*University of California, Berkeley*
U.C. Berkeley Division of Biostatistics Working Paper Series

*Year* 2007 *Paper* 228

# Resampling-Based Empirical Bayes Multiple Testing Procedures for Controlling Generalized Tail Probability and Expected Value Error Rates:

Sandrine Dudoit* Houston N. Gilbert†
Mark J. van der Laan‡

*Division of Biostatistics and Department of Statistics, University of California, Berkeley, sandrine@stat.berkeley.edu

†Division of Biostatistics, University of California, Berkeley, houstong@berkeley.edu

‡Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu



http://biostats.bepress.com/ucbbiostat/paper228

# Resampling-Based Empirical Bayes Multiple Testing Procedures for Controlling Generalized Tail Probability and Expected Value Error Rates:

Sandrine Dudoit, Houston N. Gilbert, and Mark J. van der Laan

**Abstract**

This article proposes resampling-based empirical Bayes multiple testing procedures for controlling a broad class of Type I error rates, defined as generalized tail probability (gTP) error rates, $gTP(q,g) = \Pr(g(V_n,S_n) > q)$, and generalized expected value (gEV) error rates, $gEV(g) = [g(V_n,S_n)]$, for arbitrary functions $g(V_n,S_n)$ of the numbers of false positives $V_n$ and true positives $S_n$. Of particular interest are error rates based on the proportion $g(V_n,S_n) = V_n/(V_n + S_n)$ of Type I errors among the rejected hypotheses, such as the false discovery rate (FDR), $FDR = [V_n/(V_n + S_n)]$. The proposed procedures offer several advantages over existing methods. They provide Type I error control for general data generating distributions, with arbitrary dependence structures among variables. Gains in power are achieved by deriving rejection regions based on guessed sets of true null hypotheses and null test statistics randomly sampled from joint distributions that account for the dependence structure of the data. The Type I error and power properties of an FDR-controlling version of the resampling-based empirical Bayes approach are investigated and compared to those of widely-used FDR-controlling linear step-up procedures in a simulation study. The Type I error and power trade-off achieved by the empirical Bayes procedures under a variety of testing scenarios allows this approach to be competitive with or outperform the Storey and Tibshirani [2003] linear step-up procedure, as an alternative to the classical Benjamini and Hochberg [1995] procedure.

# 1 Introduction

## 1.1 Motivation and overview

Current statistical inference problems in areas such as astronomy, genomics, and marketing, routinely involve the simultaneous test of thousands, or even millions, of null hypotheses. These hypotheses concern a wide range of parameters, for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables.

Type I error rates based on the *proportion* $V_n/(V_n+S_n)$ of false positives among the rejected hypotheses (e.g., false discovery rate, $FDR = \mathrm{E}[V_n/(V_n+S_n)]$) are especially appealing for large-scale testing problems, compared to traditional error rates based on the *number* $V_n$ of false positives (e.g., family-wise error rate, $FWER = \Pr(V_n > 0)$), as they do not increase exponentially with the number $M$ of tested hypotheses.

However, only a handful of multiple testing procedures (MTP) are currently available for controlling such Type I error rates. Furthermore, existing methods suffer from a variety of limitations. Firstly, marginal procedures can lack power by failing to account for the dependence structure of the test statistics [Benjamini and Hochberg, 1995, Lehmann and Romano, 2005]. Secondly, even for some of the marginal procedures, Type I error control relies on restrictive and hard to verify assumptions concerning the joint distribution of the test statistics, e.g., independence, dependence in finite blocks, ergodic dependence, positive regression dependence, Simes' Inequality [Benjamini and Hochberg, 1995, 2000, Benjamini and Yekutieli, 2001, Benjamini et al., 2006, Genovese and Wasserman, 2004a,b, Lehmann and Romano, 2005, Storey, 2002, Storey and Tibshirani, 2003, Storey et al., 2004]. Thirdly, some procedures err conservatively by counting rejected hypotheses as Type I errors or estimating the proportion $h_0/M$ of true null hypotheses by its upper bound of one [Benjamini and Hochberg, 1995, Dudoit and van der Laan, 2007, Dudoit et al., 2004a, van der Laan et al., 2004b].

Motivated by these observations, van der Laan et al. [2005] propose a resampling-based empirical Bayes procedure for controlling the tail probability for the proportion of false positives (TPPFP) among the rejected hypotheses, $TPPFP(q) = \Pr(V_n/(V_n+S_n) > q)$. The approach is extended in Dudoit and van der Laan [2007, Chapter 7] to control generalized tail probability (gTP) error rates, $gTP(q,g) = \Pr(g(V_n,S_n) > q)$, for arbitrary functions $g(V_n,S_n)$ of the numbers of false positives $V_n$ and true positives $S_n$. Dudoit and van der Laan [2007, Section 7.8] further remark that empirical Bayes procedures may be used to control generalized expected value (gEV) error rates, $gEV(g) = \mathrm{E}[g(V_n,S_n)]$, such as the false discovery rate, $FDR = \mathrm{E}[V_n/(V_n+S_n)]$, and other parameters of the distribution of functions $g(V_n,S_n)$.

The two main ingredients in a resampling-based empirical Bayes procedure are the following distributions.

- A null distribution $Q_0$ (or estimator thereof, $Q_{0n}$) for $M$-vectors of *null test statistics* $T_{0n}$.
- A distribution $Q_0^{\mathcal{H}}$ (or estimator thereof, $Q_{0n}^{\mathcal{H}}$) for *random guessed sets of true null hypotheses* $\mathcal{H}_{0n}$.

By randomly sampling null test statistics $T_{0n}$ and guessed sets of true null hypotheses $\mathcal{H}_{0n}$, one obtains a distribution for a random variable representing the *guessed g-specific function of the numbers of false positives and true positives* (given the empirical distribution $P_n$), for any given rejection region. Rejection regions can then be chosen to control tail probabilities and expected values for this distribution at a user-supplied Type I error level $\alpha$.

Our proposed empirical Bayes procedures seek to gain power by taking into account the joint distribution of the test statistics and by "guessing" the set $\mathcal{H}_0$ of true null hypotheses instead of conservatively setting $\mathcal{H}_0 = \{1,\ldots,M\}$. In addition, unlike most MTPs controlling the proportion of false positives, they provide Type I error control for general data generating distributions, with arbitrary dependence structures among variables.

Note that the empirical Bayes approach outlined above is very general and modular, in the sense that it can be applied to *any* distribution pair $(Q_{0n}, Q_{0n}^{\mathcal{H}})$. In particular, the common marginal non-parametric mixture model of Section 3.3 is only one among many reasonable candidate models for $Q_{0n}^{\mathcal{H}}$ that does not assume independence of the test statistics.

## 1.2 Outline

This article proposes resampling-based empirical Bayes multiple testing procedures for controlling a broad class of Type I error rates, defined as generalized tail probability error rates, $gTP(q,g) = \Pr(g(V_n, S_n) > q)$, and generalized expected value error rates, $gEV(g) = \mathrm{E}[g(V_n, S_n)]$, for arbitrary functions $g(V_n, S_n)$ of the numbers of false positives $V_n$ and true positives $S_n$.

The article is organized as follows. The remainder of this section provides a brief overview of the multiple hypothesis testing framework developed in Dudoit and van der Laan [2007]. Section 2 focusses on the special case of the false discovery rate, $FDR = \mathrm{E}[V_n/(V_n + S_n)]$, and summarizes widely-used FDR-controlling linear step-up procedures [Benjamini and Hochberg, 1995, 2000, Benjamini et al., 2006, Storey, 2002, Storey and Tibshirani, 2003]. Section 3 presents the resampling-based empirical Bayes multiple testing procedures proposed in Dudoit and van der Laan [2007, Chapter 7] and van der Laan et al. [2005] for controlling generalized tail probability and expected value error rates. In the simulation study of Sections 4 and 5, the Type I error and power properties of an FDR-controlling version of the resampling-based empirical Bayes approach are investigated and compared to those of FDR-controlling linear step-up procedures introduced in Section 2. Finally, Section 6 summarizes our findings and outlines ongoing work.

## 1.3 Multiple hypothesis testing framework

This section, based on Dudoit and van der Laan [2007, Chapter 1], introduces a general statistical framework for multiple hypothesis testing and discusses in turn the main ingredients of a multiple testing problem.

### 1.3.1 Null and alternative hypotheses

Consider a *data generating distribution* $P \in \mathcal{M}$, belonging to a *model* $\mathcal{M}$, i.e., a set of possibly non-parametric distributions.

Suppose one has a *learning set* $\mathcal{X}_n \equiv \{X_i : i = 1, \ldots, n\} \overset{IID}{\sim} P$, of $n$ independent and identically distributed (IID) random variables (RV) from $P$. Let $P_n$ denote the *empirical distribution* of the learning set $\mathcal{X}_n$, which places probability $1/n$ on each $X_i$, $i = 1, \ldots, n$.

*Hypothesis testing* is concerned with using observed data to make decisions regarding properties of, i.e., hypotheses for, the unknown distribution that generated these data.

Define $M$ pairs of null and alternative hypotheses in terms of a collection of $M$ *submodels*, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \ldots, M$, for the data generating distribution $P$ [Dudoit and van der Laan, 2007, Section 1.2.4]. Specifically, the $M$ *null hypotheses* and corresponding *alternative hypotheses* are defined, respectively, as

$$H_0(m) \equiv \mathrm{I}(P \in \mathcal{M}(m)) \qquad \text{and} \qquad H_1(m) \equiv \mathrm{I}(P \notin \mathcal{M}(m)). \tag{1}$$

In many testing problems, the submodels concern parameters, i.e., functions $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M) \in \mathbb{R}^M$ of the data generating distribution $P$, and each null hypothesis $H_0(m)$ refers to a single parameter, $\psi(m) = \Psi(P)(m) \in \mathbb{R}$.

The *complete null hypothesis* $H_0^C$ states that the data generating distribution $P$ belongs to the intersection $\cap_{m=1}^M \mathcal{M}(m)$ of the $M$ submodels,

$$H_0^C \equiv \prod_{m=1}^{M} H_0(m) = \prod_{m=1}^{M} \mathrm{I}(P \in \mathcal{M}(m)) = \mathrm{I}\left(P \in \cap_{m=1}^M \mathcal{M}(m)\right). \tag{2}$$

Let

$$\mathcal{H}_0 = \mathcal{H}_0(P) \equiv \{m : H_0(m) = 1\} = \{m : P \in \mathcal{M}(m)\} \tag{3}$$

denote the set of $h_0 \equiv |\mathcal{H}_0|$ *true null hypotheses*, where the longer notation $\mathcal{H}_0(P)$ emphasizes the dependence of this set on the data generating distribution $P$. Likewise, let

$$\mathcal{H}_1 = \mathcal{H}_1(P) \equiv \{m : H_1(m) = 1\} = \{m : P \notin \mathcal{M}(m)\} = \mathcal{H}_0^c(P) \tag{4}$$

be the set of $h_1 \equiv |\mathcal{H}_1| = M - h_0$ *false null hypotheses*.

### 1.3.2 Test statistics

A *testing procedure* is a *data-driven*, i.e., *random*, *rule* for estimating the set of false null hypotheses $\mathcal{H}_1$, i.e., for deciding which null hypotheses should be *rejected*.

The decisions to reject or not the null hypotheses are based on an $M$-vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$, that are functions $T_n(m) = T(m; \mathcal{X}_n) = T(m; P_n)$ of the data $\mathcal{X}_n$, i.e., of the empirical distribution $P_n$ [Dudoit and van der Laan, 2007, Section 1.2.5]. Denote the typically unknown (finite sample) *joint distribution of the test statistics* $T_n$ by $Q_n = Q_n(P)$.

Single-parameter null hypotheses of the form $H_0(m) = \mathrm{I}(\psi(m) \leq \psi_0(m))$ or $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$, $m = 1, \ldots, M$, may be tested based on *t-statistics* (i.e., standardized differences),

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null value}}{\text{Standard error}} = \sqrt{n}\frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}. \tag{5}$$

Here, $\hat{\Psi}(P_n) = \psi_n = (\psi_n(m) : m = 1, \ldots, M)$ denotes an *estimator* of the parameter $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$ and $(\sigma_n(m)/\sqrt{n} : m = 1, \ldots, M)$ denotes the estimated *standard errors* for elements $\psi_n(m)$ of $\psi_n$.

This general representation for the test statistics covers standard one-sample and two-sample $t$-statistics for testing hypotheses concerning mean parameters, but also test statistics for correlation coefficients and regression coefficients in linear and non-linear models. Test statistics for other types of null hypotheses include $F$-statistics, $\chi^2$-statistics, and likelihood ratio statistics.

### 1.3.3 Multiple testing procedures

A *multiple testing procedure* (MTP) provides *rejection regions* $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the corresponding null hypothesis $H_0(m)$ and declare that $P \notin \mathcal{M}(m)$, $m = 1, \ldots, M$ [Dudoit and van der Laan, 2007, Sections 1.2.6 and 1.2.7]. In other words, a MTP produces a random (i.e., data-dependent) set of rejected null hypotheses $\mathcal{R}_n$ that estimates the set of false null hypotheses $\mathcal{H}_1$,

$$\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \equiv \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : H_0(m) \text{ is rejected}\}, \tag{6}$$

where $\mathcal{C}_n(m) = \mathcal{C}(m; T_n, Q_{0n}, \alpha)$, $m = 1, \ldots, M$, denote possibly random test statistic rejection regions.

The long notation $\mathcal{R}(T_n, Q_{0n}, \alpha)$ and $\mathcal{C}(m; T_n, Q_{0n}, \alpha)$ emphasizes that a MTP depends on the following three ingredients.

1. The *data*, $\mathcal{X}_n = \{X_i : i = 1, \ldots, n\}$, through the $M$-vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$ (Section 1.3.2).

2. An (estimated) $M$-variate *test statistics null distribution*, $Q_{0n}$, which replaces the unknown test statistics distribution $Q_n = Q_n(P)$ (Section 1.3.5), for the purpose of deriving rejection regions for the test statistics, confidence regions for parameters of interest, and adjusted $p$-values (Section 1.3.6).

3. The *nominal Type I error level* $\alpha$, i.e., a user-supplied upper bound for a suitably defined Type I error rate (Section 1.3.4).

We focus without loss of generality on one-sided rejection regions of the form $\mathcal{C}_n(m) = (c_n(m), +\infty)$, where $c_n = (c_n(m) : m = 1, \ldots, M) \in \mathbb{R}^M$ is an $M$-vector of *critical values* or *cut-offs*.

### 1.3.4 Type I error rate and power

**Errors in multiple hypothesis testing** In any testing problem, two types of errors can be committed [Dudoit and van der Laan, 2007, Section 1.2.8]. A *Type I error*, or *false positive*, is committed by rejecting a true null hypothesis ($\mathcal{R}_n \cap \mathcal{H}_0$). A *Type II error*, or *false negative*, is committed by failing to reject a false null hypothesis ($\mathcal{R}_n^c \cap \mathcal{H}_1$). The situation can be summarized as in Table 1.

