FDR CONTROL WITH ADAPTIVE PROCEDURES AND FDR MONOTONICITY

By Amit Zeisel[*,‡], Or Zuk[*,§] and Eytan Domany[†,‡]

Weizmann Institute[‡] and Broad Institute[§]

The steep rise in availability and usage of high-throughput technologies in biology brought with it a clear need for methods to control the False Discovery Rate (FDR) in multiple tests. Benjamini and Hochberg (BH) introduced in 1995 a simple procedure and proved that it provided a bound on the expected value, $FDR \leq q$. Since then, many authors tried to improve the BH bound, with one approach being designing *adaptive* procedures, which aim at estimating the number of true null hypothesis in order to get a better FDR bound. Our two main rigorous results are: (i) a theorem that provides a bound on the FDR for adaptive procedures that use any estimator for the number of true hypotheses (m_0), (ii) a theorem that proves a monotonicity property of general BH-like procedures, both for the case where the hypothesis are independent. We also propose two improved procedures for which we prove FDR control for the independent case, and demonstrate their advantages over several available bounds, on simulated data and on a large number of gene expression datasets. Both applications are simple and involve a similar amount of computation as the original BH procedure. We compare the performance of our proposed procedures with BH and other procedures and find that in most cases we get more power for the same level of statistical significance.

1. Introduction. The main goal of statistical comparisons (tests) is to calculate the level of statistical significance at which a given null hypothesis is rejected on the basis of available data. Researchers use this tool in order to present their findings and support their conclusions. Uncontrolled application of single inference procedures in a multiple comparison setting can cause a high false positive rate. Special multiple comparison procedures are used in order to control the probability of committing such a type I error in families of comparisons.

The need for improved control over the multiplicity effect in biological experiments became acute in the nineties, when the amount of data that could be measured and stored increased thousands fold. Many new experimental techniques, which allowed taking a large number of measurements simultaneously were developed, along with improved data acquisition and storage capabilities.

For example, in the case of gene expression microarray measurements, a typical aim is to identify the genes whose expression levels differentiate between healthy (type A) and diseased (type B) subjects. Genes are tested one by one for differential expression; the formal way to do this is by posing several thousand null hypotheses. A null hypothesis states that a particular variable (e.g. expression level of gene i) is sampled from the same distribution for both types A, B; one is interested in identifying variables (genes) for which the null hypothesis is rejected (i.e. genes whose expression *does* differentiate between types A, B). Such a finding is referred to as a *discovery*. Denote by m the total number of hypotheses (e.g. the number of genes whose expression levels were measured), and assume that the null hypothesis is true for m_0 out of the m (i.e. m_0 genes' expression levels do not differentiate the two types). For $m_1 = m - m_0$ the null hypothesis is false (the expression levels of

types A and B are sampled from different distributions). A statistical test is performed independently for each variable, producing a p-value p_i, $i = 1, 2, ...m$. On the basis of some thresholding operation on the p_i's, the null hypothesis is rejected for R tests. The decision to reject (or not) can be correct or false; When the null hypothesis is rejected for one of the m_0 variables for which it is actually true, we have a "false discovery" (type I error). Table 1 presents the possible categories to which rejected and non-rejected hypotheses can belong, and the number of hypotheses in each category.

"ground truth"	non-rejected hypotheses	rejected hypotheses	total
null hypothesis is true	U	V	m_0
null hypothesis is false	T	S	m_1
total	$m - R$	R	m

TABLE 1

Numbers of true/false decisions taken when testing m null hypotheses

Out of the R rejected hypotheses the fraction V/R is falsely rejected. The expected value of this fraction was termed by [4], (referred to as BH95) as the False Discovery Rate (FDR),

$$(1.1) \qquad FDR \equiv E\left(\frac{V}{R}\middle| R > 0\right) \Pr(R > 0) \equiv E\left(\frac{V}{R^+}\right)$$

where here and later in the paper the term $R^+ \equiv \max(R, 1)$ is used for brevity. It is required since V/R is undefined when $R = 0$ and thus this case should be treated separately - we follow [4] and replace V/R by 0 in this case. The original BH95 procedure to control the FDR is given as follows:

1. Denote by q the desired level, $0 < q \leq 1$, of the FDR and define the following set of constants:

$$(1.2) \qquad \alpha_i = \frac{iq}{m}, \quad i = 1, 2, ..., m$$

2. Sort the p-values p_i and re-label the hypotheses accordingly, $p_{(1)} \leq p_{(2)} \leq .... \leq p_{(m)}$, such that (i) is the index of the hypothesis with the i-th smallest p-value.
3. Identify R as

$$(1.3) \qquad R = \max\left\{i : p_{(i)} \leq \alpha_i\right\}$$

If no such $R \geq 1$ exists, no hypothesis is rejected; otherwise reject all R hypotheses $(i) = 1, 2, ...R$.

This procedure has a simple graphical implementation, depicted in Fig. 1. It is referred to in BH95 as *"step-up"*; in general there could be more than one intersection point (of the $p_{(i)}$ and α_i lines), in which case the step-up procedure identifies the intersection with the largest p-value as R, whereas the more conservative *"step-down"* procedure identifies the lowest one, replacing eq. (1.3) by

$$(1.4) \qquad R = \min\left\{i : p_{(i)} > \alpha_i\right\} - 1$$

The bound

$$(1.5) \qquad FDR = E\left(\frac{V}{R^+}\right) \leq \frac{m_0}{m}q$$

was proved by BH95 for independent tests, and by [6] for a certain type of 'positive dependency' called PRDS (*Positive Regression Dependency on each one from a Subset*). The value of m_0 is unknown to the researcher, but since $m_0 \leq m$, this procedure leads to the bound:

$$(1.6) \qquad FDR = E\left(\frac{V}{R^+}\right) \leq \frac{m_0}{m}q \leq q$$
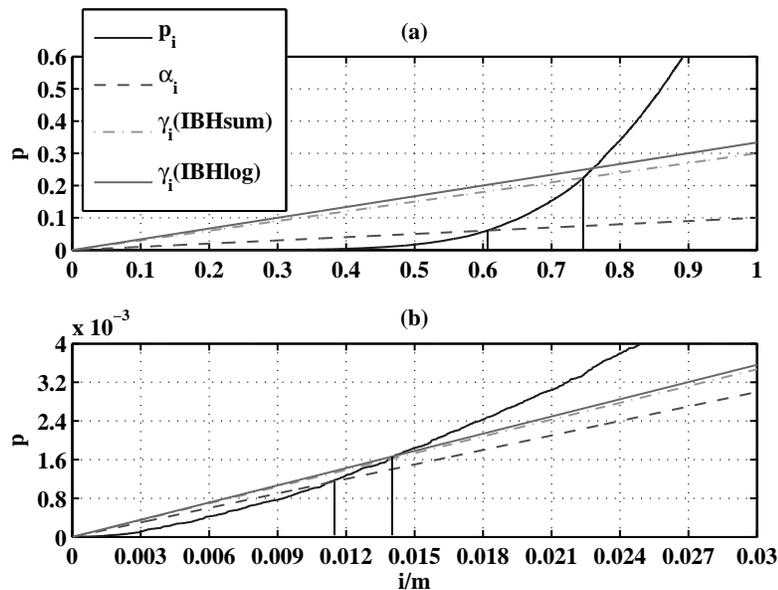
FIG 1. *Typical examples for the use of the BH95 and our IBH procedures, for a desired FDR value of $q = 0.1$. The sorted p-values (solid line), the α_i of eq. (1.2) (dashed line) and the γ_i from eq. (2.1) (dot-dashed line for IBHsum and solid light for IBHlog) are shown, for (a) leukemia data from [1] and (b) breast cancer data from [18]. As indicated in (a), the number of rejections is determined for each procedure by locating the (maximal) value $i = R$ at which the corresponding lines intersect $p_{(i)}$, (the vertical lines mark the intersection point between the lines).*

