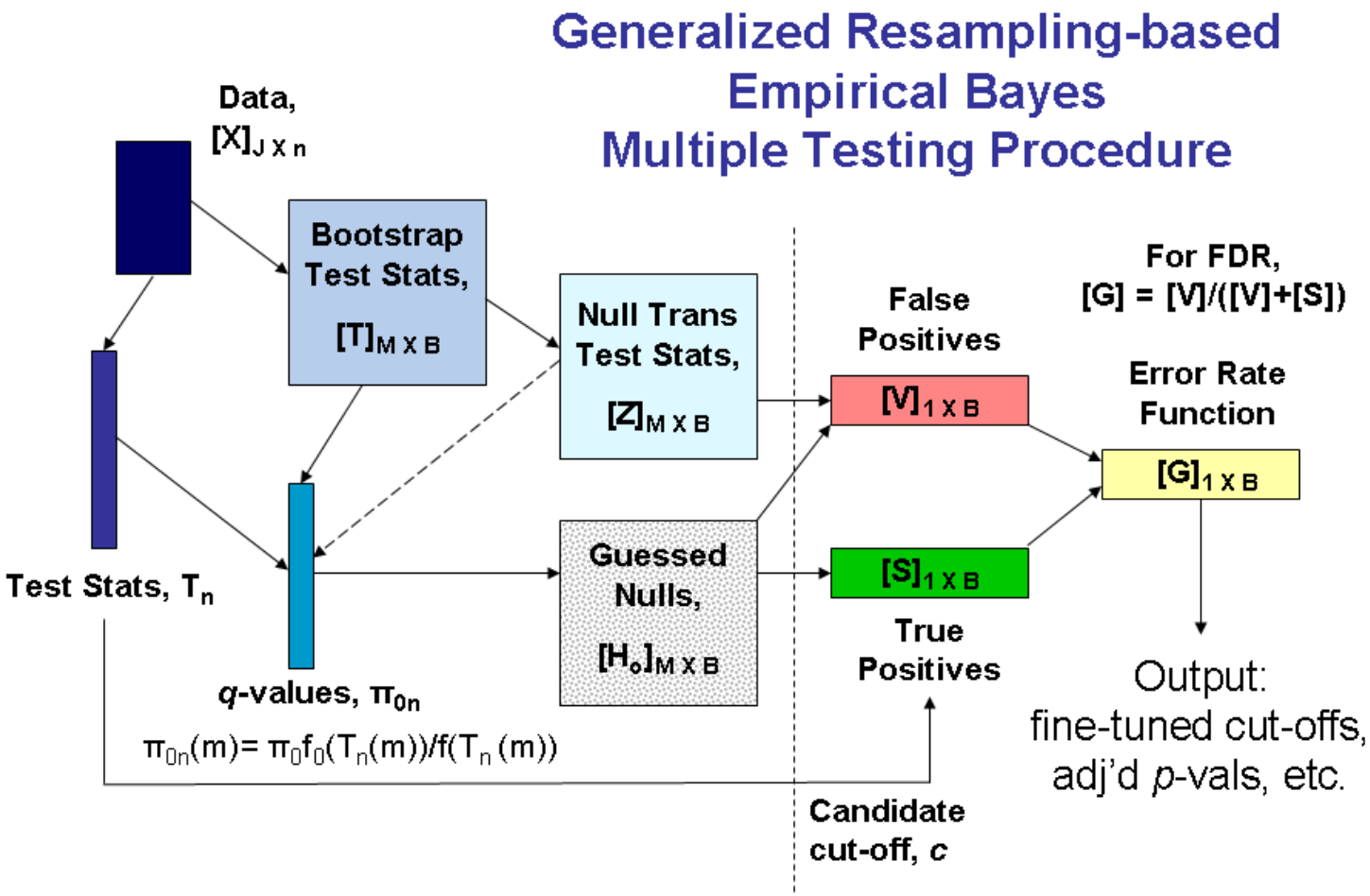
q-values

- *q*-values are "Bayesian *p*-values".
- One typically estimates the prior probability π_0 of a hypothesis being true.
- Storey and colleagues propose at least two methods for estimating this prior from a common marginal mixture distribution of test statistics (see previous).