Ideally, one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors. Unfortunately, this is not feasible and one seeks a trade-off between the two types of errors. A standard approach is to specify an acceptable level $\alpha$ for a suitably defined Type I error rate and derive testing procedures (i.e., rejection regions) that aim to minimize a Type II error rate (i.e., maximize power) within the class of tests with Type I error rate at most $\alpha$.

**Type I error rate** When testing multiple hypotheses, there are many possible definitions for the Type I error rate and power of a testing procedure. Accordingly, we define a *Type I error rate* as a parameter $\theta_n = \Theta(F_{V_n,R_n})$ of the joint distribution $F_{V_n,R_n}$ of the numbers of Type I errors $V_n = |\mathcal{R}_n \cap \mathcal{H}_0|$ and rejected hypotheses $R_n = |\mathcal{R}_n|$ [Dudoit and van der Laan, 2007, Section 1.2.9].

Such a representation covers a broad class of Type I error rates, defined as *generalized tail probability* (gTP) error rates,

$$gTP(q,g) \equiv \Pr(g(V_n, S_n) > q), \tag{7}$$

and *generalized expected value* (gEV) error rates,

$$gEV(g) \equiv \mathrm{E}[g(V_n, S_n)], \tag{8}$$

for functions $g(V_n, S_n)$ of the numbers of false positives $V_n$ and true positives $S_n = R_n - V_n$.

Consider functions $g$ that satisfy the following two natural monotonicity assumptions.

**Assumption MgV.** The function $g_s : v \to g(v,s)$ is continuous and strictly increasing for any given $s$.

**Assumption MgS.** The function $g_v : s \to g(v,s)$ is continuous and non-increasing for any given $v$.

Of particular interest are the following two special cases, corresponding, respectively, to $g$-functions for the number and proportion of false positives among the rejected hypotheses. When $g(v,s) = v$, one recovers the *generalized family-wise error rate* (gFWER) and the *per-family error rate* (PFER). When $g(v,s) = v/(v+s)$, with the convention that $v/(v+s) \equiv 0$ if $v+s=0$, one obtains the *tail probability for the proportion of false positives* (TPPFP) and the *false discovery rate* (FDR). Specifically, the FDR is defined as

$$FDR \equiv \mathrm{E}\left[\frac{V_n}{\max\{R_n,1\}}\right] = \mathrm{E}\left[\frac{V_n}{R_n}\middle| V_n > 0\right]\Pr(V_n > 0) = \mathrm{E}\left[\frac{V_n}{R_n}\middle| R_n > 0\right]\Pr(R_n > 0), \tag{9}$$

where $R_n = V_n + S_n$. Under the complete null hypothesis $H_0^C = \mathrm{I}\left(P \in \cap_{m=1}^M \mathcal{M}(m)\right)$, all $R_n$ rejected hypotheses are Type I errors, hence $V_n/R_n = 1$ and $FDR = FWER = \Pr(V_n > 0)$.

Storey and Tibshirani [2003] and related articles [Storey, 2002, Storey et al., 2004] consider a variant of the FDR, termed the *positive false discovery rate* (pFDR),

$$pFDR \equiv \mathrm{E}\left[\frac{V_n}{R_n}\middle| R_n > 0\right]. \tag{10}$$

Note that $FDR = pFDR \times \Pr(R_n > 0)$, so that, in general, $FDR \leq pFDR$. An immediate flaw of the pFDR is that it is equal to one under the complete null hypothesis and therefor cannot be controlled under this testing scenario. By contrast, the FDR reduces to the family-wise error rate, $FWER = \Pr(V_n > 0)$.

The *actual* Type I error rate $\Theta(F_{V_n,R_n})$ of a multiple testing procedure typically differs from its *nominal* Type I error level $\alpha$, i.e., the level at which it claims to control Type I errors. Discrepancies between actual and nominal Type I error rates can be attributed to a number of factors, including the choice of a test statistics null distribution $Q_{0n}$ and the type of rejection regions for a given choice of $Q_{0n}$. A testing procedure is said to be *conservative* if the nominal Type I error level $\alpha$ is greater than the actual Type I error rate, i.e., $\Theta(F_{V_n,R_n}) < \alpha$, and *anti-conservative* if the nominal Type I error level $\alpha$ is less than the actual Type I error rate, i.e., $\Theta(F_{V_n,R_n}) > \alpha$.

**Power** Likewise, we define *power* as a parameter $\vartheta_n = \Theta(F_{U_n,R_n})$ of the joint distribution $F_{U_n,R_n}$ of the numbers of Type II errors $U_n = |\mathcal{R}_n^c \cap \mathcal{H}_1|$ and rejected hypotheses $R_n = |\mathcal{R}_n|$ [Dudoit and van der Laan, 2007, Section 1.2.10].

The *average power*, i.e., the expected value of the proportion of rejected hypotheses among the false null hypotheses, is defined as

$$AvgPwr \equiv \frac{1}{h_1} \mathrm{E}[S_n] = 1 - \frac{1}{h_1} \mathrm{E}[U_n]. \tag{11}$$

### 1.3.5 Test statistics null distribution

One of the main tasks in specifying a multiple testing procedure is to derive rejection regions for the test statistics such that Type I errors are probabilistically controlled at a user-supplied level. However, one is immediately faced with the problem that the distribution of the test statistics is usually unknown.

In practice, the test statistics distribution $Q_n = Q_n(P)$ is replaced by a *null distribution* $Q_0$ (or estimator thereof, $Q_{0n}$) in order to derive rejection regions and resulting adjusted $p$-values. The choice of a proper null distribution is crucial in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the assumed null distribution does indeed provide the desired control under the true distribution.

Dudoit and van der Laan [2007, Chapter 2] provide a general characterization for a proper test statistics null distribution, which leads to the explicit construction of two main types of test statistics null distributions.

The first original proposal of Dudoit et al. [2004b], van der Laan et al. [2004a], and Pollard and van der Laan [2004], defines the null distribution as the asymptotic distribution of a vector of *null shift and scale-transformed test statistics*, based on user-supplied upper bounds for the means and variances of the test statistics for the true null hypotheses [Dudoit and van der Laan, 2007, Section 2.3].

The second and most recent proposal of van der Laan and Hubbard [2006] defines the null distribution as the asymptotic distribution of a vector of *null quantile-transformed test statistics*, based on user-supplied test statistic marginal null distributions [Dudoit and van der Laan, 2007, Section 2.4].

For a broad class of testing problems, such as the test of single-parameter null hypotheses using $t$-statistics, a proper null distribution is the $M$-variate Gaussian distribution $\mathrm{N}(0, \sigma^*)$, with mean vector zero and covariance matrix $\sigma^* = \Sigma^*(P)$ equal to the correlation matrix of the vector influence curve for the estimator $\psi_n$ of the parameter of interest $\psi$ [Dudoit and van der Laan, 2007, Section 2.6].

Resampling procedures (e.g., non-parametric or model-based bootstrap) are available to conveniently obtain consistent estimators of the null distribution and of the corresponding test statistic cut-offs, parameter confidence regions, and adjusted $p$-values [Dudoit and van der Laan, 2007, Procedures 2.3 and 2.4].

### 1.3.6 Adjusted $p$-values

*Adjusted* $p$-values, for the test of multiple hypotheses, are defined as straightforward extensions of *unadjusted* $p$-values, for the test of single hypotheses [Dudoit and van der Laan, 2007, Section 1.2.12]. Consider any multiple testing procedure $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_{0n}, \alpha)$, with rejection regions $\mathcal{C}_n(m;\alpha) = \mathcal{C}(m; T_n, Q_{0n}, \alpha)$. One can define an $M$-vector of *adjusted p-values*, $\widetilde{P}_{0n} = (\widetilde{P}_{0n}(m) : m = 1, \ldots, M) = \widetilde{P}(T_n, Q_{0n}) = \widetilde{P}(\mathcal{R}(T_n, Q_{0n}, \alpha) : \alpha \in [0,1])$, as

$$\begin{aligned} \widetilde{P}_{0n}(m) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_0(m) \text{ at nominal MTP level } \alpha\} \\ &= \inf\{\alpha \in [0,1] : m \in \mathcal{R}_n(\alpha)\} \\ &= \inf\{\alpha \in [0,1] : T_n(m) \in \mathcal{C}_n(m;\alpha)\}, \qquad m = 1, \ldots, M. \end{aligned} \tag{12}$$

That is, the adjusted $p$-value $\widetilde{P}_{0n}(m)$, for null hypothesis $H_0(m)$, is the smallest nominal Type I error level (e.g., gFWER, TPPFP, or FDR) of the multiple hypothesis testing procedure at which one would reject $H_0(m)$, given $T_n$.

As in single hypothesis tests, the smaller the adjusted $p$-value $\widetilde{P}_{0n}(m)$, the stronger the evidence against the corresponding null hypothesis $H_0(m)$. Thus, one rejects $H_0(m)$ for small adjusted $p$-values $\widetilde{P}_{0n}(m)$.

## 2 FDR-controlling linear step-up multiple testing procedures

The following commonly-used FDR-controlling linear step-up procedures, such as Benjamini and Hochberg's [1995] classical procedure and Storey and Tibshirani's [2003] so-called $q$-value procedure, require as their primary input an $M$-vector $(P_{0n}(m) : m = 1, \ldots, M)$ of unadjusted $p$-values, computed under a test statistics null distribution $Q_0$ (or estimator thereof, $Q_{0n}$). The procedures are listed in Table 2.

### 2.1 Benjamini and Hochberg [1995] classical linear step-up procedure

In their seminal article, Benjamini and Hochberg [1995] propose the following FDR-controlling procedure.

**Procedure 1 [FDR-controlling linear step-up Benjamini and Hochberg [1995] procedure]**
*Given an $M$-vector $(P_{0n}(m) : m = 1, \ldots, M)$ of unadjusted $p$-values, let $O_n(m)$ denote the indices for the ordered unadjusted $p$-values, so that $P_{0n}(O_n(1)) \leq \cdots \leq P_{0n}(O_n(M))$. For controlling the FDR at nominal level $\alpha$, the linear step-up procedure of Benjamini and Hochberg [1995] yields the following set of rejected null hypotheses,*

$$\mathcal{R}_n(\alpha) = \left\{ O_n(m) : \exists\ h \geq m \text{ such that } P_{0n}(O_n(h)) \leq \frac{h}{M}\,\alpha \right\}. \tag{13}$$

*That is, the $m$th most significant null hypothesis $H_0(O_n(m))$, with the $m$th smallest unadjusted $p$-value $P_{0n}(O_n(m))$, is rejected if and only if it or at least one of the less significant null hypotheses $H_0(O_n(h))$, $h \geq m + 1$, has an unadjusted $p$-value less than or equal to the corresponding cut-off $\alpha h/M$. Adjusted $p$-values can be derived as*

$$\widetilde{P}_{0n}(O_n(m)) = \min_{h=m,\ldots,M} \left\{ \min \left\{ \frac{M}{h}\, P_{0n}(O_n(h)), 1 \right\} \right\}, \qquad m = 1, \ldots, M. \tag{14}$$

Following the characterization of MTPs in Dudoit and van der Laan [2007, Section 1.2.7], the Benjamini and Hochberg [1995] procedure is a marginal step-up common-quantile procedure: it is *marginal*, in the sense that it is solely based on the marginal distributions of the test statistics and does not account for their dependence structure; it is a *step-up* procedure, in the sense that as soon as one null hypothesis is rejected, all remaining more significant hypotheses are rejected; it is a *common-quantile* procedure, in the sense that it is based on a $p$-value transformation of the test statistics.

Note, however, that although Procedure 1 is a marginal procedure, proofs of FDR control rely on assumptions concerning the joint distribution of the test statistics. Benjamini and Hochberg [1995] prove that Procedure 1 controls the FDR for independent test statistics. The subsequent article of Benjamini and Yekutieli [2001] establishes FDR control for test statistics with more general dependence structures, such as positive regression dependence.

### 2.2 Adaptive linear step-up procedures

Classical linear step-up Benjamini and Hochberg [1995] Procedure 1 can be conservative, as Type I error control results show that it satisfies $\mathrm{E}[V_n/R_n] \leq \alpha h_0/M \leq \alpha$, under certain assumptions on the joint distribution of the test statistics (e.g., independence, positive regression dependence). To remedy this conservativeness, Benjamini and colleagues have developed various adaptive procedures, involving the estimation of the number $h_0$ of true null hypotheses. Benjamini et al. [2006, Section 3] provide a nice review of such methods.

### 2.2.1 Generic adaptive linear step-up procedure

**Procedure 2 [FDR-controlling generic adaptive linear step-up Benjamini et al. [2006, Definition 2] procedure]**
*Given an estimator $h_{0n}$ of the number of true null hypotheses $h_0$, the generic adaptive linear step-up procedure of Benjamini et al. [2006, Definition 2] replaces the nominal Type I error level $\alpha$ in Benjamini and Hochberg [1995] Procedure 1 by the less conservative level of $\alpha M/h_{0n} \geq \alpha$ .*

*Provided $h_{0n}$ does not depend on the nominal Type I error level $\alpha$, the adjusted p-values of an adaptive linear step-up procedure are simply the adjusted p-values of Procedure 1 scaled by $M/h_{0n}$,*

$$\widetilde{P}_{0n}(O_n(m)) = \frac{h_{0n}}{M} \min_{h=m,\ldots,M} \left\{ \min \left\{ \frac{M}{h} P_{0n}(O_n(h)), 1 \right\} \right\}, \qquad m = 1, \ldots, M. \tag{15}$$

Since $h_{0n}/M \leq 1$, adaptive procedures lead to a larger number of rejected hypotheses than the standard Benjamini and Hochberg [1995] procedure (with $h_{0n}/M = 1$) applied with the same nominal FDR level $\alpha$.

### 2.2.2 Benjamini and Hochberg [2000] adaptive linear step-up procedure

The adaptive linear step-up procedure of Benjamini and Hochberg [2000], summarized in Benjamini et al. [2006, Definition 3], derives the following estimator of the number of true null hypotheses based on graphical considerations.

$$h_{0n}^{ABH} = \lceil \min \{ h_{0n}(m_n), M \} \rceil, \tag{16}$$

where

$$\begin{aligned} h_{0n}(m) &= \frac{M + 1 - m}{1 - P_{0n}(O_n(m))}, \\ m_n &= \min \{ m = 2, \ldots, M : h_{0n}(m) > h_{0n}(m-1) \}, \end{aligned}$$

and the ceiling $\lceil x \rceil$ denotes the least integer greater than or equal to $x$, i.e., $\lceil x \rceil \in \mathbb{Z}$ and $\lceil x \rceil - 1 < x \leq \lceil x \rceil$.

Benjamini and Hochberg [2000] prove that this adaptive procedure controls the FDR for independent test statistics.

### 2.2.3 Storey and Tibshirani [2003] adaptive linear step-up procedure

Benjamini et al. [2006, Definition 5] show that the so-called $q$-value procedure of Storey [2002] and Storey and Tibshirani [2003], further discussed in Section 2.3, below, is a particular type of adaptive linear step-up procedure, with estimated number of true null hypotheses defined as

$$h_{0n}^{ST}(\lambda) = \frac{|\{ m : P_{0n}(m) > \lambda \}|}{1 - \lambda}, \tag{17}$$

in terms of a to-be-determined tuning parameter $\lambda \in [0, 1]$, as in Procedure 3.