Clearly, had we known m_0, we could have defined a different set of constants (compare to eq. (1.2))

$$\alpha_i' = \frac{iq}{m_0} \tag{1.7}$$

and defining

$$R' = max\left\{ i : p_{(i)} \le \alpha_i' \right\} \tag{1.8}$$

would have obtained a larger number $R' \ge R$ of rejected hypotheses (still with $FDR \le q$), than the number R given by the original BH95 procedure, which used m as an upper-bound on m_0. This procedure, based on knowledge of m_0, is called "oracle" (ORC), see [11]. Subsequently various improved (also called 'adaptive') procedures were proposed, based on the idea of estimating the unknown m_0 in order to get a more accurate handle on the FDR. These procedures can be divided into two major classes:

1. Procedures for local FDR *estimation*: This approach, previously suggested and applied by [30],[24],[20], can be used when one has an estimator $\hat{m}_0$ of m_0, that satisfies:

$$m_0 \le E(\hat{m}_0) \le m \tag{1.9}$$

In procedures of this type one can write the local FDR (lFDR) estimate as (see [20]):

$$t_{(i)} = \frac{\hat{v}(p_{(i)})}{\hat{F}(p_{(i)})} \tag{1.10}$$

where, $p_{(i)}$ is the ordered p-value, $\hat{v}(\alpha)$ is the estimator for the type I errors (in the rejection region), and $\hat{F}(\alpha)$ is the estimator for the probability $Pr(p \le \alpha)$ (often estimated by $R(p_{(i)})/m$).

Since for $\hat{v}(\alpha)$ most methods use:

$$\hat{v}(\alpha) = \alpha \frac{\hat{m}_0}{m} \tag{1.11}$$

any estimator that satisfies eq. (1.9) can provide an improved estimator and yield:

$$\alpha \frac{m_0}{m} \le E(\hat{v}(\alpha)) = \alpha \frac{E(\hat{m}_0)}{m} \le \alpha \tag{1.12}$$

This approach is the preferred one in many biological contexts, when the investigator wishes to control R, the number of discoveries made (e.g. differentiating genes to be used in further experiments).

2. Procedures for FDR *control*: In this approach, one wishes to control the FDR at a preset level q. This is achieved by defining $\gamma_i = iq/\hat{m}_0$ to be used in the same way as α_i and α'_i (see eq. (1.2) and eq. (1.7)), leading typically to a larger number R' of rejected hypotheses (compared to BH95), with the FDR still being bound by the desired value q. The advantage of this procedure (presented in Sec. 5) is that one retains control of q, the desired level of FDR.

We present in this paper two estimators, $\hat{m}_0$ and $\tilde{m}_0$, that satisfy eq. (1.9), and hence can be used trivially for FDR estimation. As opposed to FDR estimation, proving *control* of the FDR is far more involved, and constitutes a significant portion of this paper. We provide two new proven procedures for control of the FDR. We first prove control for these procedures when employed in a step-up manner. Then, by using a new general monotonicity result for the FDR which we derive, we show that the step-down versions of our procedures also control the FDR. Designing better procedures for FDR estimation and control has drawn a great deal of attention in recent years, as is demonstrated by the abundance of proposed procedures and many theoretical and experimental papers. However, as far as we know only for a few such procedures has control of the FDR been rigorously established: the original BH95 procedure [4], the two-stage and multiple-stage adaptive BH procedures [5] (we refer to the latter as BKY), and Storey's procedure [24] (referred to as STS). All these procedures (except, of course, BH95) claim to give improved power over BH95. All are derived from a better estimation of m_0. Almost all proofs for FDR control assume independence of the p-values (with the notable exception [6]). Thus, far less is known about the behavior of FDR procedures under dependency, where most of our understanding comes from simulation studies. In addition, the FDR, by its definition (eq. (1)), is an *expected* value. However, the fraction of the false discoveries V/R^+ is a *random variable*. While the mean value (FDR) was extensively studied, far less attention has been devoted in the literature to the behavior of this random variable, its variance and entire distribution. We therefore perform simulations whose purposes are: a. To study the behavior of the various procedures under dependence, where analytical results are harder to establish, and b. study the distribution of the fraction of false rejections (V/R^+), which has implications on possible violation of the bound for a particular realization. Our simulations provide a comparison of our new procedures to the known ones mentioned above and we show that our new procedures compare favorably in most cases of interest. We analyze simulated and real data, and show that for both the new procedures almost always reject more hypotheses than BH95, while maintaining control even under dependence, and we therefore refer to these procedures as 'Improved BH' (IBH). The real data which we use is gene expression data obtained from various cancer studies, and we show that our new procedures allow rejection of more hypotheses at a given confidence level and thus increase discovery power.

A Matlab package implementing our proposed procedures, including examples and datasets analyzed in the paper is provided in the supplementary information and in the following URL:
http://www.broadinstitute.org/~orzuk/matlab/libs/stats/fdr/matlab_fdr_utils.html

2. Preliminaries and theorem on control. In this section we present a theorem which provides a general way to build an improved bound for controlling the FDR using an estimator for m_0. Two examples of practical implementation of the theorem lead to useful procedures described in the next section. The working assumptions we use here is that the p-values are independent. The theorem is not proven for dependent variables but our simulations indicate that in most cases we do control the FDR even under dependence (see Sec. 5). Our first step is defining mathematically a family of estimators $\hat{m}_0$ for m_0. We define a general modified BH procedure, in which any one of these estimators is used by replacing m in the original BH95 procedure (see eqs. (1.2,1.3)) by $\hat{m}_0$. Throughout this section and the rest of the paper we denote for convenience $p_{i..j} \equiv p_i, .., p_j$. We also denote $\vec{p} = (p_1, .., p_m)$ the vector of all p-values.