The *q*-value for the *m*th ordered *p*-value is given by:

$$\widetilde{P}_{0n}(O_n(m)) = \min\left\{\frac{\hat{\pi}_0 M P_{0n}(O_n(m))}{m}, \widetilde{P}_{0n}(O_n(m+1))\right\}, \quad m = 1, \ldots, M \tag{6}$$

The Procedure



Resampling-Based Empirical Bayes MTPs - PROs

- Seek to gain power by using random guessed sets of true null hypotheses vs. conservatively setting $\mathcal{H}_0 = \{1, \ldots, M\}$. We do not assume the complete null hypothesis.
- Based on a test statistics joint null distribution.*
- Applicable to general data generating distributions with arbitrary dependence structures among variables.
  - No assumptions about independence, pos. reg'n. dependence, ergodic dependence, dependence in finite blocks, etc.
  - Do not rely on subset pivotality condition of Westfall and Young.
- Do not require estimation of a prior probability (can use most conservative prior of 1).*
- Work for arbitrary tail probability and expected value error rates. The FDR is only one error rate that can be controlled using this methodology (e.g., gFWER-k, TPPFP-q, etc.).
- Intuitive structure lends itself to transparent software implementation.

Simulation Design

The BH, QV and EB procedures were compared in a simulation study testing two-sided null hypotheses for difference of means. Test statistics (effect sizes) were sampled directly (MVN, $B = 10,000$).

- *Number of null hypotheses*, $M \in \{24, 400, 2000\}$.
- *Proportion of true null hypotheses*, $h_o/M \in \{0.50, 0.75, 0.95, 1\}$.
- *Alternative test statistic shift parameters*, $d_n(m) \in \{2, 3\}$.
- *Correlation matrix*, σ^*:
  - *Independent*, $\sigma^* = I_M$, the $M \times M$ identity matrix.
  - *Constant*, σ^* same as above, with $\sigma^*_{ij} = 0.5, i \neq j$.
  - *Empirical*, σ^* a random $M \times M$ submatrix of the genes $\times$ genes correlation matrix for a publicly available microarray data set (Golub 1999).
- *Estimated test statistics null distributions, Q_{0n} and $\breve{Q}_{0n}$*: Both the null shift and scaled-transformed and null quantile-transformed test statistics null distributions were implemented (including for marginal procedures).
- *"Best Case" Controls*. Allow BH and EB to know the proportion of true nulls, h_0/M.

Numerical Simulation Results

h_0/M	BH-TIE	BH-P	QV-TIE	QV-P	EB-TIE	EB-P
Independent						
0.95	0.080	0.042	0.086	0.042	0.055	0.034
0.75	0.039	0.130	0.048	0.162	0.043	0.146
0.50	0.025	0.224	0.045	0.342	0.036	0.296
Constant						
0.95	0.027	0.057	0.052	0.084	0.025	0.088
0.75	0.040	0.157	0.070	0.188	0.039	0.219
0.50	0.045	0.262	0.070	0.347	0.048	0.378
Empirical						
0.95	0.038	0.062	0.049	0.071	0.038	0.076
0.75	0.036	0.139	0.050	0.173	0.042	0.175
0.50	0.024	0.231	0.046	0.339	0.036	0.319

Table 1: *Simulation Study – Average TIE and power of the BH, QV and EB FDR-controlling procedures over* $A = 250$ *data sets,* $M = 2000$ *hypotheses, mean shift* $d_m = 2$*, controlling at nominal level* $\alpha = 0.05$. TIE = Type I error, P = Power. Results are based on the null shifted and scaled test statistics null distribution.