### 2.2.4 Benjamini et al. [2006] adaptive two-stage linear step-up procedure

Benjamini et al. [2006, Section 4, Definition 6] propose an adaptive two-stage linear step-up procedure (TST), whereby the estimator of the number of true null hypotheses $h_0$ is obtained from a one-stage application of standard linear step-up Benjamini and Hochberg [1995] Procedure 1. Specifically, the estimator of $h_0$

is defined in terms of the number $R_n^1(\alpha/(1+\alpha))$ of rejected hypotheses from a one-stage application of Procedure 1 with nominal FDR level $\alpha/(1+\alpha)$,

$$h_{0n}^{TST}(\alpha) = (1+\alpha)(M - R_n^1(\alpha/(1+\alpha))). \tag{18}$$

Benjamini et al. [2006, Section 5] prove that the TST procedure controls the FDR for independent test statistics.

A multi-stage extension of Procedure 1 is also proposed [Benjamini et al., 2006, Definition 7].

Note that the estimated number of true null hypotheses $h_{0n}^{TST}(\alpha)$ depends on the nominal Type I error level $\alpha$. As a result, one cannot obtain closed form expressions (e.g., as in Equation (15)) for the adjusted $p$-values of the two-stage procedure .

A practical question of interest is the nature and strength of the dependence of the estimated number of true null hypotheses $h_{0n}^{TST}(\alpha)$ on the nominal Type I error level $\alpha$. In general, $h_{0n}^{TST}(\alpha)$ is not monotonic in $\alpha$, as $M - R_n^1(\alpha/(1+\alpha))$ decreases with $\alpha$, while $1+\alpha$ increases with $\alpha$. Extreme cases are $h_{0n}^{TST}(0) = M$ and $h_{0n}^{TST}(1) = 2(M - R_n^1(1/2))$.

## 2.3 Storey and Tibshirani [2003] adaptive linear step-up procedure

As argued in Benjamini et al. [2006, Definition 5] and below, the $q$-value method proposed in Storey and Tibshirani [2003] and related articles [Storey, 2002, Storey et al., 2004] can be viewed simply as a special case of adaptive linear step-up Procedure 2, with a particular type of estimator for the number of true null hypotheses $h_0$.

The procedure requires as input an $M$-vector $(P_{0n}(m) : m = 1, \ldots, M)$ of unadjusted $p$-values and returns so-called $q$-values, which correspond in fact to adjusted $p$-values for the false discovery rate and a variant thereof, the positive false discovery rate, defined in Equation (10).

### 2.3.1 Algorithm

The $q$-value algorithm, provided in Storey and Tibshirani [2003, Remark B] and implemented in the Bioconductor R package qvalue, is summarized below using the notation introduced in Section 1.3.

**Procedure 3 [FDR-controlling adaptive linear step-up Storey and Tibshirani [2003] procedure]**
*For controlling the FDR at nominal level $\alpha$, the adaptive linear step-up procedure of Storey and Tibshirani [2003] proceeds as follows.*

1. *Given an $M$-vector $(P_{0n}(m) : m = 1, \ldots, M)$ of unadjusted p-values, let $O_n(m)$ denote the indices for the ordered unadjusted p-values, so that $P_{0n}(O_n(1)) \leq \cdots \leq P_{0n}(O_n(M))$.*

2. *For a range of values for the tuning parameter $\lambda$, e.g., $\lambda \in \{0, 0.01, 0.02, \ldots, 0.95\}$, compute the following candidate estimators $\pi_{0n}(\lambda)$ of the proportion of true null hypotheses $\pi_0 = h_0/M$,*

$$\pi_{0n}(\lambda) = \frac{|\{m : P_{0n}(m) > \lambda\}|}{M(1-\lambda)}. \tag{19}$$

3. *Fit a natural cubic spline to $(\lambda, \pi_{0n}(\lambda))$ and let the estimator $\pi_{0n}$ be the fitted value at $\lambda = 1$.*

4. *Compute q-values $(\widetilde{P}_{0n}(O_n(m)) : m = 1, \ldots, M)$ recursively, from the least significant null hypothesis $H_0(O_n(M))$ to the most significant null hypothesis $H_0(O_n(1))$. That is,*

$$\widetilde{P}_{0n}(O_n(M)) = \inf_{\delta \geq P_{0n}(O_n(M))} \frac{\pi_{0n} M \delta}{|\{m : P_{0n}(m) \leq \delta\}|} \approxeq \pi_{0n} P_{0n}(O_n(M))$$

*and, for* $m = M-1, \ldots, 1$,

$$\begin{aligned} \widetilde{P}_{0n}(O_n(m)) &= \inf_{\delta \geq P_{0n}(O_n(m))} \frac{\pi_{0n} M \delta}{|\{m : P_{0n}(m) \leq \delta\}|} \\ &\approxeq \min\left\{ \frac{\pi_{0n} M P_{0n}(O_n(m))}{m}, \widetilde{P}_{0n}(O_n(m+1)) \right\}. \end{aligned} \tag{20}$$

5. *Reject null hypotheses with q-value less than or equal to* $\alpha$*, that is,*

$$\mathcal{R}_n(\alpha) = \{m : \widetilde{P}_{0n}(m) \leq \alpha\}. \tag{21}$$

Note that the $q$-values of Equation (20) may be rewritten as

$$\widetilde{P}_{0n}(O_n(m)) = \pi_{0n} \min_{h=m,\ldots,M} \left\{ \min\left\{ \frac{M}{h} P_{0n}(O_n(h)), 1 \right\} \right\}, \qquad m = 1, \ldots, M, \tag{22}$$

and are therefore simply the adjusted $p$-values for classical linear step-up Benjamini and Hochberg [1995] Procedure 1, multiplied by the estimated proportion of true null hypotheses $\pi_{0n}$. Indeed, Benjamini et al. [2006, Definition 5] argue that the Storey and Tibshirani [2003] method can be viewed as a special case of adaptive linear step-up Procedure 2.

Storey and Tibshirani's [2003] procedure is a joint step-up common-quantile procedure: it is a *joint* procedure, only in the sense that the $q$-values $\widetilde{P}_{0n}(m)$ are based on all $M$ unadjusted $p$-values $P_{0n}(m)$, via the estimator $\pi_{0n}$ of the proportion of true null hypotheses; it is a *step-up* procedure, in the sense that as soon as one null hypothesis is rejected, all remaining more significant hypotheses are rejected; it is a *common-quantile* procedure, in the sense that it is based on a $p$-value transformation of the test statistics.

### 2.3.2 Motivation

Storey and Tibshirani's [2003] adaptive linear step-up procedure can be motivated as follows. Consider a common unadjusted $p$-value cut-off $\delta$ and a set of rejected null hypotheses defined as

$$\mathcal{R}_n(\delta) = \{m : P_{0n}(m) \leq \delta\}. \tag{23}$$

**Estimation of the false discovery rate** For a large number of hypotheses $M$, the false discovery rate can be approximated as

$$FDR(\delta) = \mathrm{E}\left[\frac{V_n(\delta)}{R_n(\delta)}\right] \approxeq \frac{\mathrm{E}\left[V_n(\delta)\right]}{\mathrm{E}\left[R_n(\delta)\right]}. \tag{24}$$

The expected number of rejected hypotheses $\mathrm{E}\left[R_n(\delta)\right]$ can simply be estimated by the observed number $R_n(\delta)$. Under $\mathrm{U}(0,1)$ marginal distributions for the unadjusted $p$-values $(P_{0n}(m) : m \in \mathcal{H}_0)$ corresponding to the true null hypotheses $\mathcal{H}_0$, the expected number of Type I errors is

$$\mathrm{E}\left[V_n(\delta)\right] = \sum_{m \in \mathcal{H}_0} \Pr(P_{0n}(m) \leq \delta) = h_0 \delta.$$

Given an estimator $\pi_{0n}$ of the proportion $\pi_0 = h_0/M$ of true null hypotheses, this leads to the following estimator of the FDR,

$$\widehat{FDR}(\delta) = \frac{\pi_{0n} M \delta}{R_n(\delta)} = \frac{\pi_{0n} M \delta}{|\{m : P_{0n}(m) \leq \delta\}|}. \tag{25}$$

**Estimation of the proportion of true null hypotheses** A trivial, conservative estimator of the proportion $\pi_0 = h_0/M$ of true null hypotheses is one, $\pi_{0n} = 1$.

Storey and Tibshirani [2003] propose a less conservative estimator of $\pi_0$ by arguing that unadjusted $p$-values $(P_{0n}(m) : m \in \mathcal{H}_0)$ for the true null hypotheses have $\mathrm{U}(0,1)$ marginal distributions, whereas unadjusted $p$-values $(P_{0n}(m) : m \in \mathcal{H}_1)$ for the false null hypotheses should be close to zero. This leads to the following Bayesian heuristics.

$$\begin{aligned} \Pr(P_{0n}(m) > \lambda) &= \Pr(P_{0n}(m) > \lambda | H_0(m) = 1) \Pr(H_0(m) = 1) \\ &\quad + \Pr(P_{0n}(m) > \lambda | H_0(m) = 0) \Pr(H_0(m) = 0) \\ &\approxeq (1-\lambda)\pi_0 + 0(1-\pi_0), \end{aligned} \tag{26}$$

for $\lambda \in [0,1]$ above which the unadjusted $p$-values appear to be uniformly distributed. Thus,

$$\pi_0 \approxeq \frac{\Pr(P_{0n}(m) > \lambda)}{1-\lambda}. \tag{27}$$

The Storey and Tibshirani [2003] estimator of the proportion of true null hypotheses, based on the empirical survivor function of the $M$ unadjusted $p$-values evaluated at $\lambda$, is then given by,

$$\pi_{0n}(\lambda) = \frac{|\{m : P_{0n}(m) > \lambda\}|}{M(1-\lambda)}, \tag{28}$$

where $\lambda \in [0,1]$ is a to-be-determined tuning parameter.

As noted in Storey and Tibshirani [2003, Remark B], there is a bias-variance trade-off in selecting $\lambda$. The larger $\lambda$, the smaller the bias but the larger the variance. In particular, one recovers the conservative estimator $\pi_{0n}(\lambda) = 1$ when $\lambda = 0$. By contrast, for large values of $\lambda$, the estimator $\pi_{0n}(\lambda)$ is based on only a small fraction of the unadjusted $p$-values and is therefore variable. Storey and Tibshirani [2003] propose fitting a natural cubic spline to $(\lambda, \pi_{0n}(\lambda))$ for a range of values of the tuning parameter $\lambda$ and estimating $\pi_0$ by the fitted value at $\lambda = 1$.

**$q$-values** The so-called $q$-values appear to be nothing more than FDR-specific adjusted $p$-values. Indeed, the $q$-value for the $m$th null hypothesis $H_0(m)$ is defined as the smallest nominal FDR level at which this hypothesis is rejected, that is,

$$\widetilde{P}_{0n}(m) = \inf_{\delta \geq P_{0n}(m)} \widehat{FDR}(\delta) = \inf_{\delta \geq P_{0n}(m)} \frac{\pi_{0n} M \delta}{|\{m : P_{0n}(m) \leq \delta\}|}. \tag{29}$$

Note that in Procedure 3, the infimum over intervals $[P_{0n}(O_n(m)), 1]$ is approximated by a minimum over finite sets of unadjusted $p$-values $\{P_{0n}(O_n(h)) : h = m, \ldots, M\}$.

The $q$-values $\widetilde{P}_{0n}(m)$ lead to the same significance ranking as the unadjusted $p$-values $P_{0n}(m)$ and the set of rejected null hypotheses for controlling the FDR at nominal level $\alpha$ is given as usual by

$$\mathcal{R}_n(\alpha) = \{m : \widetilde{P}_{0n}(m) \leq \alpha\}. \tag{30}$$

Note that technically $\widetilde{P}_{0n}(m)$ is based on an estimator of the pFDR rather than FDR. However, as noted in Storey and Tibshirani [2003, Remark A], the approximation is reasonable for large $M$ because $\Pr(R_n > 0) \approxeq 1$ and $FDR \approxeq pFDR \approxeq \mathrm{E}[V_n]/\mathrm{E}[R_n]$.

**Properties** Storey and Tibshirani [2003, Remark D] summarize theoretical properties of adaptive linear step-up Procedure 3 [Storey et al., 2004]. In particular, it is argued that the method provides control of the FDR for large numbers of hypotheses $M$ and weak dependence structures.

In summary, the Storey and Tibshirani [2003] procedure for controlling the FDR at nominal level $\alpha$ is nothing more than the classical linear step-up procedure of Benjamini and Hochberg [1995] for controlling the FDR at nominal level $\alpha/\pi_{0n} \geq \alpha$, where $\pi_{0n}$ is an estimated proportion of true null hypotheses as in Equation (19).

# 3 Resampling-based empirical Bayes multiple testing

This section presents the resampling-based empirical Bayes multiple testing approach proposed in Dudoit and van der Laan [2007, Chapter 7] and van der Laan et al. [2005], for controlling generalized tail probability and expected value error rates. The interested reader is referred to these earlier publications for further detail, including a proof of Type I error control, the derivation of adjusted $p$-values, and connections to the frequentist FDR-controlling linear step-up procedure of Benjamini and Hochberg [1995].

## 3.1 Resampling-based empirical Bayes multiple testing procedure

Given random $M$-vectors of test statistics $Z_0 = (Z_0(m) : m = 1, \ldots, M)$ and $Z = (Z(m) : m = 1, \ldots, M)$, a set of null hypotheses $\mathcal{H} \subseteq \{1, \ldots, M\}$, and an $M$-vector of cut-offs $c = (c(m) : m = 1, \ldots, M) \in \mathbb{R}^M$ that define one-sided rejection regions of the form $\mathcal{C}(m) = (c(m), +\infty)$, introduce the following notation for the number of false positives (i.e., Type I errors), the number of true positives, the number of rejected hypotheses, and a function $g$ of the numbers of false positives and true positives,

$$\begin{aligned} V(c; \mathcal{H}, Z) &\equiv \sum_{m \in \mathcal{H}} \mathrm{I}\left(Z(m) > c(m)\right), \\ S(c; \mathcal{H}, Z) &\equiv \sum_{m \notin \mathcal{H}} \mathrm{I}\left(Z(m) > c(m)\right), \\ R(c; \mathcal{H}, Z_0, Z) &\equiv V(c; \mathcal{H}, Z_0) + S(c; \mathcal{H}, Z), \end{aligned} \tag{31}$$

and

$$G(c; \mathcal{H}, Z_0, Z) \equiv g(V(c; \mathcal{H}, Z_0), S(c; \mathcal{H}, Z)).$$

In addition, define the following $g$-specific function for the generalized tail probability $gTP(q, g) = \Pr(g(V_n, S_n) > q)$ and expected value $gEV(g) = \mathrm{E}[g(V_n, S_n)]$ error rates,

$$\widetilde{G}(c; \mathcal{H}, Z_0, Z) \equiv \begin{cases} \mathrm{I}(G(c; \mathcal{H}, Z_0, Z) > q) & \text{for } gTP(q, g) \\ G(c; \mathcal{H}, Z_0, Z) & \text{for } gEV(g) \end{cases}, \tag{32}$$

so that these error rates can be expressed as

$$\theta_n(c) \equiv \mathrm{E}[\widetilde{G}(c; \mathcal{H}_0, T_n, T_n)]. \tag{33}$$

In order to control $gTP(q, g)$ and $gEV(g)$ at level $\alpha$, one seeks cut-offs $c_n = (c_n(m) : m = 1, \ldots, M)$, for the test statistics $T_n = (T_n(m) : m = 1, \ldots, M) \sim Q_n$, so that the following Type I error constraint is satisfied,

$$\begin{aligned} \theta_n(c_n) = \mathrm{E}[\widetilde{G}(c_n; \mathcal{H}_0, T_n, T_n)] &\leq \alpha \qquad \text{[finite sample control]} \\ \limsup_{n \to \infty} \theta_n(c_n) = \limsup_{n \to \infty} \mathrm{E}[\widetilde{G}(c_n; \mathcal{H}_0, T_n, T_n)]) &\leq \alpha \qquad \text{[asymptotic control]}. \end{aligned} \tag{34}$$

However, one is immediately faced with the problem that the distribution of $G(c_n; \mathcal{H}_0, T_n, T_n)$ depends on the unknown data generating distribution $P$, via the unknown set of true null hypotheses $\mathcal{H}_0$ and joint distribution $Q_n$ of the test statistics $T_n$.