DEFINITION 2.1. *An estimator for m_0 is a family of functions $\hat{m}_0 \equiv \hat{m}_0^{(m)} : [0,1]^m \to \mathbb{R}$, $\hat{m}_0 \equiv \hat{m}_0(\vec{p})$. We usually omit the index $^{(m)}$ as it is obvious from the context. We say that $\hat{m}_0$ is a monotonic estimator if it satisfies:*

1. *$\hat{m}_0^{(m)}(p_1, .., p_i, .., p_m) \geq \hat{m}_0^{(m)}(p_1, .., p_i', .., p_m), \quad \forall\, p_i \geq p_i', \; i = 1, 2, , , m, \; m \geq 1$*
2. *$\hat{m}_0^{(m)}(p_1, .., p_i, .., p_m) \geq \hat{m}_0^{(m-1)}(p_1, .., p_{i-1}, p_{i+1}, .., p_m), \quad \forall\, i = 1, 2, , , m, \; m \geq 2$*

DEFINITION 2.2. *Assume w.l.o.g that we have m hypotheses the first m_0 of which are null. Let $\vec{p} = (p_1, .., p_m)$ be the corresponding p-values. The modified step-up BH procedure with estimator $\hat{m}_0$ is defined as follows:*

1. *Compute $\hat{m}_0 \equiv \hat{m}_0(\vec{p})$.*
2. *For each i define:*

$$(2.1) \qquad\qquad \gamma_i = \frac{iq}{\hat{m}_0}$$

3. *Order the p-values in an increasing order: $p_{(1)} \leq .. \leq p_{(m)}$.*
4. *Let $R = \max\{i : p_{(i)} \leq \gamma_i\}$, and reject the hypotheses $(1), (2), ...(R)$ (If no such R exists, don't reject any hypothesis).*

This procedure is similar to the original BH95 procedure, with the addition initial step of estimating m_0, and the different set of constants used to determine R. The modified step-down BH procedure is defined in the same way, except that in step 4 we take $R = \min\{i : p_{(i)} > \alpha_i\} - 1$.

The next theorem gives the bound on the FDR for the above procedure under the above assumptions (a very similar result was given by [5]):

THEOREM 2.3. *Let $\hat{m}_0 \equiv \hat{m}_0(\vec{p})$ be a monotonic estimator for m_0. Consider the modified step-up BH procedure defined above. Let $\hat{m}_0^{(\backslash 1)}(\vec{p}) \equiv \hat{m}_0(p_2, .., p_m)$ be the same estimator, but disregarding the first (null) p-value p_1. Assume that the null p-values are i.i.d. $U[0,1]$. Then the procedure satisfies:*

$$(2.2) \qquad\qquad FDR = E\left[\frac{V}{R^+}\right] \leq m_0 q E\left[\frac{1}{\hat{m}_0^{(\backslash 1)}}\right]$$

Here p_1 is a representative of one of the true null p-values. The modified estimator $\hat{m}_0^{(\backslash 1)}$ which excludes p_1 cannot be implemented in practice, as the researcher does not know which of the p-values are null, but for any estimator $\hat{m}_0$ we can still consider this hypothetical estimator (in similar vain to the 'oracle' procedure sometimes considered in the literature) and study it's statistical properties - it only serves for an hypothetical auxiliary procedure which is used in the proof of the theorem,

and the theorem applies to the practical original procedure with the estimator $\hat{m}_0$ which does use p_1 (as well as all other p-values). The proof of Thm. 2.3 is given in Supplement A for completeness. In general, a direct computation, or bounding of the FDR for a given procedure is a demanding task, which depends heavily on the procedure's details, and suffers from complicated dependence on the rejection of different hypotheses, reflected in the computation of $E[V/R^+]$ (this is true even if the p-values themselves are independent) and therefore there is no general way to prove FDR controlling properties of various procedures. The advantage of Thm. 2.3 is that it provides a direct method for proving control for a wide class of procedures, by simply bounding the reciprocal mean of the estimator for m_0. In the next section we use this theorem to prove control of the FDR for two procedures, based on different estimators $\hat{m}_0$, and $\tilde{m}_0$ which we propose. We are not aware of a direct way for proving control of the FDR for these procedures, thus demonstrating the power and generality of the theorem.

3. The proposed procedures. In this section we propose two FDR controlling procedures. We show that they achieve direct control of q, the desired value of the FDR, while producing a list of R' discoveries satisfying almost always $R' \geq R$, the corresponding BH95 value. The procedures are particular cases of Def. 2.2. According to Thm. 2.3 any estimator that satisfies our monotonicity assumption bounds the FDR by $FDR \leq m_0 q E[1/\hat{m}_0^{(\mathcal{V})}]$. Therefore, in order to show that the FDR is controlled, it suffices to bound $E[1/\hat{m}_0^{(\mathcal{V})}]$. In particular, if we want to achieve a certain FDR control level q, we need to verify that

$$(3.1) \qquad E\left[\frac{1}{\hat{m}_0^{(\mathcal{V})}}\right] \leq \frac{1}{m_0}$$

Our first estimator is based on

$$(3.2) \qquad \hat{m}_0' = 2\sum_{j=1}^{m} p_j$$

$\hat{m}_0'$ was used by [19] for estimation but without proving control of the FDR. The second estimator is based on

$$(3.3) \qquad \tilde{m}_0' = -\sum_{i=1}^{m} \log(1 - p_i)$$

For both estimators we first show that eq. (1.9) is satisfied and hence both can be used for FDR *estimation*. Next we describe the procedure to be used for *control* of the FDR, which is proved by showing, for slightly modified versions of both estimators (see blow), $\hat{m}_0$ and $\tilde{m}_0$ that the bound eq. (3.1) is satisfied. Both $\hat{m}_0', \tilde{m}_0'$ are monotonic estimators according to Def. 2.1. Our claims are:

1. Both estimators are conservative, i.e. their expectation is at least m_0. Moreover, as the statistical power of each individual test increases, and the p_i of the alternative hypothesis approach zero, our estimators converge (in expectation) to the true value of m_0.
2. Both procedures control the FDR - for the list of R' discoveries we have $FDR \leq q$.
3. In nearly all cases of interest the number of discoveries obtained by our procedures exceeds the number obtained (for the same value of q) by the BH95 procedure, i.e. $R' \geq R$. This holds since nearly always $\hat{m}_0 \leq m$ (exceptions occur when there are almost no false hypotheses, i.e. m and m_0 are very close).

A reasonable requirement from an estimator for m_0 should be that it is conservative (i.e. larger than m_0 in expectation). We would also like our estimator to be (approximately) unbiased, at least

when all hypotheses are null. For otherwise we will get a systematic over-estimation of m_0 and a corresponding under-estimation of the FDR. Finally, a desirable property is being *asymptotically unbiased* - that is, even when there are non-null hypothesis, when the sample size of the individual tests grows to infinity, we would want the estimator to converge, on expectation, to the true value m_0. These properties were dealt with in [19], where it was shown that $\hat{m}_0'$ indeed satisfy them. Here we show them for both our procedures:

CLAIM 3.1.   *(a.) Both estimators are conservative:*

$$(3.4) \qquad\qquad\qquad\qquad E[\hat{m}_0'], E[\tilde{m}_0'] \geq m_0$$

(b.) Assume that the sample size of all tests goes to infinity, and thus $E[p_i] \to 0$ for $i = m_0+1, .., m$. Then both estimators converge in expectation to m_0:

$$(3.5) \qquad\qquad\qquad\qquad E[\hat{m}_0'] \to m_0, \ E[\tilde{m}_0'] \to m_0$$

Proof:
(a.)

$$(3.6) \qquad E[\hat{m}_0'] = 2\sum_{j=1}^{m} E[p_j] = 2(\sum_{j=1}^{m_0} E[p_j] + \sum_{j=m_0+1}^{m} E[p_j]) = m_0 + 2\sum_{j=m_0+1}^{m} E[p_j] \geq m_0.$$

$$(3.7) \qquad \begin{aligned} E[\tilde{m}_0'] &= \sum_{j=1}^{m} E[\log(1-p_j)] = \sum_{j=1}^{m_0} E[\log(1-p_j)] + \sum_{j=m_0+1}^{m} E[\log(1-p_j)] = \\ &= m_0 + \sum_{j=m_0+1}^{m} E[\log(1-p_j)]) \geq m_0 \end{aligned}$$

(b.) From the two equations above it is clear that as all the alternative $E[p_j]$ approache zero, the expectation of both estimators converges to m_0.