Graphical Simulation Results - Type I Error

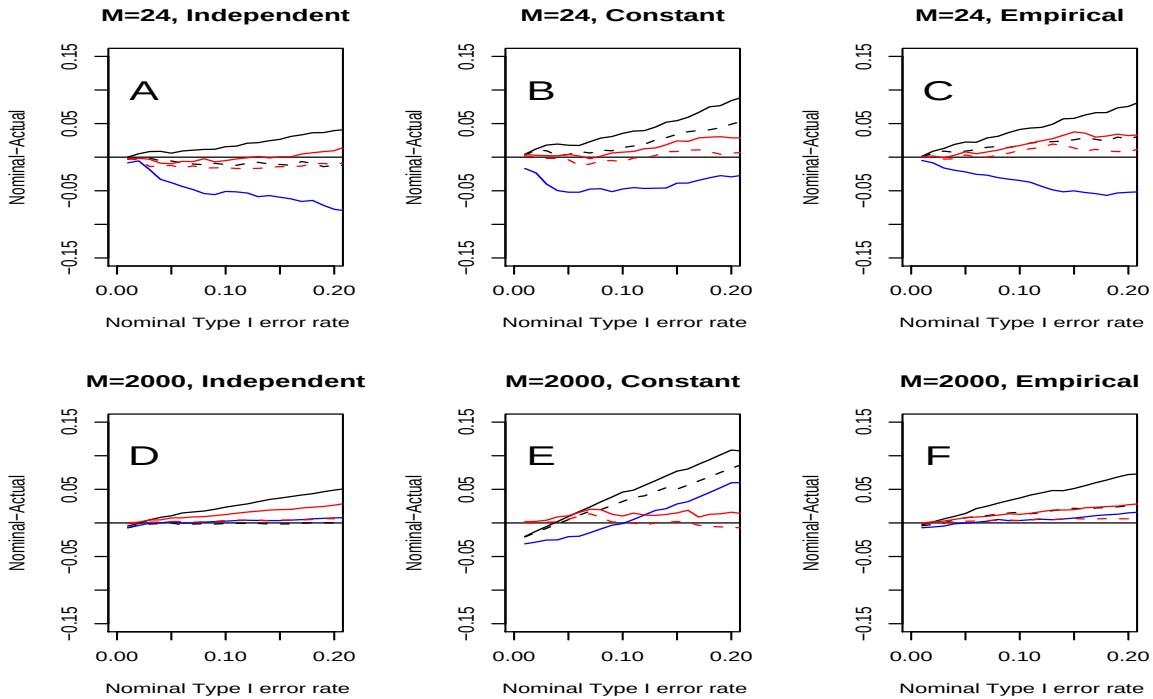


Figure 1: *Simulation Study – Type I error control comparison:* $M = 24$ *and* $M = 2000$ *hypotheses, proportion of true nulls* $h_0/M = 0.75$*, mean shift* $d_m = 2$*.* Performance of BH, QV and EB FDR-controlling procedures under the null shifted and scaled test statistics null distribution. Lines above the horizontal represent conservative behavior.