The resampling-based empirical Bayes approach replaces the unknown $g$-specific function of the numbers of false positives and true positives $G(c; \mathcal{H}_0, T_n, T_n)$ by the corresponding guessed function $G(c; \mathcal{H}_{0n}, T_{0n}, T_n)$, where $T_n \sim Q_n$ is the $M$-vector of observed test statistics, $T_{0n} \sim Q_{0n}$ is an $M$-vector of null test statistics, and $\mathcal{H}_{0n} \sim Q_{0n}^{\mathcal{H}}$ is a guessed set of true null hypotheses.

The null test statistics $T_{0n}$ and the guessed sets $\mathcal{H}_{0n}$ are sampled independently, given the empirical distribution $P_n$, from distributions $Q_{0n}$ and $Q_{0n}^{\mathcal{H}}$, chosen conservatively so that the guessed function $G(c; \mathcal{H}_{0n}, T_{0n}, T_n)$ is asymptotically stochastically greater than the corresponding true function $G(c; \mathcal{H}_0, T_n, T_n)$.

**Procedure 4 [gTP- and gEV-controlling resampling-based empirical Bayes procedure]**
*Consider the simultaneous test of $M$ null hypotheses $H_0(m)$, $m = 1, \ldots, M$, based on an $M$-vector of test statistics $T_n = (T_n(m) : m = 1, \ldots, M)$, with distribution $Q_n = Q_n(P)$. Given a function $g$, that satisfies monotonicity Assumptions MgV and MgS, the following resampling-based empirical Bayes procedure may be used to control the generalized tail probability error rate, $gTP(q, g) = \Pr(g(V_n, S_n) > q)$, and the generalized expected value error rate, $gEV(g) = \mathrm{E}[g(V_n, S_n)]$.*

1. *Generate $B$ pairs $\{(T_{0n}^b, \mathcal{H}_{0n}^b) : b = 1, \ldots, B\}$ of null test statistics $T_{0n}^b$ and random guessed sets $\mathcal{H}_{0n}^b$ of true null hypotheses as follows.*

   (a) *The $M$-vectors of* null test statistics *$T_{0n}^b$ have a null distribution $Q_{0n}$, such as the bootstrap-based null-transformed test statistics null distributions described in Section 3.2 and Dudoit and van der Laan [2007, Chapter 2].*

   (b) *The* random guessed sets of true null hypotheses *$\mathcal{H}_{0n}^b$ have a distribution $Q_{0n}^{\pi}$ that corresponds to $M$ independent Bernoulli random variables with parameters $\pi_{0n}(T_n(m))$. That is, generate binary random $M$-vectors $H_{0n}^b = (H_{0n}^b(m) : m = 1, \ldots, M)$ of null hypotheses as*

$$H_{0n}^b(m) \overset{\perp}{\sim} Bernoulli(\pi_{0n}(T_n(m))), \qquad m = 1, \ldots, M, \tag{35}$$

   *and define sets*

$$\mathcal{H}_{0n}^b \equiv \left\{m : H_{0n}^b(m) = 1\right\}. \tag{36}$$

   *Here, $\pi_{0n}(t)$ is an estimated true null hypothesis posterior probability function, such as the estimated local q-value function*

$$\pi_{0n}(t) = \min\left\{1, \frac{\pi_{0n} f_{0n}(t)}{f_n(t)}\right\}, \tag{37}$$

   *corresponding to the marginal non-parametric mixture model of Section 3.3.*

   (c) *Null test statistics $T_{0n}^b$ and guessed sets $\mathcal{H}_{0n}^b$ are independent, given the empirical distribution $P_n$.*

2. *For any given test statistic cut-off vector $c = (c(m) : m = 1, \ldots, M)$, compute, for each of the $B$ pairs $(T_{0n}^b, \mathcal{H}_{0n}^b)$, the corresponding* guessed *$g$*-specific function of the numbers of false positives and true positives,

$$G(c; \mathcal{H}_{0n}^b, T_{0n}^b, T_n) = g(V(c; \mathcal{H}_{0n}^b, T_{0n}^b), S(c; \mathcal{H}_{0n}^b, T_n)). \tag{38}$$

   *An estimator of the (gTP or gEV) Type I error rate $\theta_n(c) = \mathrm{E}[\widetilde{G}(c; \mathcal{H}_0, T_n, T_n)]$ is then given by*

$$\hat{\theta}_n(c) = \frac{1}{B} \sum_{b=1}^{B} \widetilde{G}(c; \mathcal{H}_{0n}^b, T_{0n}^b, T_n). \tag{39}$$

3. *For user-supplied Type I error level $\alpha \in (0, 1)$, derive a cut-off vector $c_n$ that satisfies the* empirical Type I error constraint

$$\hat{\theta}_n(c_n) \leq \alpha. \tag{40}$$

   **Common-cut-off procedure.** *The* common cut-off *$\gamma_n$ is the smallest (i.e., least conservative) value $\gamma$ for which the constraint in Equation* (40) *is satisfied. That is,*

$$\gamma_n \equiv \inf\left\{\gamma \in I\!R : \hat{\theta}_n\left(\gamma^{(M)}\right) \leq \alpha\right\}, \tag{41}$$

*where $\gamma^{(M)}$ denotes the $M$-vector with all elements equal to $\gamma$, i.e., $\gamma^{(M)}(m) = \gamma$, $\forall\ m = 1, \ldots, M$. The adjusted p-values may be approximated as*

$$\widetilde{p}_{0n}(o_n(m)) \cong \min_{h \in \overline{\mathcal{O}}_n(m)} \hat{\theta}_n \left( (t_n(h))^{(M)} \right), \tag{42}$$

*where $O_n(m)$ denote the indices for the ordered test statistics $T_n(O_n(m))$, so that $T_n(O_n(1)) \geq \cdots \geq T_n(O_n(M))$, and $\overline{\mathcal{O}}_n(m) \equiv \{O_n(m), \ldots, O_n(M)\}$.*

**Common-quantile procedure.** *The* common quantile probability $\delta_n$, *corresponding to the test statistics null distribution $Q_{0n}$, is the smallest (i.e., least conservative) value $\delta$ for which the constraint in Equation* (40) *is satisfied. That is,*

$$\delta_n \equiv \inf \left\{ \delta \in [0,1] : \hat{\theta}_n \left( q_{0n}^{-1}(\delta) \right) \leq \alpha \right\}, \tag{43}$$

*where $q_{0n}^{-1}(\delta) = (Q_{0n,m}^{-1}(\delta) : m = 1, \ldots, M)$ denotes the $M$-vector of $\delta$-quantiles for the null distribution $Q_{0n}$.*
*The adjusted p-values may be approximated as*

$$\widetilde{p}_{0n}(o_n(m)) \cong \min_{h \in \overline{\mathcal{O}}_n(m)} \hat{\theta}_n \left( q_{0n}^{-1}(1 - p_{0n}(h)) \right), \tag{44}$$

*where $p_{0n}(m) = 1 - Q_{0n,m}(t_n(m))$ is the unadjusted p-value for null hypothesis $H_0(m)$, $O_n(m)$ denote the indices for the ordered unadjusted p-values $P_{0n}(O_n(m))$, so that $P_{0n}(O_n(1)) \leq \cdots \leq P_{0n}(O_n(M))$, and $\overline{\mathcal{O}}_n(m) \equiv \{O_n(m), \ldots, O_n(M)\}$.*

Following the characterization of MTPs in Dudoit and van der Laan [2007, Section 1.2.7], Procedure 4 is a *joint single-step common-cut-off* or *common-quantile* procedure.

The two main ingredients of a resampling-based empirical Bayes procedure are discussed next: the null distribution $Q_0$ (or estimator thereof, $Q_{0n}$) for the $M$-vectors of null test statistics $T_{0n}$ (Section 3.2) and the distribution $Q_0^{\mathcal{H}}$ (or estimator thereof, $Q_{0n}^{\mathcal{H}}$) for the random guessed sets of true null hypotheses $\mathcal{H}_{0n}$ (Section 3.3). Further detail can be found in Dudoit and van der Laan [2007, Chapter 7] and van der Laan et al. [2005].

## 3.2 Distribution for the null test statistics

Test statistics null distributions are briefly discussed in Sections 1.3.5 and 4.2.3 of the present article and in depth in Dudoit and van der Laan [2007, Chapters 2 and 7].

## 3.3 Distribution for the guessed sets of true null hypotheses

The following is only one among many reasonable candidate distributions $Q_{0n}^{\mathcal{H}}$ for the guessed sets of true null hypotheses, that does not assume independence of the test statistics.

### 3.3.1 Common marginal non-parametric mixture model

Consider $M$ identically distributed pairs of test statistics and null hypotheses $((T_n(m), H_0(m)) : m = 1, \ldots, M)$. Test statistics are assumed to have the following *common marginal non-parametric mixture distribution*,

$$T_n(m) \sim f \equiv \pi_0 f_0 + (1 - \pi_0) f_1, \qquad m = 1, \ldots, M, \tag{45}$$

where $\pi_0$ denotes the *prior probability of a true null hypothesis*, $f_0$ the *marginal null density of the test statistics*, and $f_1$ the *marginal alternative density of the test statistics*, i.e., $\pi_0 \equiv \Pr(H_0(m) = 1)$, $T_n(m) | \{H_0(m) =$

$1\} \sim f_0$, and $T_n(m)|\{H_0(m) = 0\} \sim f_1$.

### 3.3.2 Local $q$-values

A parameter of interest, for generating guessed sets of true null hypotheses under the marginal non-parametric mixture model of Equation (45), is the *local q-value function*, i.e., the posterior probability function for a true null hypothesis $H_0(m)$, given the corresponding test statistic $T_n(m)$,

$$\pi_0(t) \equiv \Pr(H_0(m) = 1|T_n(m) = t) = \frac{\pi_0 f_0(t)}{f(t)}, \qquad m = 1, \ldots, M. \tag{46}$$

Empirical Bayes $q$-values are similar in some sense to frequentist $p$-values: the smaller the $q$-value $\pi_0(T_n(m))$, the stronger the evidence against the corresponding null hypothesis $H_0(m)$.

In practice, the local $q$-value function $\pi_0(t)$ is unknown, as it depends on the unknown true null hypothesis prior probability $\pi_0$, test statistic marginal null density $f_0$, and test statistic marginal density $f$. Estimators of $\pi_0(t)$ may be obtained by the *plug-in* method, from estimators of the three main parameters, $\pi_0$, $f_0$, and $f$, of the mixture model of Equation (45).

Note that the $q$-values defined here in Equation (46) are different in nature from the $q$-values of Equations (20) and (29) for the linear step-up procedure of Storey and Tibshirani [2003], as the latter are actually adjusted $p$-values for FDR control.

### 3.3.3 Estimation of the true null hypothesis prior probability $\pi_0$

A trivial estimator $\pi_{0n}$ of the prior probability $\pi_0$ of a true null hypothesis is the conservative value of one, i.e., $\pi_{0n} = 1$.

Alternately, $\pi_0$ may be estimated from prior knowledge or as a by-product of a computationally convenient procedure, such as the FDR-controlling adaptive linear step-up procedure of Benjamini and Hochberg [2000] or two-stage linear step-up procedure of Benjamini et al. [2006].

Various approaches are summarized in Section 4.2.5 and Table 2.

### 3.3.4 Estimation of the test statistic marginal null density $f_0$

For the test of single-parameter null hypotheses using $t$-statistics, the common marginal null density $f_0$ is simply a standard Gaussian density, i.e., $T_n(m)|\{H_0(m) = 1\} \sim \mathrm{N}(0, 1)$ (Section 4.2.5).

For other types of test statistics, one may estimate $f_0$ by kernel density smoothing of the $M \times B$ pooled elements of a matrix $\mathbf{Z}_n^B$ of null-transformed bootstrap test statistics [Dudoit and van der Laan, 2007, Procedures 2.3 and 2.4].

### 3.3.5 Estimation of the test statistic marginal density $f$

For the test of single-parameter null hypotheses using $t$-statistics, the common marginal density $f$ may be estimated based on an estimator of the asymptotic $M$-variate Gaussian distribution of the $M$-vector of $t$-statistics $T_n$ (Section 4.2.5).

For other types of test statistics, one may estimate $f$ by kernel density smoothing of the $M \times B$ pooled elements of a matrix $\mathbf{T}_n^B$ of raw (before null transformation) bootstrap test statistics [Dudoit and van der Laan, 2007, Procedures 2.3 and 2.4].

## 3.4 Estimation of the proportion of true null hypotheses

A parameter of interest in multiple hypothesis testing is the number of true null hypotheses $h_0$. The following two estimators of $h_0$ may be obtained as by-products of the resampling-based empirical Bayes approach.

### 3.4.1 $q$-value-based empirical Bayes estimator

In the Bayesian context of Section 3.3, the local $q$-value function $\pi_0(t)$, used to generate the random guessed sets of true null hypotheses in Procedure 4, is a posterior probability function for the true null hypotheses (Equation (46)).