∎

In order to show control of the FDR using Thm. 2.3, we have to apply small corrections to both estimators, turning them into conservative estimators (i.e. over-estimating m_0). This is due to two reasons: the first is that the bound on the FDR given in Thm. 2.3 uses $\hat{m}_0^{(\mathcal{Y})}$ (rather than $\hat{m}_0$) and thus we 'lose' one of the p-values and need to correct for that. The second reason is that $\hat{m}_0^{(\mathcal{Y})}$ appears in the denominator, and its fluctuations have asymmetric influence on the FDR bound. This can be illustrated by using Jensen's inequality which gives $E[1/\hat{m}_0^{(\mathcal{Y})}] \geq 1/E[\hat{m}_0^{(\mathcal{Y})}]$, thus showing that an unbiased estimator for m_0 will typically show a bias when its reciprocal is used. Nevertheless, we show that these two effects can be overcome by applying a small correction, which becomes negligible as the number of hypotheses go to infinity.

3.1. *The IBHsum procedure.*   Our first estimator is based on $\hat{m}_0'$ (see eq. (3.2)) that was also used by [19] but only for *estimation* and not for *control*. Since for the m_0 variables for which the null hypothesis holds we have $p_i^{true} \sim U[0,1] \Rightarrow E[p_i^{true}] = \frac{1}{2}$, it is trivial to see that $E[\hat{m}_0'] \geq m_0$. To show that $E[\hat{m}_0'] \leq m$, we have to make a further assumption regarding the alternative p-values p_i^{false}: We denote the distribution of p_i^{false} by f_i^{false}, i.e. $p_i^{false} \sim f_i^{false}$. If all the f_i's are *stochastically smaller* [2] than the uniform distribution, ($f_i^{false} \leq_{st} U[0,1]$), we have $E[p_i^{false}] \leq \frac{1}{2}$ which immediately implies

$E[\hat{m}'_0] \leq m$, (a probability density function f is said to be stochastically smaller than a probability density function g, $f \leq_{st} g$, if $F(x) = \int_{-\infty}^{x} f(t)dt \geq G(x) = \int_{-\infty}^{x} g(t)dt \quad \forall x \in (-\infty, \infty)$ [2]).

We introduce the following modified estimator:

$$(3.8) \qquad \hat{m}_0 = C(m) \cdot min \left[m, max\big(s(m), \hat{m}'_0\big) \right],$$

where $C(m), s(m)$ are universal correction factors that ensure that the condition (3.1) is satisfied (for details see Supplement B). The correction factors were computed numerically and are presented in Fig. 2. When $m \to \infty$, $C \to 1$ and $s/m \to 0$, and therefore the corrections become negligible and the estimator $\hat{m}_0$ reduces to $\hat{m}'_0$.
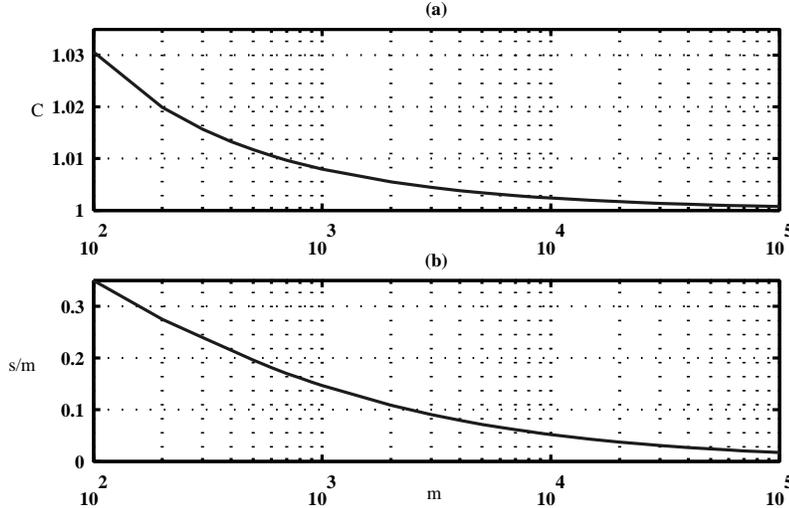
FIG 2. *The correction functions $C(m)$ and $s(m)/m$ (see eq. (3.8)). As $m \to \infty$ the multiplicative correction $C(m)$ approaches one, while the (normalized) threshold $s(m)/m$ (used when $\hat{m}'_0 \leq s(m)$) goes to zero, thus $\hat{m}_0$ reduces to the un-corrected $\hat{m}'_0$*

3.2. *The IBHlog estimator.* In this section we propose another estimator for m_0, based on $\tilde{m}'_0$, (see eq. (3.3)). Again, since for $i = 1, 2...m_0$ we have $p_i^{true} \sim U[0, 1] \Rightarrow E[-log(1 - p_i)] = 1$ and therefore $E[\tilde{m}'_0] \geq m_0$. Furthermore, if all the alternative p-values p_i^{false} have a distribution which is stochastically smaller than the uniform distribution ($f_{p_i}^{false}(p) \leq_{st} U[0, 1]$), then $E[-log(1 - p_i^{\text{false}})] \leq 1$, and therefore $E[\tilde{m}'_0] \leq m$.

The advantage of using the second estimator $\tilde{m}'_0$ over $\hat{m}'_0$ is that when $f_{p_i}^{false}(p) \leq_{st} U[0, 1]$, the alternative hypothesis generates p-values skewed to the left. Since $-\log(1 - p) < 2p, \forall p < \frac{1}{2}$ (see eqs. (3.2) and (3.3)), this typically implies $\tilde{m}'_0 \leq \hat{m}'_0$ and thus $\tilde{m}'_0$ is typically closer to the true m_0. A possible drawback is that the variance of $\tilde{m}'_0$ is typically larger than that of $\hat{m}'_0$, which might result in an instability in the estimation of m_0.

Proving control of the FDR for $\hat{m}_0$ is difficult since we need to bound $1/\hat{m}_0$ which has a complicated distribution. Here we show that the distribution of $\tilde{m}'_0$ is much simpler, and this enables us to prove control of the FDR by introducing only a slight additive correction.

CLAIM 3.2. *Define the (corrected) estimator:*

$$(3.9) \qquad \tilde{m}_0 \equiv 2 + \tilde{m}'_0 = 2 - \sum_{i=1}^{m} \log(1 - p_i)$$

Assume that the null p-values are i.i.d $U[0,1]$. Then the modified BH procedure with estimator $\tilde{m}_0$ and parameter q controls the FDR at level $\leq q$.