Graphical Simulation Results - Power

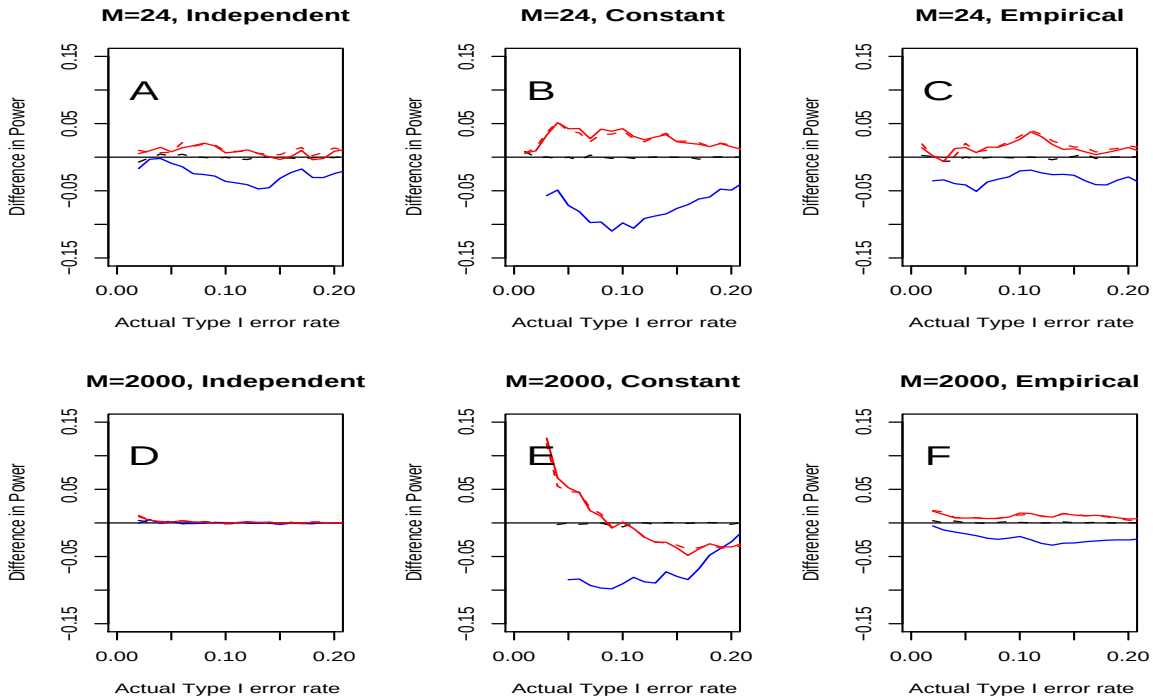


Figure 2: *Simulation Study – Power comparison: $M = 24$ and $M = 2000$ hypotheses, proportion of true nulls $h_0/M = 0.75$, mean shift $d_m = 2$.* Performance of the QV and EB FDR-controlling procedures relative to BH under the null shifted and scaled test statistics null distribution. Lines above the horizontal represent more powerful procedures when benchmarked against actual Type I error rate.

- BH is robust to a variety of conditions, but becomes conservative as h_0/M decreases (as expected).
- QV performs quickly and competitively when assumptions hold, fails under high correlation and small M.
- EB performs consistently well, particularly for small M, high correlation, high h_0/M.
- EB results in fewer "0/0" instances (more rejections).
- EB performs slightly better in terms of the spread of the FDR estimate across simulated data sets.
- EB controls TIE under the complete null hypothesis at the same or similar rates as vanilla (non-EB) FWER and FDR-controlling procedures (Bonferroni, ssmaxT, BH, QV, etc., as well as EB-ssmaxT). Higher M might require more bootstrap replicates, B.

Genetic Cross Experiment in *S. cerevisiae*

- Data are from a genetic cross experiment between a lab strain and a wild isolate of *Saccharomyces cerevisiae* (Brem et al., 2005), for n=112 segregants (progeny).
- Phenotypes of interest are gene expression levels themselves. 6,215 cDNA microarray gene expression measures (ψ).
- Genotypes are biallelic (0/1) SNPs. 2,957 positions covering ca. 95% of the genome.
- "Two-dimensional" multiple testing problem. In practice, one dimension is often reduced, pooled or completely ignored.
- Potential loss of information for complex genetic traits!
- Ideally, we would identify significant sets of genotypes associated with sets of phenotypes.

Chromosome	ATP7	ATP15	COX4	COX6
IIa	QV, EB	EB		
IIb			EB	EB(2)
III		EB(2)		
IV	EB	BH, QV(3), EB(8)		
V		EB		EB
VIIa		EB		
VIIb	EB			
VIII		BH, QV, EB	QV, EB	QV, EB
XVI	QV, EB(2)	QV, EB(2)	QV, EB	QV, EB(2)

Table 2: *Comparative linkage results for individual genes of interest over 1,181 informative SNPs using the BH, QV and EB FDR-controlling procedures.* Adjp < 0.25 for two-sample t-tests using the $N(0,1)$ null quantile-transformed test statistics null distribution. Numbers in parentheses () indicate the number of markers called significant on a given chromosome (if greater than one).

For each gene displayed, EB finds more markers in regions within a chromosome as well as more markers on different chromosomes than the other two methods.

Summary and Future Work

- EB MTPs provide a general, flexible methodology for controlling a large class of Type I error rates.
- EB-FDR procedure exhibits a positive trade-off between TIE control and power across a variety of conditions.
- Promise in high correlation setting should be of practical interest to researchers in several fields.
- Attempting to define parameters to account for "two-dimensional" genotype/phenotype testing problems, graphical structure, etc. *Data adaptive approaches for identifying sets of associated phenotypes and genotypes (transcripts and SNPs).*
- Structure allows for transparent software development - COMING SOON!

Acknowledgements

- Alan Hubbard, UC Berkeley Division of Biostatistics
- Merrill Birkner, Genentech, Inc.
- Katie Pollard, UC Davis Genome Center & Dept. of Statistics
- Rachel Brem, UC Berkeley Dept. of Mol. & Cell Biology
- Bonnie Hutchings, UC Berkeley Division of Biostatistics

THANK YOU