The prior probability $\pi_0 = \Pr(H_0(m) = 1)$ of a true null hypothesis yields an *a priori*, i.e., non data-driven, estimator of the number $h_0$ of true null hypotheses. Indeed, the a priori expected value of $h_0$ is

$$\mathrm{E}[h_0] = \mathrm{E}\left[\sum_{m=1}^{M} \mathrm{I}(H_0(m) = 1)\right] = \sum_{m=1}^{M} \Pr(H_0(m) = 1) = M\pi_0. \tag{47}$$

The local $q$-values $\pi_0(T_n(m)) = \Pr(H_0(m) = 1|T_n(m))$ are posterior probabilities for the true null hypotheses and in turn lead to the following *a posteriori*, i.e., data-driven, estimator of $h_0$. The a posteriori expected value of $h_0$ is

$$\begin{aligned} \mathrm{E}[h_0|\mathcal{X}_n] &= \mathrm{E}\left[\sum_{m=1}^{M} \mathrm{I}(H_0(m) = 1)\middle| \mathcal{X}_n\right] \\ &= \sum_{m=1}^{M} \Pr(H_0(m) = 1|\mathcal{X}_n) \\ &= \sum_{m=1}^{M} \Pr(H_0(m) = 1|T_n(m)) \\ &= \sum_{m=1}^{M} \pi_0(T_n(m)), \end{aligned} \tag{48}$$

under the assumption that the null hypotheses $H_0(m)$ are conditionally independent of the data $\mathcal{X}_n$ given the corresponding test statistics $T_n(m)$.

Thus, the number of true null hypotheses $h_0$ may be estimated by the sum of the estimated local $q$-values,

$$h_{0n}^{QV} = \sum_{m=1}^{M} \pi_{0n}(T_n(m)). \tag{49}$$

### 3.4.2 Resampling-based empirical Bayes estimator

A resampling-based empirical Bayes estimator of the number of true null hypotheses $h_0$ can also be obtained as a by-product of Procedure 4, by averaging the cardinality $|\mathcal{H}_{0n}^b|$ of the guessed sets of true null hypotheses,

$$h_{0n}^{REB} = \frac{1}{B}\sum_{b=1}^{B} |\mathcal{H}_{0n}^b|. \tag{50}$$

Keeping track of the $B$ guessed numbers $|\mathcal{H}_{0n}^b|$ of true null hypotheses provides some indication of the stability of the guessed sets.

The above two estimators should be very similar. Indeed, the $q$-value-based estimator $h_{0n}^{QV}$ is the expected value of the guessed numbers $|\mathcal{H}_{0n}^b|$ of true null hypotheses and, for a large number $B$ of resampled datasets, the empirical mean $h_{0n}^{REB}$ should converge to its expected value of $h_{0n}^{QV}$.

# 4 Simulation study

## 4.1 Simulation model

Simulated data consist of learning sets $\mathcal{X}_n = \{X_i : i = 1, \ldots, n\} \overset{IID}{\sim} \mathrm{N}(\psi, \sigma)$, of $n$ independent and identically distributed random variables from an $M$-variate Gaussian data generating distribution $P$, with mean vector $\psi = (\psi(m) : m = 1, \ldots, M) = \Psi(P) = \mathrm{E}[X]$ and covariance matrix $\sigma = (\sigma(m, m') : m, m' = 1, \ldots, M) = \Sigma(P) = \mathrm{Cov}[X]$. The shorter notation $\sigma^2(m) \equiv \sigma(m, m)$ may be used for variances and the correlation matrix corresponding to $\sigma$ is denoted by $\sigma^* = \Sigma^*(P) = \mathrm{Cor}[X]$.

Both the mean vector $\psi$ and the covariance matrix $\sigma$ are treated as unknown parameters; the parameter of interest is the mean vector $\psi$.

## 4.2 Multiple testing procedures

### 4.2.1 Null and alternative hypotheses

The simulation study concerns the two-sided test of the $M$ null hypotheses $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$ vs. the alternative hypotheses $H_1(m) = \mathrm{I}(\psi(m) \neq \psi_0(m))$, $m = 1, \ldots, M$. For simplicity, and without loss of generality, the null values are set equal to zero, i.e., $\psi_0(m) = 0$.

### 4.2.2 Test statistics

The $M$ null hypotheses are tested based on usual one-sample $t$-statistics,

$$T_n(m) \equiv \sqrt{n}\frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}, \tag{51}$$

where $\psi_n(m) = \bar{X}_n(m) = \sum_i X_i(m)/n$ and $\sigma_n^2(m) = \sum_i (X_i(m) - \bar{X}_n(m))^2/(n-1)$ denote, respectively, the empirical means and variances for the $M$ elements of $X$.

### 4.2.3 Test statistics null distribution

The unknown asymptotic joint null distribution $Q_0$ of the $t$-statistics of Equation (51) is the $M$-variate Gaussian distribution $\mathrm{N}(0, \sigma^*)$, with mean vector zero and covariance matrix equal to the unknown correlation matrix $\sigma^*$ of $X$.

A parametric estimator $Q_{0n}$ of $Q_0$ is the Gaussian distribution $\mathrm{N}(0, \sigma_n^*)$, where $\sigma_n^*$ is the empirical correlation matrix of the learning set $\mathcal{X}_n$.

This joint distribution $Q_{0n}$ can be approximated by the empirical distribution of the $B$ columns $\{Z_n^B(\cdot, b) \sim \mathrm{N}(0, \sigma_n^*) : b = 1, \ldots, B\}$ of a matrix $\mathbf{Z}_n^B$ simulated from $\mathrm{N}(0, \sigma_n^*)$ ($B = 10,000$ in the present simulation study).

### 4.2.4 FDR-controlling linear step-up procedures

The simulation study examines the following five linear step-up procedures, summarized in Table 2.

1. LSU.BH: Benjamini and Hochberg [1995] classical linear step-up Procedure 1.
2. LSU.O: Oracle linear step-up procedure, using the unknown number of true null hypotheses $h_0$ in place of $h_{0n}$ in Procedure 2.
3. LSU.ABH: Benjamini and Hochberg [2000] adaptive linear step-up procedure, using $h_{0n}^{ABH}$ from Equation (16) in Procedure 2.
4. LSU.TST: Benjamini et al. [2006] adaptive two-stage linear step-up procedure, using $h_{0n}^{TST}(\alpha)$, $\alpha = 0.05, 0.10$, from Equation (18) in Procedure 2.

5. LSU.ST: Storey and Tibshirani [2003] adaptive linear step-up Procedure 3, using $h_{0n}^{ST}(\lambda)$ from Equation (17) in Procedure 2.

Each of the five linear step-up procedures is given as input two-sided unadjusted $p$-values $P_{0n}(m)$ computed under a standard Gaussian test statistic marginal null distribution. Specifically,

$$P_{0n}(m) = 2(1 - \Phi(T_n(m))), \tag{52}$$

where $\Phi$ is the $\mathrm{N}(0,1)$ cumulative distribution function (CDF).

Estimators of the number $h_0$ of true null hypotheses are examined for the last three adaptive procedures.

The first four procedures are implemented using the function `mt.rawp2adjp` from the Bioconductor R package multtest. The LSU.ST procedure of Storey and Tibshirani [2003] is implemented using the function `qvalue` from the R package qvalue, with default argument values.

### 4.2.5 FDR-controlling resampling-based empirical Bayes procedures

The above linear step-up procedures are compared to FDR-controlling resampling-based empirical Bayes Procedure 4, with common cut-offs for the test statistics defined as

$$\gamma_n = \inf\left\{\gamma \in I\!R : \frac{1}{B}\sum_{b=1}^{B} \frac{V(\gamma^{(M)}; \mathcal{H}_{0n}^b, T_{0n}^b)}{\max\left\{V(\gamma^{(M)}; \mathcal{H}_{0n}^b, T_{0n}^b) + S(\gamma^{(M)}; \mathcal{H}_{0n}^b, T_n), 1\right\}} \leq \alpha\right\}. \tag{53}$$

In the simulation study, the common cut-offs $\gamma_n$ are selected based on $B = 10,000$ pairs $\{(T_{0n}^b, \mathcal{H}_{0n}^b) : b = 1, \ldots, B\}$ of null test statistics and guessed sets of true null hypotheses, from the discrete set $\{0,\ 0.05,\ 0.10,\ \ldots,\ 4.50\}$, i.e., from the interval $[0, 4.50]$, with a resolution of 0.05.

The two main ingredients for Procedure 4 are the null distribution $Q_0$ (or estimator thereof, $Q_{0n}$) for the $M$-vectors of null test statistics $T_{0n}$ (Section 3.2) and the distribution $Q_0^{\mathcal{H}}$ (or estimator thereof, $Q_{0n}^{\mathcal{H}}$) for the random guessed sets of true null hypotheses $\mathcal{H}_{0n}$ (Section 3.3). In the case of the common marginal non-parametric mixture model of Section 3.3, $Q_0^{\mathcal{H}}$ is specified by three parameters: the true null hypothesis prior probability $\pi_0$, the test statistic marginal null density $f_0$, and the test statistic marginal density $f$.

The following four versions of empirical Bayes Procedure 4 are considered in terms of the estimator $\pi_{0n}$ of the true null hypothesis prior probability $\pi_0$ (Table 2).

1. EB.C: Conservative prior $\pi_{0n} = 1$.
2. EB.O: Oracle prior $\pi_{0n} = h_0/M$, based on the unknown number $h_0$ of true null hypotheses.
3. EB.ABH: Data-adaptive prior $\pi_{0n} = h_{0n}^{ABH}/M$, based on the Benjamini and Hochberg [2000] estimator $h_{0n}^{ABH}$ of the number of true null hypotheses (Equation (16)).
4. EB.QV: Data-adaptive prior $\pi_{0n} = h_{0n}^{QV}/M$, based on the sum of the local $q$-values $\pi_{0n}(T_n(m))$ computed with an initial conservative prior $\pi_{0n} = 1$ (Equation (49)).

For each of these procedures, the estimators remaining to be specified, $Q_{0n}$, $f_{0n}$, and $f_n$, are as follows.

- *Test statistics joint null distribution, $Q_{0n}$.* $M$-variate Gaussian distribution $\mathrm{N}(0, \sigma_n^*)$, where $\sigma_n^*$ is the empirical correlation matrix of the learning set $\mathcal{X}_n$, as in Section 4.2.3.
- *Test statistic marginal null density, $f_{0n}$.* Standard Gaussian density $f_{0n} \sim \mathrm{N}(0,1)$.
- *Test statistic marginal density, $f_n$.* Kernel density smoothed function of the $M \times B$ pooled elements of a matrix $\mathbf{T}_n^B$, with columns $T_n^B(\cdot, b) \sim \mathrm{N}(T_n, \sigma_n^*)$, $b = 1, \ldots, B$ ($B = 10,000$ in the present simulation study).

Estimators of the number of true null hypotheses $h_0$, based on the sum of the local $q$-values $\pi_{0n}(T_n(m))$ (Equation (49)), are examined for each of the four empirical Bayes procedures, namely, EB.C, EB.O, EB.ABH, and EB.QV.

## 4.3 Simulation study design

### 4.3.1 Simulation parameters

Although a simple Gaussian data generating distribution is used, a broad range of testing scenarios (including extreme ones) are covered by varying the following model parameters. The simulation results should therefore provide a fairly complete assessment of the Type I error and power properties of the FDR-controlling procedures of Table 2.

- *Sample size, $n$.* $n = 10,\ 30,\ 100,\ 250,\ +\infty$.
- *Number of null hypotheses, $M$.* $M = 40,\ 400,\ 2,000$.
- *Proportion of true null hypotheses, $h_0/M$.* $h_0/M = 0.50,\ 0.75,\ 0.95,\ 1.00$.
- *Shift parameter vector, $d_n$.* The elements of the mean vector $\psi$ are expressed as $\psi(m) = d_n(m)\sigma(m)/\sqrt{n}$, in terms of a shift vector $d_n$. For the true null hypotheses, i.e., for $m \in \mathcal{H}_0$, $d_n(m) = 0$. For the false null hypotheses, i.e., for $m \in \mathcal{H}_1$, $d_n(m) = 2,\ 3,\ 4$.
- *Correlation matrix, $\sigma^*$.* The following three correlation structures are considered.
  - *No correlation*, where $\sigma^* = I_M$, the $M \times M$ identity matrix.
  - *Constant correlation*, where all off-diagonal elements of $\sigma^*$ are set to a common value: $\sigma^*(m,m) = 1$, for $m = 1, \ldots, M$; $\sigma^*(m,m') = 0.50$, for $m \neq m' = 1, \ldots, M$.
  - *Empirical microarray correlation*, where $\sigma^*$ corresponds to a random $M \times M$ submatrix of the probes $\times$ probes correlation matrix for the Golub et al. [1999] leukemia microarray dataset. [1]

Detailed results for some parameter combinations are reported in Section 5. Results for other parameter values are only briefly discussed in the present article and are posted on the website companion.

### 4.3.2 Simulated datasets

For each simulation scenario (i.e., each combination of values for parameters $n$, $M$, $h_0/M$, $d_n$, and $\sigma^*$, from Section 4.3.1), generate $A = 500$ learning sets $\mathcal{X}_n^a = \{X_i^a : i = 1, \ldots, n\} \overset{IID}{\sim} \mathrm{N}(\psi, \sigma)$, $a = 1, \ldots, A$, where the elements of the $M$-dimensional mean vector $\psi = (\psi(m) : m = 1, \ldots, M)$ are defined as $\psi(m) = d_n(m)\sigma(m)/\sqrt{n}$, in terms of a shift vector $d_n = (d_n(m) : m = 1, \ldots, M)$.

For each simulated dataset $\mathcal{X}_n^a$, compute cut-offs (resampling-based empirical Bayes procedures EB) and adjusted $p$-values $\widetilde{P}_{0n}^a(m)$ (linear step-up procedures LSU) for each of the multiple testing procedures summarized in Table 2.

### 4.3.3 Type I error control and power comparison

**Estimation of Type I error rate and power** For each simulated dataset $\mathcal{X}_n^a$ and given nominal Type I error level $\alpha$, compute, for each MTP, the numbers of false positives $V_n^a(\alpha)$ and true positives $S_n^a(\alpha)$. Specifically, given adjusted $p$-values $\widetilde{P}_{0n}^a(m)$, define

$$V_n^a(\alpha) \equiv \sum_{m \in \mathcal{H}_0} \mathrm{I}\left(\widetilde{P}_{0n}^a(m) \leq \alpha\right) \qquad \text{and} \qquad S_n^a(\alpha) \equiv \sum_{m \notin \mathcal{H}_0} \mathrm{I}\left(\widetilde{P}_{0n}^a(m) \leq \alpha\right). \tag{54}$$

Likewise for procedures whose results are expressed in terms of rejection regions for the test statistics.

[1] The following three pre-processing steps were applied to the $7,129 \times 38$ probes $\times$ patients matrix of expression measures corresponding to the training set of 38 patients (object `Golub_Train` in the Bioconductor R package `golubEsets`): (i) *thresholding*, floor of 100 and ceiling of 16,000; (ii) *filtering*, exclusion of probes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer, respectively, to the maximum and minimum intensities for a particular probe across the 38 mRNA samples; (iii) *base-2 logarithmic transformation*. These pre-processing steps resulted in a $3,051 \times 38$ probes $\times$ patients matrix of expression measures, from which one can compute a $3,051 \times 3,051$ probe correlation matrix and extract a random $M \times M$ submatrix $\sigma^*$.