The proof is achieved by bounding $E[1/\tilde{m}_0^{(\gamma)}]$ and then using Thm. 2.3. See Supplement C for full details

4. Is the FDR monotonic?.   In this section we take a slight detour from the study of our proposed procedures to investigate the following question: is it generally true that by modifying an FDR procedure to be more stringent, one is guaranteed to obtain a more conservative control on the FDR? The motivation for dealing with this question in the context of the current paper (which deals with the control property of a modified BH procedure) comes from the fact that Thm. 2.3 was proved only for step-up procedures, which leads us to ask whether it holds also for the more conservative step-down case. Monotonicity is a natural property that one might expect when performing statistical tests, as it allows the researcher to choose a trade-off between maximizing the statistical power and minimizing the risk of making false discoveries. The analogous question for a single hypothesis is whether taking a more conservative (lower) p-value cutoff guarantees to reduce the risk of making a type-I error, and is trivially answered in the affirmative. Our formulation of the question in the multiple-hypothesis settings using FDR is as follows: Given two procedures, $B^{(1)}, B^{(2)}$ (possibly parameterized by q or other parameters), and assuming that for any realization of the p-values, $B^{(2)}$ passes more hypotheses than $B^{(1)}$, is it true that $FDR^{(1)} \leq FDR^{(2)}$? while this statement seems a natural and plausible property of FDR procedures, we are not aware of any previous treatment of it in the literature. Here we show that under certain monotonicity conditions on the alternative hypothesis p-values distribution, one can prove this monotonicity property of the FDR.

THEOREM 4.1.   *Let $\vec{p} = (p_{1..m})$ be a set of independent p-values. Assume that f, the marginal probability density function of the alternatives, is monotonically non-increasing and differentiable. Let $B^{(i)}$ be two threshold FDR procedures rejecting $R^{(i)}(\vec{p})$ hypotheses and each having $FDR^{(i)}$, $i = 1, 2$. Assume that for any q, $R^{(1)}(\vec{p}) \leq R^{(2)}(\vec{p})$, $\forall \vec{p}$. Then it also holds that $FDR^{(1)} \leq FDR^{(2)}$.*

The proof is given in Supplement D. A particular application of the above theorem is showing that step-down procedures give better FDR then step-up procedures. Thus, we immediately get:

COROLLARY 4.2.   *The statement of Thm. 2.3 holds also for the step-down procedure, provided that the alternative f is monotonically decreasing.*

The above conditions for monotonicity might appear a bit restrictive, and one could hope to relax them - for example require only $f \leq_{st} U[0,1]$ instead of monotonicity. We have found that, perhaps surprisingly, monotonicity of the FDR does not hold under such relaxed conditions, by giving an example in which FDR monotonicity is violated, even for a simples case of independent test statistics (both null and non-null), when $f \leq_{st} U[0,1]$, and when the FDR procedures themselves are monotonic. It is thus not obvious at all that in practice we will always observe a monotonic behavior of the FDR, and thus it is possible to get a higher FDR for a more conservative procedure.

EXAMPLE 4.3.   *Let $m = 3$ and $m_0 = 1$. Let the two alternative hypotheses p-values be taken from a mixture distribution, $p_i \sim \epsilon U[0,\epsilon] + (1-\epsilon)\delta(p_i - \epsilon)$ for some $0 < \epsilon < 1$. Thus, p_2, p_3 are 'truncated' uniform r.v.s., having $1-\epsilon$ of their mass concentrated at ϵ, and the rest (ϵ) uniformly distributed on $[0,\epsilon]$; their distributions are stochastically smaller than $U[0,1]$. For simplicity of computations we assume that $\epsilon << 1$ and thus look only at the first order in ϵ, although the example's conclusion*

holds for any $\epsilon > 0$. Let $P^{(1)}$ be the procedure always rejecting the lowest p-value and $P^{(2)}$ be the procedure rejecting the two lowest p-values (we assume that ties are handled in the same way by both procedures, e.g. by taking p-values in lexicographic order - the precise tie-breaking rule does not change the example's results). We next compute the FDR for both procedures:

$$(4.1) \qquad FDR^{(1)} = Pr(p_1 < p_2, p_3) = \epsilon[(\epsilon^2/3 + 2\epsilon(1-\epsilon)/2 + (1-\epsilon)^2] = \epsilon + O(\epsilon^2)$$

$$(4.2) \qquad FDR^{(2)} = (1 - Pr(p_1 > p_2, p_3))/2 = [1 - (1-\epsilon) - \epsilon^3/3]/2 = \epsilon/2 + O(\epsilon^3)$$

Thus for ϵ small enough $FDR^{(1)} > FDR^{(2)}$ and the more conservative procedure leads, in fact, to a higher FDR.

5. Synthetic data obtained by simulations. We applied our method, as well as several others (see below), to synthetic data obtained by simulations performed along the lines of gavrilov:2009, with full details presented in Supplement E. The advantage of working with synthetic data is that several parameters of interest are under full control, and one can investigate their effect on the quality of different procedures and bounds. Furthermore, by performing repeated simulations, one can determine not only the (*expected* value) FDR but also the entire *distribution* of V/R^+. One should bear in mind that results based on specific simulations might have limited applicability and are hard to generalize, since the simulations use specific configurations (e.g. data distribution, test to determine p-values, hypothesis dependency structure etc.). A comprehensive simulation capturing all possible behaviors of the hypothesis is infeasible, but we have tried to explore various different plausible scenarios which might be encountered in practice, by changing the number of (total and null) hypothesis and their dependency structure, with both positive and negative correlations. The simulations produce two kinds of Gaussian random variables: $Z_1, ... Z_{m_0}$, sampled from the standard normal distribution $P_0 \equiv N(0,1)$, and $Z_{m_0+1}, ... Z_m$, sampled from $P_1 \equiv N(\mu_1, 1)$, centered on $\mu_1 > 0$. All variables (both null and non-null) are sampled with covariance ρ, $(0 \leq \rho \leq 1)$: at the extreme cases, setting $\rho = 0$ corresponds to independent variables whereas $\rho = 1$ to full (deterministic) dependency. For each Z_i the corresponding *two-tailed* p-value is obtained, $p_i = 2\Phi(-|Z_i|)$, where Φ is the standard Gaussian cumulative distribution function. The obtained p_i's have a uniform $U[0,1]$ distribution for $i = 1, .., m_0$ (corresponding to the null hypothesis) and a distribution stochastically smaller than uniform for $i = m_0 + 1, .., m$ (the alternative hypothesis).