The *actual Type I error rate* is estimated as follows and compared to the *nominal Type I error level* $\alpha$,

$$FDR(\alpha) \equiv \frac{1}{A}\sum_{a=1}^{A}\frac{V_n^a(\alpha)}{\max\{V_n^a(\alpha)+S_n^a(\alpha),1\}}. \tag{55}$$

The *average power* of a given MTP is estimated by

$$AvgPwr(\alpha) \equiv \frac{1}{h_1}\frac{1}{A}\sum_{a=1}^{A}S_n^a(\alpha). \tag{56}$$

The simulation error for the actual Type I error rate and power is of the order $1/\sqrt{A} = 1/\sqrt{500} \approxeq 0.045$.

Table 3 reports numerical summaries of the actual Type I error rate and average power of FDR-controlling procedures from Table 2, for a nominal Type I error level $\alpha = 0.05$.

**Type I error control comparison** For a given simulation scenario, plot, for each MTP, the difference between the nominal and actual Type I error rates vs. the nominal Type I error level, that is, plot

$$\alpha - FDR(\alpha) \qquad \text{vs.} \qquad \alpha,$$

for $\alpha \in \{0.01, 0.02, \ldots, 0.50\}$. Positive (negative) differences correspond to (anti-) conservative MTPs; the higher the curve, the more conservative the procedure.

**Power comparison** For a given simulation scenario, *receiver operator characteristic* (ROC) curves may be used for a fair comparison of different MTPs in terms of power. ROC curves are obtained by plotting, for each MTP, power vs. *actual* Type I error rate, i.e., $AvgPwr(\alpha)$ vs. $FDR(\alpha)$, for a range of nominal Type I error levels $\alpha$.

However, due to possibly large variations in power between simulation scenarios, we consider instead the following modified display, which facilitates comparisons across scenarios. For a given scenario and MTP, a linear interpolation of the power $AvgPwr(\alpha)$ as a function of the actual Type I error rate $FDR(\alpha)$ is obtained using the R function `approxfun` (with default argument values). The difference in power between each procedure of interest and a baseline procedure (without loss of generality, procedure LSU.BH) is then taken and plotted against the actual Type I error rate.

#### 4.3.4 Estimation of the proportion of true null hypotheses

A parameter of interest in multiple hypothesis testing is the proportion of true null hypotheses $h_0/M$. Accordingly, the properties of the following six estimators of $h_0/M$ are investigated and compared, using boxplots of the corresponding estimates over the $A = 500$ simulated datasets: estimator of Equation (16) for the adaptive linear step-up LSU.ABH procedure of Benjamini and Hochberg [2000] (Section 2.2); estimator of Equation (17) for the adaptive linear step-up LSU.ST procedure of Storey and Tibshirani [2003] (Sections 2.2 and 2.3); and $q$-value-based estimator of Equation (49) for resampling-based empirical Bayes procedures EB.C, EB.O, EB.ABH, and EB.QV, each corresponding to a particular estimator $\pi_{0n}$ of the true null hypothesis prior probability $\pi_0$, as summarized in Table 2 (Section 3.4).

# 5 Results

## 5.1 Type I error control and power comparison

### 5.1.1 Actual Type I error rate and power at a given nominal FDR level

Table 3 reports numerical summaries of the *actual* Type I error rate $FDR(\alpha)$ and average power $AvgPwr(\alpha)$ of FDR-controlling procedures from Table 2, for a *nominal* Type I error level $\alpha = 0.05$.

The original linear step-up procedure of Benjamini and Hochberg [1995] and adaptive versions thereof [Benjamini and Hochberg, 2000, Benjamini et al., 2006] consistently offer conservative Type I error control across combinations of simulation parameters, with the adaptive procedures being, as expected, less conservative and more powerful (Table 3, LSU.BH, LSU.ABH, and LSU.TST). Two-stage linear step-up procedure LSU.TST appears to be more conservative than adaptive procedure LSU.ABH.

The adaptive linear step-up procedure of Storey and Tibshirani [2003], as implemented in the R package qvalue, is typically anti-conservative, particularly for smaller numbers of hypotheses $M$ and more complex correlation structures $\sigma^*$ (Table 3, LSU.ST). When assumptions underlying the method are met (i.e., independent test statistics and a large number of hypotheses $M$), the LSU.ST procedure outpowers all but the oracle procedures at a given nominal Type I error level $\alpha = 0.05$.

The performance of the resampling-based empirical Bayes procedures varies with the correlation structure $\sigma^*$ and proportion of true null hypotheses $h_0/M$ (Table 3, EB.C, EB.ABH, and EB.QV). For the empirical microarray correlation structure, the empirical Bayes procedures and Storey and Tibshirani's [2003] linear step-up procedure LSU.ST offer significant gains in power over the procedures of Benjamini and colleagues (LSU.BH, LSU.ABH, and LSU.TST). The empirical Bayes procedure EB.C, with the most conservative true null hypothesis prior probability $\pi_{0n} = 1$, demonstrates this increase in power while maintaining equal or better Type I error control than the LSU.ST procedure. Using a data-adaptive prior $\pi_{0n}$ for the empirical Bayes method (EB.ABH and EB.QV) further increases power (equal to or over that of LSU.ST), without sacrificing much with respect to Type I error control. Under constant, heavy correlation, the empirical Bayes procedures yield the highest average power when testing at nominal Type I error level $\alpha = 0.05$. This increase in power comes, however, at the expense of Type I error control. It is therefore not advisable to relax the prior under conditions of heavy correlation, as doing so may lead to anti-conservative behavior.

Oracle procedures, given the unknown proportion of/prior for the true null hypotheses, tend to be more powerful than their empirical counterparts, possibly at the detriment of Type I error control (LSU.O vs. LSU.BH, LSU.ABH, LSU.TST, and LSU.ST; EB.O vs. EB.C, EB.ABH, and EB.QV). This is of course to be expected when comparing oracle procedures to conservative procedures with $\pi_{0n} = h_{0n}/M = 1$ (LSU.O vs. LSU.BH; EB.O vs. EB.C). However, as discussed below and illustrated in Figure 3, estimators of the proportion of true null hypotheses also tend to be conservatively biased, i.e., $h_{0n} \geq h_0$.

### 5.1.2 Type I error control comparison

The Type I error properties of five non-oracle FDR-controlling procedures are illustrated in Figure 1, for a range of nominal FDR levels $\alpha \in [0, 0.20]$.

Overall, procedures tend to be more conservative for weaker correlation structures $\sigma^*$ and smaller proportions of true null hypotheses $h_0/M$, with the resampling-based empirical Bayes procedures (EB.C and EB.QV) and Storey and Tibshirani's [2003] linear step-up procedure LSU.ST remaining closer (in absolute value) to the target nominal Type I error level $\alpha$ (horizontal line) than the linear step-up procedures of Benjamini and colleagues (LSU.BH and LSU.ABH). The LSU.BH and LSU.ABH procedures are conservative over the range of simulation parameters, while the empirical Bayes EB.C and EB.QV procedures and the Storey and Tibshirani [2003] LSU.ST procedure become anti-conservative with stronger correlation structures and higher proportions of true null hypotheses.

As expected, under no correlation, the classical linear step-up procedure LSU.BH of Benjamini and Hochberg [1995] becomes conservative at a rate commensurate with the proportion of true null hypotheses $h_0/M$ (Figure 1, Panels A and D). The adaptive procedures relax this conservatism, with the LSU.ST procedure hovering closest to the target nominal Type I error level $\alpha$.

The results for the empirical microarray correlation structure are similar to those for no correlation, although the empirical Bayes procedures are somewhat less conservative when compared to the linear step-up procedures (Figure 1, Panels B and E).

Under constant, heavy correlation, the procedures of Benjamini and colleagues remain conservative, while the LSU.ST procedure is anti-conservative for small nominal FDR levels $\alpha$ and conservative for less stringent levels. The empirical Bayes procedures display anti-conservative behavior, particularly with a relaxed prior and as the proportion of true null hypotheses increases (Figure 1, Panels C and F).

#### 5.1.3 Power comparison

As argued in Section 4.3.3, fair power comparisons between multiple testing procedures are best performed by benchmarking power against *actual*, rather than *nominal*, Type I error rate (Figure 2).

For the no correlation structure, no method is more powerful outright than the original linear step-up procedure LSU.BH of Benjamini and Hochberg [1995] (Figure 2, Panels A and D). In this case, all gains in power observed in Table 3 for the adaptive linear step-up procedures or resampling-based empirical Bayes procedures (when benchmarking against the *nominal* FDR, $\alpha = 0.05$) are due to these procedures selecting less conservative cut-offs, with higher *actual* FDR, rather than being more powerful *per se*.

Under empirical microarray correlation, the empirical Bayes procedures (EB.C and EB.QV) are as powerful as standard linear step-up procedure LSU.BH, whereas the Storey and Tibshirani [2003] linear step-up LSU.ST procedure is slightly less powerful (Figure 2, Panels B and E).

In the constant, heavy correlation scenario, all procedures loose power relative to the Benjamini and Hochberg [1995] LSU.BH procedure, the largest loss occurring for the Storey and Tibshirani [2003] LSU.ST procedure (Figure 2, Panels C and F).

### 5.2 Estimation of the proportion of true null hypotheses

The properties of six estimators of the proportion $h_0/M$ of true null hypotheses are illustrated in Figure 3, using boxplots over $A = 500$ simulated datasets (Section 4.3.4).

Overall, the estimators tend to be conservatively biased, with decreasing bias for higher proportions of true null hypotheses. Variability tends to increase with increasing correlation levels.

The LSU.ABH estimator, used in the adaptive linear step-up procedure of Benjamini and Hochberg [2000], is consistently the most conservative. The LSU.TST estimator ($\alpha = 0.05, 0.10$), from the two-stage linear step-up procedure of Benjamini et al. [2006], is similar to the LSU.ABH estimator, with a slightly less conservative bias for the higher nominal Type I error level $\alpha = 0.10$ (results not shown). These observations reinforce earlier findings that the procedures of Benjamini and colleagues are capable of maintaining desired levels of Type I error control across a variety of conditions (Figure 1 and Table 3).

As expected, the $q$-value-based empirical Bayes estimators of $h_0/M$ become less conservative as the estimated prior $\pi_{0n}$ is relaxed. These estimators are still conservatively biased, although the lower tails of their distributions dip below the true value $h_0/M$ more frequently as the correlation and/or proportion of true null hypotheses increase.

Although least biased among non-oracle estimators of $h_0/M$, the LSU.ST estimator, from the adaptive linear step-up procedure of Storey and Tibshirani [2003], is by far the most variable. In particular, for a small number of hypotheses $M = 40$ and/or constant, heavy correlation structure $\sigma^*$, the qvalue software returns errors for roughly 1 to 5 percent of simulated datasets, indicating that a negative estimate of the proportion $h_0/M$ is produced. Moreover, as noted by Benjamini et al. [2006], the LSU.ST method can yield estimates that exceed one. Specifically, for $M = 400$ hypotheses and $h_0/M = 0.75$, estimates of $h_0/M$ had to be bounded by one in ten (2.0%) simulated datasets with no correlation among variables, 54 (10.8%) datasets with empirical microarray correlation structure, and 161 (32.2%) datasets with constant, heavy correlation structure.

Estimators of $h_0/M$ are slightly less conservative for a smaller number of hypotheses $M = 40$, but vary between the $M = 40$ and $M = 400$ scenarios by only ca. 1% for empirical Bayes EB.C, EB.O, EB.ABH, and EB.QV estimators, ca. 2–3% for Benjamini and Hochberg [2000] LSU.ABH estimator, and ca. 4–6% for Storey and Tibshirani [2003] LSU.ST estimator (results not shown).

### 5.3 Additional simulation results

Additional simulations were performed to investigate and compare the FDR-controlling procedures of Table 2. The results are summarized below and posted on the website companion.

### 5.3.1 Simulation scenarios

For some combinations of simulation parameters, the Gaussian approximation to the test statistics null distribution seems appropriate for a sample size as low as $n = 30$. However, when $n = 30$, all procedures tend to be anti-conservative for most simulation scenarios. When increasing the sample size to $n = 100$, Central Limit Theorem convergence is observed for all simulation scenarios, except those with constant, heavy correlation. For the latter, Storey and Tibshirani [2003] linear step-up Procedure 3 and resampling-based empirical Bayes Procedure 4 tend to be anti-conservative. Results for the asymptotic or "infinite sample" scenario closely resemble those reported above for a sample size of $n = 250$.

Simulation results for high proportions of true null hypotheses ($h_0/M = 0.95$) are similar to those for $h_0/M = 0.75$, with slightly more anti-conservative behavior for $h_0/M = 0.95$. The empirical Bayes procedures tend to be anti-conservative for the complete null hypothesis ($h_0/M = 1$).

### 5.3.2 Resampling-based empirical Bayes procedures: Test statistic marginal distribution

For the resampling-based empirical Bayes procedures considered in the present article, the guessed sets of true null hypotheses $\mathcal{H}_{0n}^b$ are generated from a distribution $Q_{0n}^{\pi}$ which is based on a common marginal non-parametric mixture model for the test statistics (Section 3.3).

Pilot simulations used an oracle estimator $f_n$ of the common marginal mixture density $f$, obtained by smoothing pooled random vectors from the *unknown* test statistics joint distribution $\mathrm{N}(d_n, \sigma^*)$ (results not shown and van der Laan et al. [2005]). Only minimal anti-conservative behavior is observed for these oracle procedures, under any correlation structure $\sigma^*$ or proportion of true null hypotheses $h_0/M$, including the complete null hypothesis ($h_0/M = 1$).

In the more realistic simulation setting presented here, where a $\mathrm{N}(T_n, \sigma_n^*)$ estimator of the test statistics joint distribution is used to estimate $f$, we witness anti-conservative bias under heavy correlation structures and high proportions of true null hypotheses.

We observe, however, that this anti-conservative bias decreases as the alternative shift parameters $(d_n(m) : m \in \mathcal{H}_1)$ increase (e.g., $d_n(m) = 3, 4$, for $m \in \mathcal{H}_1$). Furthermore, for $h_0/M = 0.95$ and the complete null hypothesis, the anti-conservative behavior does not appear to depend on either the number of hypotheses $M$ (e.g., $M = 2,000$) or the number $B$ of resampled pairs $(T_{0n}^b, \mathcal{H}_{0n}^b)$ (e.g., $B = 20,000$ or $30,000$).

Comparing previous simulation results with those presented here suggests that the anti-conservative behavior of the non-oracle empirical Bayes procedures stems from the increased variability of the estimators of the common marginal density $f$. This variability results in smaller local $q$-values $\pi_{0n}(T_n(m))$ and hence less conservative guessed sets of true null hypotheses $\mathcal{H}_{0n}^b$, which do not as consistently contain the true set of true null hypotheses $\mathcal{H}_0$.

# 6 Discussion

We have proposed resampling-based empirical Bayes procedures for controlling generalized tail probability error rates, $gTP(q, g) = \Pr(g(V_n, S_n) > q)$, and generalized expected value error rates, $gEV(g) = \mathrm{E}[g(V_n, S_n)]$, for arbitrary functions $g(V_n, S_n)$ of the numbers of false positives $V_n$ and true positives $S_n$.

The simulation study of Sections 4 and 5 illustrates the competitive Type I error and power properties of the resampling-based empirical Bayes procedures when compared to widely-used FDR-controlling linear step-up procedures. These results for FDR control are consistent with previous results for TPPFP control in the original article of van der Laan et al. [2005].

For a variety of testing scenarios, the resampling-based empirical Bayes approach exhibits Type I error and power properties intermediate between those of the linear step-up procedures of Benjamini and colleagues and Storey and colleagues (Figures 1 and 2, Table 3). Specifically, empirical Bayes procedures control the false discovery rate less conservatively than the classical Benjamini and Hochberg [1995] procedure and adaptive versions thereof [Benjamini and Hochberg, 2000, Benjamini et al., 2006], with, as for the Storey and Tibshirani [2003] procedure, the risk of anti-conservative behavior for heavy correlation structures. The

empirical Bayes procedures tend to be more powerful than the so-called $q$-value procedure of Storey and Tibshirani [2003], particularly for microarray-like correlation structures, which have been viewed in the literature as exhibiting potentially weak dependence or dependence in finite blocks [Storey, 2002, Storey and Tibshirani, 2003, Storey et al., 2004].

The simulation study indicates that gains in power can be achieved by the empirical Bayes procedures when using a data-adaptive prior $\pi_{0n}$ to estimate the local $q$-values $\pi_{0n}(T_n(m))$. The decision to deviate from the most conservative prior ($\pi_{0n} = 1$), however, should be guided by prior knowledge regarding the proportion of true null hypotheses as well as the level of correlation between test statistics. In many applications, the anti-conservative bias occurring in extreme simulation conditions will either not be present or may be of minor practical significance. Diagnostic tests suggest that the density ratio $f_0/f$ is a critical quantity to further investigate regarding proper Type I error control.

The local $q$-values, used to generate the random guessed sets of true null hypotheses in the empirical Bayes procedures, provide estimators of the proportion of true null hypotheses that tend to be less conservatively biased than the Benjamini and Hochberg [2000] estimator and less variable than the Storey and Tibshirani [2003] estimator.

Of course an issue in presenting any resampling-based procedure is the trade-off between gains in accuracy and extra computational cost. As shown in this study, for testing scenarios with no correlation and a large proportion of true null hypotheses, the empirical Bayes procedures do not improve upon the linear step-up methods of Benjamini and colleagues. If Type I error control is the primary concern, then these simpler procedures are probably the better choice. However, when the goal is to reject a larger number of hypotheses, while still maintaining adequate Type I error control, then the empirical Bayes procedures are strong contenders under various levels of correlation.

We wish to stress the benefits and generality of the proposed resampling-based empirical Bayes methodology.

- It can be used to control a broad class of Type I error rates, defined as tail probabilities and expected values of arbitrary functions $g(V_n, S_n)$ of the numbers of false positives $V_n$ and true positives $S_n$. As discussed in Dudoit and van der Laan [2007, Section 7.8], the approach can be further extended to control other parameters of the distribution of functions $g(V_n, S_n)$. Researchers can therefore select from a wide library of Type I error rates for subject-matter-relevant measures of false positives and control these error rates at little additional computational cost, using the same resampled pairs $(T_{0n}^b, \mathcal{H}_{0n}^b)$.

- Unlike most MTPs controlling the proportion of false positives, it is based on a test statistics joint null distribution and provides Type I error control in testing problems involving general data generating distributions, with arbitrary dependence structures among variables.

- Gains in power are achieved by deriving rejection regions based on guessed sets of true null hypotheses and null test statistics randomly sampled from joint distributions that account for the dependence structure of the data.

- It is modular and can be applied to any distribution pair $(Q_{0n}, Q_{0n}^{\mathcal{H}})$ for the null test statistics and guessed sets of true null hypotheses, i.e., the common marginal non-parametric mixture model of Section 3.3 is only one among many reasonable working models that does not assume independence of the test statistics.

In summary, the Type I error and power trade-off achieved by the resampling-based empirical Bayes procedures under a variety of testing scenarios (with varying degrees of correlation) allows this approach to be competitive with or outperform the Storey and Tibshirani [2003] linear step-up procedure, as an alternative to the classical Benjamini and Hochberg [1995] procedure.

Ongoing efforts include further investigating the distribution $Q_{0n}^{\mathcal{H}}$ for the guessed sets of true null hypotheses, in order to guarantee proper Type I error control by the empirical Bayes procedures for a wider range of testing scenarios. In particular, we are interested in developing less biased estimators of the density ratio $f_0/f$ in the local $q$-value function.

We are also considering improvements to the estimator of the gTP and gEV error rates in Equation (39), which is used to select test statistic cut-offs that satisfy the Type I error constraint of Equation (40). In the common-cut-off case and for testing scenarios with a large proportion of true null hypotheses $h_0/M$, we have noted that the current estimator

$$\hat{\theta}_n(\gamma^{(M)}) = \frac{1}{B}\sum_{b=1}^{B} \widetilde{G}(\gamma^{(M)}; \mathcal{H}_{0n}^b, T_{0n}^b, T_n),$$

of the Type I error function $\theta_n(\gamma^{(M)}) = \mathrm{E}[\widetilde{G}(\gamma^{(M)}; \mathcal{H}_0, T_n, T_n)]$, can be anti-conservatively biased and variable, i.e., non-monotonic in the common cut-off $\gamma$. This is especially problematic for the complete null hypothesis ($h_0/M = 1$), where the false discovery rate coincides with the family-wise error rate and one would therefore like estimators of these two error rates to be nearly equal and monotonic in the common cut-off $\gamma$. Smoothing or enforcing monotonicity constraints on the estimator $\hat{\theta}_n(\gamma^{(M)})$ may alleviate the anti-conservative bias.

Finally, we are implementing the proposed multiple testing procedures in the R package multtest, released as part of the Bioconductor Project.

## Software and website companion

The multiple testing procedures proposed in Dudoit and van der Laan [2007] and related articles [Birkner et al., 2005, Dudoit et al., 2004a,b, van der Laan et al., 2004a,b, 2005, van der Laan and Hubbard, 2006, Pollard et al., 2005a,b, Pollard and van der Laan, 2004] are implemented in the R package multtest, released as part of the Bioconductor Project, an open-source software project for the analysis of biomedical and genomic data (Dudoit and van der Laan [2007, Section 13.1]; Pollard et al. [2005b]; `www.bioconductor.org`).

The simulation study was performed in R (Release 2.5.1), using the following packages: multtest (Version 1.16.0), qvalue (Version 1.1), and golubEsets (Version 1.4.3).

The website companion for this article provides additional tables, figures, references, and software: `www.stat.berkeley.edu/~sandrine`.

## Acknowledgement

We are most grateful to Alan E. Hubbard (Division of Biostatistics, UC Berkeley) and Merrill D. Birkner (Genentech, Inc.) for valuable discussions on multiple testing methodology and software.

## References

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

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

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

M. D. Birkner, K. S. Pollard, M. J. van der Laan, and S. Dudoit. Multiple testing procedures and applications to genomics. Technical Report 168, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7360, 2005. URL `www.bepress.com/ucbbiostat/paper168`.

S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer, New York, 2007.

S. Dudoit, M. J. van der Laan, and M. D. Birkner. Multiple testing procedures for controlling tail probability error rates. Technical Report 166, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7360, 2004a. URL `www.bepress.com/ucbbiostat/paper166`.

S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, 2004b. URL `www.bepress.com/sagmb/vol3/iss1/art13`.

C. R. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061, 2004a.

C. R. Genovese and L. Wasserman. Exceedance control of the false discovery proportion. Technical Report 807, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, July 2004b. URL `www.stat.cmu.edu/tr/tr807/tr807.html`.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

E. L. Lehmann and J. P. Romano. Generalizations of the familywise error rate. *Annals of Statistics*, 33(3): 1138–1154, 2005.

K. S. Pollard and M. J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125(1–2):85–100, 2004.

K. S. Pollard, M. D. Birkner, M. J. van der Laan, and S. Dudoit. Test statistics null distributions in multiple testing: Simulation studies and applications to genomics. *Journal de la Société Française de Statistique*, 146(1–2):77–115, 2005a. URL `www.stat.berkeley.edu/~sandrine/Docs/Papers/SFdS05/SFdS.html`. Numéro double spécial *Statistique et Biopuces*.

K. S. Pollard, S. Dudoit, and M. J. van der Laan. Multiple testing procedures: The multtest package and applications to genomics. In R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 15, pages 249–271. Springer, New York, 2005b. URL `www.bioconductor.org/pubs/docs/mogr,www.bepress.com/ucbbiostat/paper164`.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64(3):479–498, 2002.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.*, 100 (16):9440–9445, 2003.

J. D. Storey, J. E. Taylor, and D. O. Siegmund. 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(1):187–205, 2004.

M. J. van der Laan and A. E. Hubbard. Quantile-function based null distribution in resampling based multiple testing. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 14, 2006. URL `www.bepress.com/sagmb/vol5/iss1/art14`.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 14, 2004a. URL `www.bepress.com/sagmb/vol3/iss1/art14`.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, 2004b. URL `www.bepress.com/sagmb/vol3/iss1/art15`.

M. J. van der Laan, M. D. Birkner, and A. E. Hubbard. Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 29, 2005. URL `www.bepress.com/sagmb/vol4/iss1/art29`.

Table 1: *Type I and Type II errors in multiple hypothesis testing.* This table summarizes the different types of decisions and errors in multiple hypothesis testing. The number of rejected null hypotheses is $R_n = |\mathcal{R}_n|$, the number of Type I errors or false positives is $V_n = |\mathcal{R}_n \cap \mathcal{H}_0|$, the number of Type II errors or false negatives is $U_n = |\mathcal{R}_n^c \cap \mathcal{H}_1|$, the number of true negatives is $W_n = |\mathcal{R}_n^c \cap \mathcal{H}_0|$, and the number of true positives is $S_n = |\mathcal{R}_n \cap \mathcal{H}_1|$. Cells corresponding to errors are enclosed in boxes.

| | | **Null hypotheses** | | |
|---|---|---|---|---|
| | | **Non-rejected,** $\mathcal{R}_n^c$ | **Rejected,** $\mathcal{R}_n$ | |
| **Null hypotheses** | **True,** $\mathcal{H}_0$ | $W_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_0\rvert$ | $\boxed{V_n = \lvert\mathcal{R}_n \cap \mathcal{H}_0\rvert}$ | $h_0$ |
| | **False,** $\mathcal{H}_1$ | $\boxed{U_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_1\rvert}$ | $S_n = \lvert\mathcal{R}_n \cap \mathcal{H}_1\rvert$ | $h_1$ |
| | | $M - R_n$ | $R_n$ | $M$ |

Table 2: *Simulation study: Multiple testing procedures.* This table summarizes the FDR-controlling procedures examined in the simulation study of Sections 4 and 5. The adaptive linear step-up procedures are based on generic Procedure 2, with specified estimators $h_{0n}$ of the number of true null hypotheses $h_0$ (Sections 2 and 4.2.4). The resampling-based empirical Bayes procedures are based on Procedure 4, with specified estimators $\pi_{0n}$ of the true null hypothesis prior probability $\pi_0$ (Sections 3 and 4.2.5).

| **LSU: Linear step-up procedures** | | |
|---|---|---|
| | $h_{0n}$ | |
| LSU.BH | $M$ | Conservative: Benjamini and Hochberg [1995]; Procedure 1 |
| LSU.O | $h_0$ | Oracle |
| LSU.ABH | $h_{0n}^{ABH}$ | Adaptive: Benjamini and Hochberg [2000]; Equation (16) |
| LSU.TST | $h_{0n}^{TST}(\alpha)$ | Adaptive two-stage: Benjamini et al. [2006]; Equation (18), $\alpha = 0.05, 0.10$ |
| LSU.ST | $h_{0n}^{ST}(\lambda)$ | Adaptive: Storey and Tibshirani [2003]; Procedure 3, Equation (17) |
| **EB: Resampling-based empirical Bayes procedures** | | |
| | $\pi_{0n}$ | |
| EB.C | 1 | Conservative |
| EB.O | $h_0/M$ | Oracle |
| EB.ABH | $h_{0n}^{ABH}/M$ | Adaptive: Benjamini and Hochberg [2000]; Equation (16) |
| EB.QV | $h_{0n}^{QV}/M$ | Adaptive $q$-value-based: Equation (49) |

Table 3: *Simulation study: Type I error control and power comparison.* This table reports the actual Type I error rate $FDR(\alpha)$ and the average power $AvgPwr(\alpha)$ for FDR-controlling procedures summarized in Table 2, applied with nominal FDR level $\alpha = 0.05$. Results correspond to the following simulation parameters: sample size $n = 250$; number of null hypotheses $M = 40, 400$; proportion of true null hypotheses $h_0/M = 0.50, 0.75$; common alternative shift parameter $d_n(m) = 2$, $m \in \mathcal{H}_1$; correlation structure $\sigma^*$="No correlation", "Empirical microarray correlation", "Constant correlation". Increasingly anti-conservative behavior, i.e., increasingly negative differences $\alpha - FDR(\alpha)$ between the target *nominal* Type I error level $\alpha = 0.05$ and the *actual* Type I error rate $FDR(\alpha)$, is indicated by the following colors: yellow for $FDR(\alpha) \in (0.050, 0.060]$, orange for $FDR(\alpha) \in (0.060, 0.070]$, and red for $FDR(\alpha) \in (0.070, 1.000]$. Increasingly conservative behavior is indicated by the following colors: green for $FDR(\alpha) \in [0.040, 0.050)$, blue for $FDR(\alpha) \in [0.030, 0.040)$, and purple for $FDR(\alpha) \in [0, 0.030)$.