A set of m such variables constitutes a single instance or realization of the data to be analyzed. To get accurate estimates of the FDR and the V/R^+ distribution, we generated for each simulation 50000 such realizations, which generally gave highly accurate and reproducible estimates. Under the null hypotheses all variables are sampled from the first distribution, m p-values are calculated accordingly and used as input to one of the procedures with a desired FDR bound q, producing a list of R rejections. As opposed to real data, here one can go back and identify those V among the R that were falsely rejected (i.e. were, in fact, selected from P_0). This way one can keep track of the true values of V/R^+, their mean (calculated over a large number of instances), variance, etc. One important goal of the simulation is comparing our procedures to existing ones. Specifically, we compare our procedure to: (1) the BH95 procedure as described in the introduction, (2) the BKY procedure which defines a local (i-dependent) estimator for m_0, given by $\hat{m}_0^{BKY} = m+1-i(1-q)$, and uses it in the step down manner of the BH95 procedure with $q^* = qm/\hat{m}_0^{BKY}$, (3) the STS procedure which introduces $\hat{m}_0^{STS} = (m + 1 - r(\lambda))/(1 - \lambda)$ as the estimator for m_0 where $r(\lambda) = \#\{p_i \leq \lambda\}$, and then uses the step-up BH95 procedure, with $q^* = qm/\hat{m}_0^{STS}$, with the requirement that all the rejected $p_i \leq \lambda$ (throughout this paper we used the STS procedure with $\lambda = 0.5$). We present here two kinds of results derived from such simulations. First we compare the values of $FDR = E(V/R^+)$ obtained by the procedures discussed above: BH95, BKY, STS, IBHsum and IBHlog when the

hypotheses are dependent. In particular, we demonstrate that for positive correlations $\rho > 0$ our IBH as well as the BKY procedures yield for a given desired value of q, an FDR that is either less than q or exceeds it slightly. On the other hand the STS method produce, for $\rho > 0$, values of FDR that exceeds q by a large margin. The second aim is to assess the extent to which the value of V/R^+, obtained for a particular realization, will violate the bound, especially for the IBH methods.

As an overview we start by presenting in Fig. 3 the performance of our proposed IBHsum procedure for fixed $m = 500$ and $q = 0.05, 0.2$, and for a wide range of the parameters m_0/m (fraction of alternative hypotheses) and μ_1 (signal strength), by estimating the expected value $FDR = E(V/R^+)$ from our simulations. Fig. 3a and c are for the independent case and show both step-down and step-up results. As we can see, the two become identical when the signal (μ_1) is strong or when m_0/m is small. Fig. 3b and d are for the positively dependent case ($\rho = 0.8$) for which the procedure is not proved to control the FDR. Indeed we can observe in Fig. 3b violation of the FDR level q for large signals (μ_1); this violation of the bound for the dependent case will be discussed later.
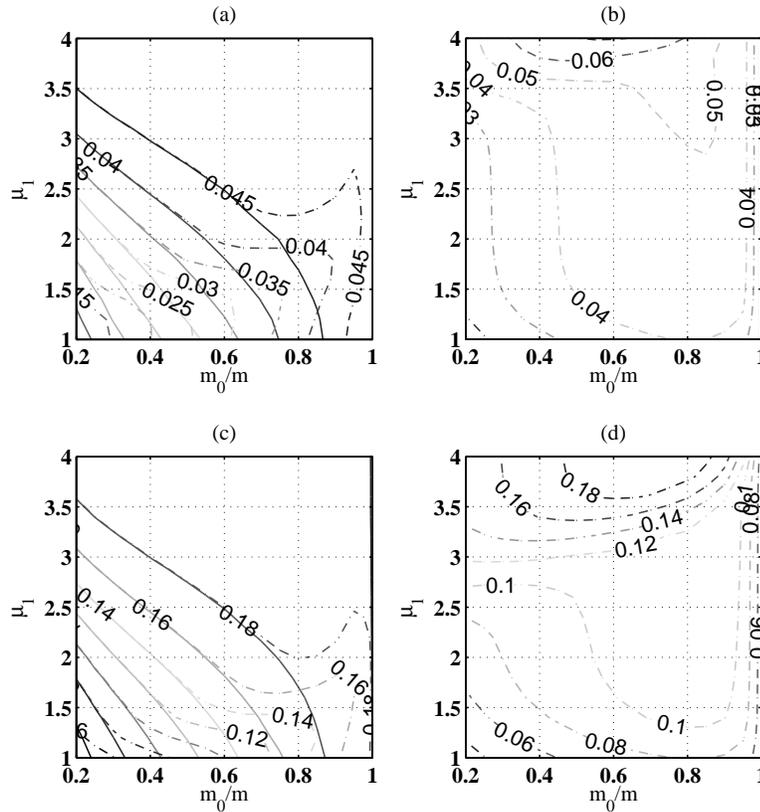
FIG 3. *Isolines of $E(V/R^+)$, measured for the IBHsum procedure by simulations, presented in the $(\mu_1, m_0/m)$ plane. The solid lines in (a) and (c) are for the step-up procedure and the dashed lines for the step-down procedure. (a) and (c) are for the independent case ($\rho = 0$). (b) and (d) are for the positive dependency case $\rho = 0.8$. The FDR levels are $q = 0.05$ in (a),(b) and $q = 0.2$ in (c),(d). In (b) we find $E(V/R^+) > 0.05$ for large μ_1, in violation of the bound $q = 0.05$. The step-up and step-down procedures tend to coincide for independent p-values and low m_0/m; the differences between them are more significant when the signal is weak (small μ_1) and m_0/m is high.*

5.1. *Comparison of several methods under dependency .*   Here we fixed the signal parameter $\mu_1 = 3.5$, and varied m_0/m between 0.2 and 1 (for $m = 500$). We present, in Fig. 4a, c and e results

obtained for $\rho = 0$ (complete independence) and in Fig. 4b, d and f for $\rho = 0.8$ (strong dependence). For each instance we applied the five procedures with $q = 0.05$. For STS we chose $\lambda = 0.5$, and our IBHsum and IBHlog were employed in a step down manner. Fig. 4a and b present for each method the mean value of V/R^+, as a function of m_0/m. These means provide excellent estimates of $E(V/R^+)$, and they reveal that as expected, for $\rho = 0$ all methods satisfy the bound $E(V/R^+) \leq q$. The STS and IBH come closest to saturating the bound, with BKY slightly lower and BH95 significantly lower. The figures show also the result obtained by an "oracle", namely the procedure that uses the known value of m_0 in order to determine R' according to eqs. (1.7) and (1.8).

For $\rho > 0$ no proved upper bound exists for either of the BKY, STS or IBH procedures. Furthermore, the proof of [6] for the BH95 procedure does not hold for two-tailed tests: indeed, as can be seen on Fig. 4b, the FDR obtained by the oracle procedure (slightly) violates the bound $q = 0.05$ for $m_0/m \leq 0.3$, in agreement with the violation reported in [21]. Therefore it is important to assess the extent to which $E(V/R^+)$ obtained by each of these methods violates the bound q in the presence of positive correlations between the hypotheses. As seen in Fig. 4b, for $\rho = 0.8$ the STS method produces a measured FDR that overshoots the value $q = 0.05$ of the bound by more than twice, for most of the range of m_0 values studied. In comparison, the other methods (BH95,BKY,IBHsum) provide FDR which remains below the bound or exceeds it slightly for a narrow range of m_0. The IBHlog procedure also violates the bound for nearly the entire range of m_0/m, but by much less than STS.

We conclude these comparisons between the different procedures by presenting, in Fig. 4c and d their power, measured as the fraction of correctly rejected hypotheses, or 'True Discovery Rate'. For each realization we calculated $S = R - V$ and plotted the ratio $S/m_1 = (R-V)/(m-m_0)$, averaged over all instances. This measure of power is one minus the type two error rate, known as the False Non-Discovery Rate T/m_1 ([12]). For the independent case $\rho = 0$ the power of the ORC, BKY, STS and both IBH procedures are very close and much better than that of BH95. For $\rho = 0.8$ STS has the highest power, followed closely by the oracle, both IBH and BKY, with a large gap to BH95. Again, one should bear in mind that STS has the largest number of discoveries R, at the cost of violating strongly the bound of 0.05 on the FDR. Interestingly, there is no simple monotonicity relationship between the values of the FDR, $E(V/R^+)$, and the True Discovery Rate $E(S/m_1)$.

Fig. 4e shows the standard deviation (st.d.) of V/R^+ for the independent case, and Fig. 4f for the positively dependent case. As can be seen when the p-values are independent the st.d. is very similar for all the procedures, but increases steeply as $m_0/m \to 1$. In the case of dependent p-values the situation becomes worse; for nearly the entire range of m_0/m the coefficient of variance $cv = st.d.(V/R^+)/E(V/R^+)$ is greater than 1. Also, as will be mentioned below, for real data the st.d. of the STS procedure is significantly higher than that of the IBH. These high values of st.d. result from the FDR definition, since the expectation of V/R^+ takes into account many realizations with $R = 0$ that give, by definition, $V/R^+ = 0$, making the distribution of V/R^+ very non-symmetric. A comparison similar to the one presented in Fig. 4 for $q = 0.05$ is presented in Supplement E Fig. 4 for $q = 0.2$, and provides similar observations. We thus conclude that for $\rho = 0$ our IBH procedures provide, an expected improvement over the BH95 in therm of power and saturation of the bound and their performance is comparable to that of the other adaptive methods tested. For dependent variables STS violate the bound on $E(V/R^+)$ much more than the IBHlog and the IBHsum which violate it only slightly.

5.2. *Applicability for a particular realization.* Controlling the FDR at a level q means that the *average* fraction of false rejections is no larger than q. It could still be the case that on average the fraction of false rejections is controlled, yet for a large percentage of the realizations one gets many false rejections and a high proportion of false discoveries. In contrast to the average behavior, captured by the FDR definition, questions involving the distribution of false rejections, affecting the
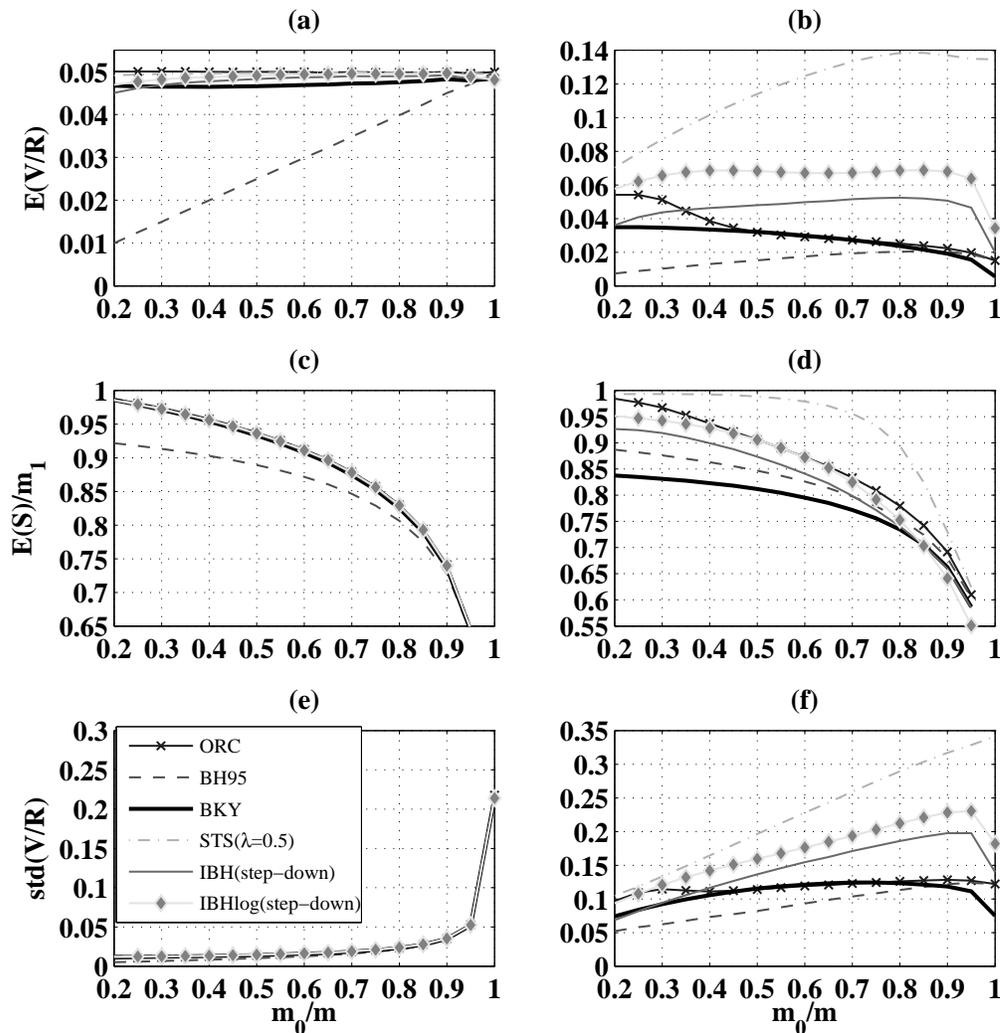
FIG 4. *Results obtained for synthetic data with $m = 500$ hypotheses; m_0 was varied, the FDR was set at $q = 0.05$, the mean of the distributions P_1 was $\mu_1 = 3.5$ and the data were drawn either with covariance $\rho = 0$ [(a), (c) and (e)] or $\rho = 0.8$ [(b), (d) and (f)]. Six methods were compared: oracle (ORC), BH95, BKY, STS and our two IBH procedures (in a step down manner), showing $E(V/R^+)$ in (a) and (b), the power $E(S)/m_1$ in (c) and (d), and the standard deviation (st.d.) of V/R^+ in (e) and (f), for the independent case and positively dependent cases, respectively.*

behavior of a particular realization, were not studied much in the literature (a notable exception is [17] who studied the variance of R). We therefore set out to address the issue of validity of the bound for a particular realization, by calculating for the synthetic data the probability $Pr(\frac{V}{R^+} \leq q)$. This was done for $q = 0.05$ for the six procedures (ORC, BH95, BKY, STS, IBHsum and IBHlog, the latter two in step-down mode). The probability $Pr(\frac{V}{R^+} \leq q)$ was estimated by computing, for each procedure, the fraction of realizations in which we indeed got $\frac{V}{R^+} \leq q$. In such a comparison one should bear in mind that a conservative procedure, such as BH95, restricts the discoveries much more than a procedure that produces tight bounds (such as the oracle). For example, looking at Fig. 4a we see that the mean value $E(V/R^+)$ of BH95 is much lower than $q = 0.05$, and hence the weight of the tail of the distribution of V/R^+ values that "leaks" to $V/R^+ > 0.05$ is very small, whereas for

the oracle, which has $E(V/R^+) \approx 0.05$, the probability of exceeding 0.05 is close to 0.5, and if we want to guarantee that $Pr(V/R^+ < B) \approx 1$, we must set B at a value which is significantly larger than the FDR bound q. As seen in Fig. 5a, the results of IBH are slightly more conservative than the oracle in the case of independence, while all improved procedures have fairly similar results. In the case of strong dependency, Fig. 5b, the differences between the procedures are more pronounced; the STS is the most permissive procedure.