| | $M = 40$ | | | | $M = 400$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $h_0/M = 0.50$ | | $h_0/M = 0.75$ | | $h_0/M = 0.50$ | | $h_0/M = 0.75$ | |
| | $FDR$ | $AvgPwr$ | $FDR$ | $AvgPwr$ | $FDR$ | $AvgPwr$ | $FDR$ | $AvgPwr$ |
| $\sigma^*$: No correlation | | | | | | | | |
| LSU.BH | 0.022 | 0.257 | 0.041 | 0.185 | 0.028 | 0.229 | 0.042 | 0.135 |
| LSU.O | 0.048 | 0.393 | 0.057 | 0.227 | 0.052 | 0.371 | 0.055 | 0.173 |
| LSU.ABH | 0.034 | 0.330 | 0.050 | 0.208 | 0.035 | 0.278 | 0.046 | 0.146 |
| LSU.TST | 0.024 | 0.278 | 0.042 | 0.192 | 0.031 | 0.250 | 0.043 | 0.139 |
| LSU.ST | 0.077 | 0.429 | 0.077 | 0.256 | 0.048 | 0.349 | 0.054 | 0.167 |
| EB.C | 0.038 | 0.344 | 0.064 | 0.242 | 0.038 | 0.300 | 0.052 | 0.162 |
| EB.O | 0.049 | 0.402 | 0.068 | 0.254 | 0.049 | 0.358 | 0.056 | 0.174 |
| EB.ABH | 0.046 | 0.374 | 0.064 | 0.248 | 0.041 | 0.317 | 0.053 | 0.166 |
| EB.QV | 0.046 | 0.375 | 0.068 | 0.251 | 0.043 | 0.326 | 0.055 | 0.170 |
| $\sigma^*$: Empirical microarray correlation | | | | | | | | |
| LSU.BH | 0.022 | 0.243 | 0.035 | 0.198 | 0.023 | 0.228 | 0.032 | 0.159 |
| LSU.O | 0.043 | 0.375 | 0.043 | 0.237 | 0.047 | 0.366 | 0.046 | 0.193 |
| LSU.ABH | 0.039 | 0.318 | 0.044 | 0.225 | 0.031 | 0.283 | 0.038 | 0.175 |
| LSU.TST | 0.027 | 0.268 | 0.039 | 0.207 | 0.027 | 0.254 | 0.035 | 0.166 |
| LSU.ST | 0.070 | 0.396 | 0.070 | 0.275 | 0.048 | 0.348 | 0.054 | 0.197 |
| EB.C | 0.038 | 0.341 | 0.058 | 0.266 | 0.038 | 0.323 | 0.055 | 0.211 |
| EB.O | 0.048 | 0.397 | 0.063 | 0.277 | 0.050 | 0.379 | 0.060 | 0.222 |
| EB.ABH | 0.045 | 0.371 | 0.062 | 0.275 | 0.042 | 0.343 | 0.057 | 0.216 |
| EB.QV | 0.043 | 0.374 | 0.064 | 0.276 | 0.044 | 0.353 | 0.059 | 0.221 |
| $\sigma^*$: Constant correlation | | | | | | | | |
| LSU.BH | 0.021 | 0.267 | 0.031 | 0.182 | 0.027 | 0.241 | 0.029 | 0.175 |
| LSU.O | 0.046 | 0.378 | 0.038 | 0.216 | 0.052 | 0.344 | 0.037 | 0.204 |
| LSU.ABH | 0.035 | 0.332 | 0.045 | 0.201 | 0.038 | 0.297 | 0.040 | 0.190 |
| LSU.TST | 0.029 | 0.295 | 0.034 | 0.188 | 0.035 | 0.271 | 0.034 | 0.184 |
| LSU.ST | 0.062 | 0.348 | 0.073 | 0.225 | 0.057 | 0.316 | 0.072 | 0.202 |
| EB.C | 0.052 | 0.382 | 0.081 | 0.269 | 0.070 | 0.346 | 0.070 | 0.251 |
| EB.O | 0.078 | 0.428 | 0.086 | 0.281 | 0.087 | 0.384 | 0.076 | 0.261 |
| EB.ABH | 0.067 | 0.404 | 0.089 | 0.276 | 0.078 | 0.363 | 0.078 | 0.256 |
| EB.QV | 0.064 | 0.408 | 0.091 | 0.379 | 0.081 | 0.370 | 0.082 | 0.263 |

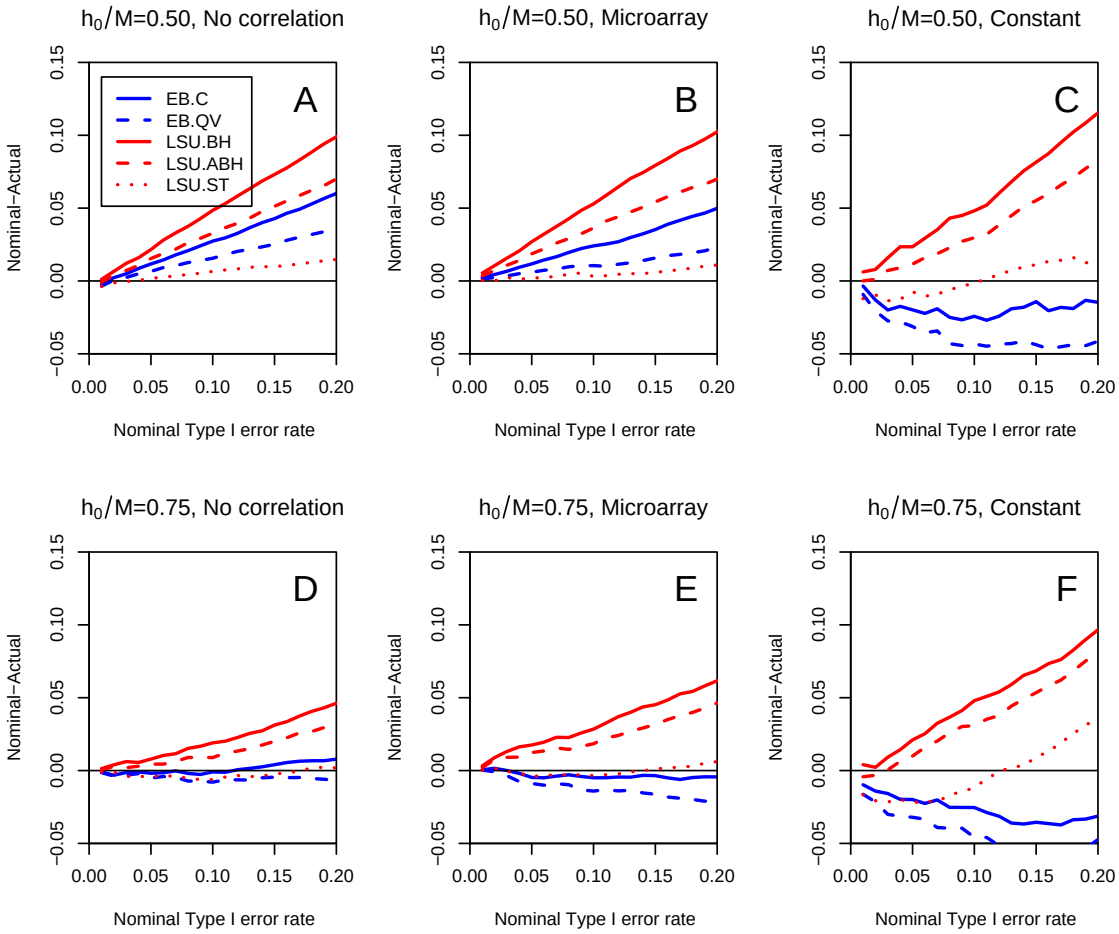


Figure 1: *Simulation study: Type I error control comparison.* Plots of differences $\alpha - FDR(\alpha)$ between nominal and actual Type I error rates vs. nominal Type I error level $\alpha \in [0, 0.20]$, for FDR-controlling procedures EB.C, EB.QV, LSU.BH, LSU.ABH, and LSU.ST, summarized in Table 2. Results correspond to the following simulation parameters: sample size $n = 250$; number of null hypotheses $M = 400$; common alternative shift parameter $d_n(m) = 2$, $m \in \mathcal{H}_1$; proportion of true null hypotheses ($h_0/M = 0.50, 0.75$) and correlation structure ($\sigma^* =$ "No correlation", "Microarray", "Constant") indicated in the panel titles. Positive (negative) differences indicate (anti-)conservative behavior.

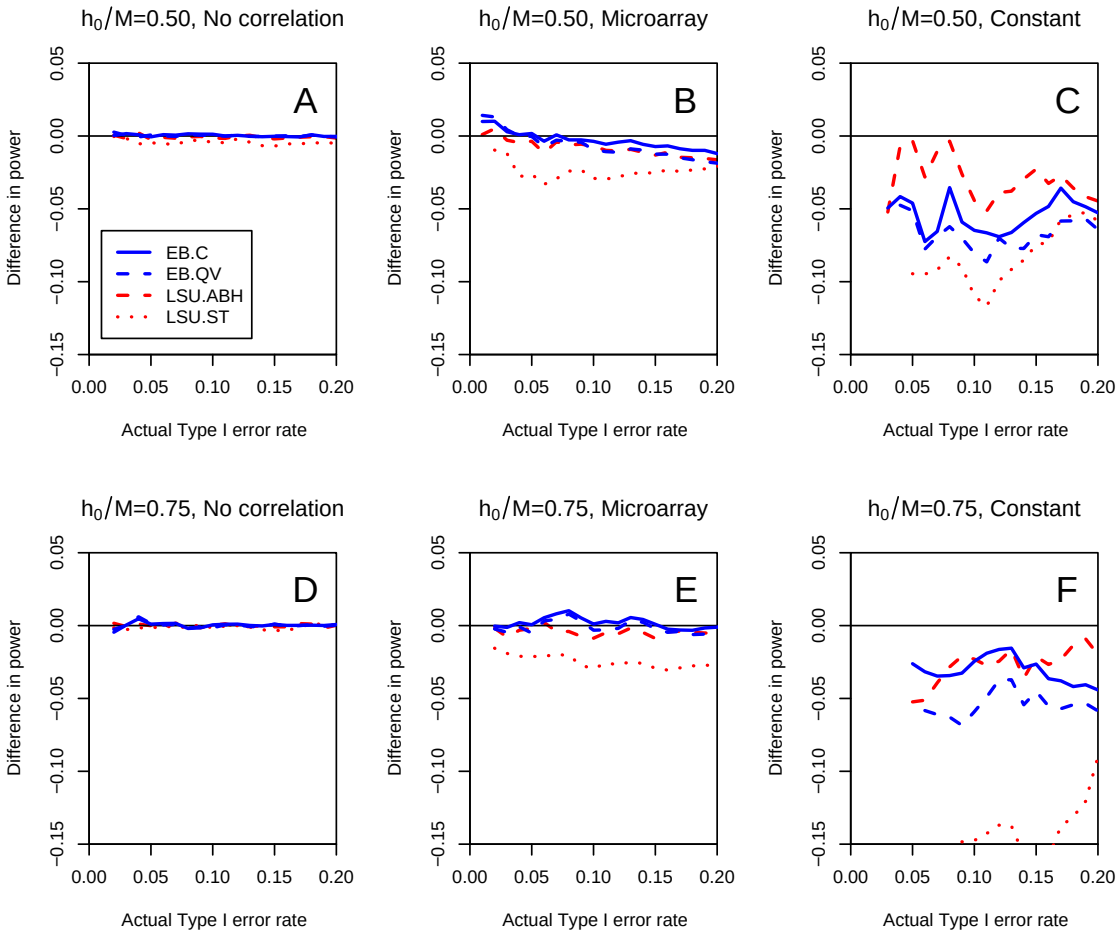


Figure 2: *Simulation study: Power comparison.* Plots of differences in power vs. actual Type I error rate, for FDR-controlling procedures EB.C, EB.QV, LSU.BH, LSU.ABH, and LSU.ST, summarized in Table 2, using LSU.BH as baseline. Results correspond to the following simulation parameters: sample size $n = 250$; number of null hypotheses $M = 400$; common alternative shift parameter $d_n(m) = 2$, $m \in \mathcal{H}_1$; proportion of true null hypotheses ($h_0/M = 0.50, 0.75$) and correlation structure ($\sigma^*$ = "No correlation", "Microarray", "Constant") indicated in the panel titles. Positive (negative) differences indicate greater (lower) power than the baseline LSU.BH procedure.

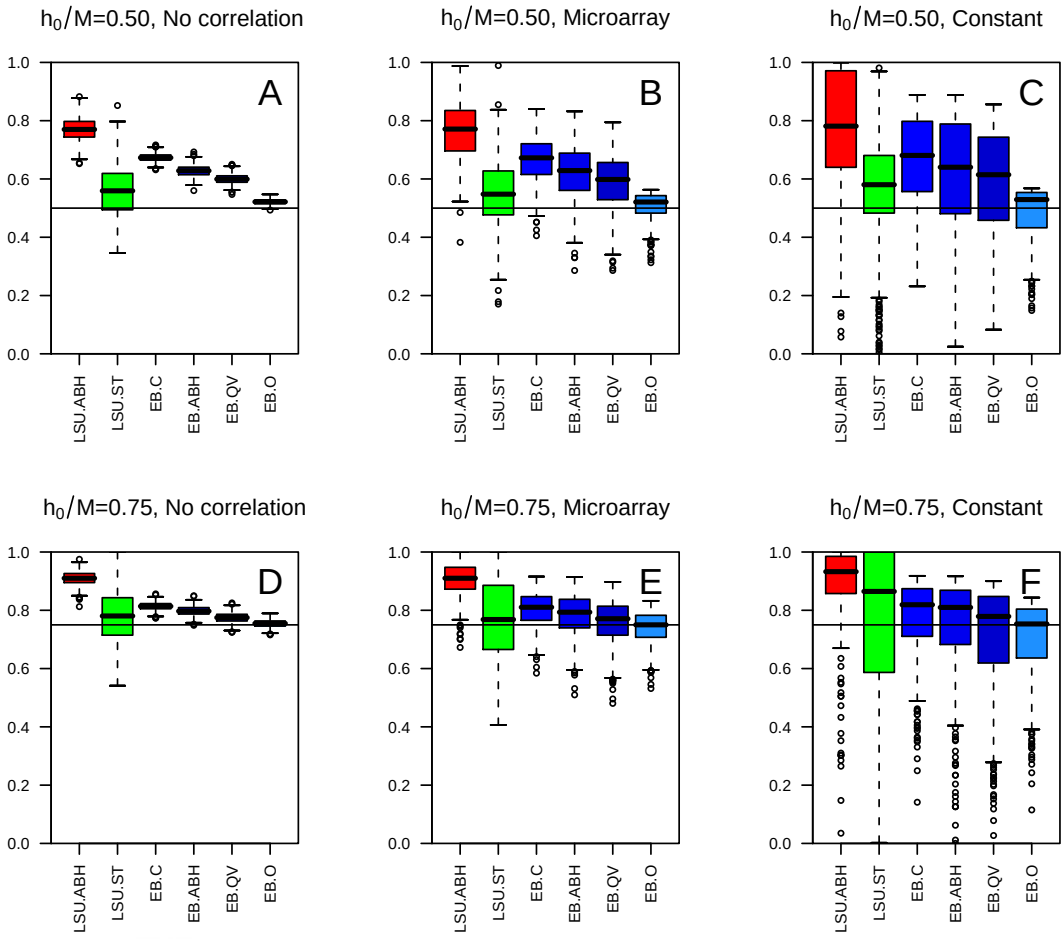


Figure 3: *Simulation study: Estimation of the proportion of true null hypotheses.* Boxplots of estimates $h_{0n}/M$ of the proportion of true null hypotheses $h_0/M$ (over $A = 500$ simulated datasets), from FDR-controlling procedures LSU.ABH, LSU.ST, EB.C, EB.ABH, EB.QV, and EB.O, as summarized in Section 4.3.4. Results correspond to the following simulation parameters: sample size $n = 250$; number of null hypotheses $M = 400$; common alternative shift parameter $d_n(m) = 2$, $m \in \mathcal{H}_1$; proportion of true null hypotheses ($h_0/M = 0.50, 0.75$) and correlation structure ($\sigma^*$ = "No correlation", "Microarray", "Constant") indicated in the panel titles. The horizontal line indicates the true, unknown proportion of true null hypotheses, $h_0/M$.