It is very interesting to see that in the case of positive dependent statistics the probability to violate the bound is smaller, although $E[V/R^+]$ is larger. This is again due to the fact that in these cases we get $R = 0$ for many realizations, which means that $V/R^+ = 0$, i.e. the variance of V/R^+ is increased for positive correlations, whereas for the independent case V/R^+ is very likely to be close to its expectation. Further study on the distribution of V/R^+ is required in order to shed light on the behavior of different procedures for particular realizations. Fig. 5c and d present the cumulative distribution function (CDF) of V/R^+ for specific set of parameters, $m = 1000, m_0/m = 0.7, \mu_1 = 3.5, q = 0.05$, and the different procedures to be compared, for the independent case (Fig. 5c) and for the positive dependence case (Fig. 5d). We would like to emphasize two points: 1. the CDF of our improved procedures have very similar behavior to the other improved procedures. 2. while in the independent case the distribution is close to symmetric, under dependency the distribution is very non-symmetric, and hence controling the mean (of V/R^+) is almost irrelevant.

6. Application to gene expression data .   As an ultimate tests for their utility, we wanted to asses the performance of our new procedures on real life data, which typically provide complex and unexpected dependency structures which are hard to capture in simulations. We therefore applied our procedures that were described in Sec. 3 to publicly available expression data. First we present in full detail how our procedures were applied to two datasets. Next, our procedures were applied to 33 datasets and results were compared with those obtained by several other procedures: the original BH95 and the improved bounds of BKY [5] and STS [25] with $\lambda = 0.5$.

6.1. *Detailed application of our procedures.*   The first dataset used is that of [1] who studied several types of childhood leukemia. We focus here on search for genes whose expression separated 6 patients with normal bone marrow from 11 T-Cell Acute Lymphoblastic Leukemia patients, which yielded a large number of discoveries (differentiating genes). The number of hypotheses (e.g. potentially differentiating probesets) was $m = 21288$; the corresponding reported p-values were ordered and plotted on Fig. 1a. Our estimators for m_0, obtained using eq. (3.8) and (3.9) for this data, were $\hat{m}_0 = 7093, \tilde{m}_0 = 6380$, and the estimated numbers of discoveries were $m - \hat{m}_0 \approx 14000, m - \tilde{m}_0 \approx 15000$.

The second study, of [18] on breast cancer, had a relatively small number of discoveries. The aim was to find genes that differentiated early discovery breast cancer cases of poor and good outcomes, i.e. were differentially expressed between tumors obtained from 38 subjects that died of the disease and from 121 patients who were alive. The number of hypotheses was $m = 44611$, and our p-values based estimators for m_0 (plotted in Fig. 1b) were $\hat{m}_0 = 38587, \tilde{m}_0 = 37580$.

For both studies we have set the desired FDR value at $q = 0.1$. We plot in Fig. 1 the sorted p-values $p_{(i)}$ versus i/m for these two datasets. In each of the two figures we show three FDR lines; the α_i of BH95 (see eq. (1.3)) and the values of γ_i corresponding to our two procedures, (see eq. (2.1)).

For the first dataset the BH95 procedure yields at $q = 0.1$ a large number of $R = 0.6065 \cdot 21288 = 12912$ discoveries (see Fig. 1a). When we apply our procedure we get, at the same FDR, $R' = 0.746 \cdot 21288 = 15884$ (for the IBHsum) discoveries, i.e. 23% more.

The BH95 procedure yields for the second dataset (at $q = 0.1$) $R = 499$ discoveries. When we apply our procedure we get, at the same FDR, $R' = 621$ (for the IBHsum) discoveries, i.e. 24% more.
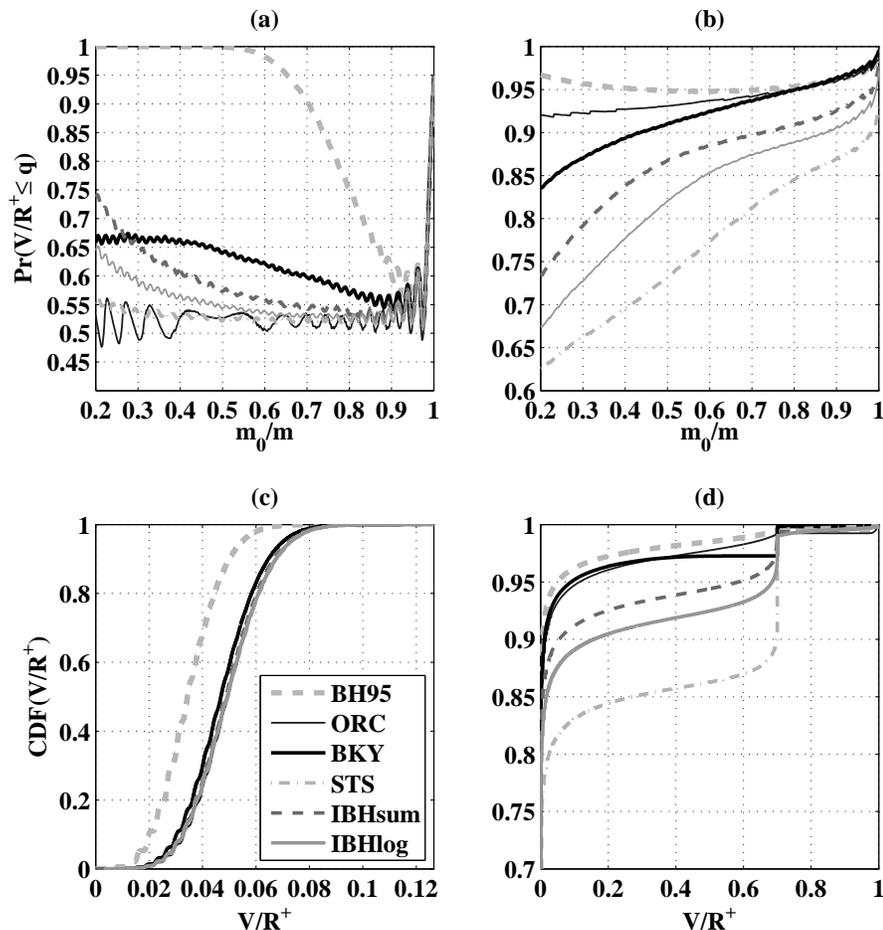
FIG 5. *(a) and (b) shows the probability that a single instance satisfies the desired FDR level q as a function of m_0/m. Results are shown for simulated data with $m = 1000$ hypotheses, the mean of the distribution P_1 was $\mu_1 = 3.5$, the FDR bound was set to $q = 0.05$. Five methods are compared: ORC, BKY, STS, and our two IBH procedures (in the step-down manner). (a) $\rho = 0$ and (b) $\rho = 0.8$. The oscillatory behavior of some bounds is caused by finite size effects. (c) and (d) shows the cumulative distribution function of V/R^+ for $m_0/m = 0.7$, (c) $\rho = 0$ and (d) $\rho = 0.8$ (obtained from 10^6 realizations).*

6.2. *Applying our procedures to many datasets.* We downloaded from the ONCOMINE website [22] p-value vectors that were obtained from 33 comparisons, performed on expression data from 19 studies of various types of cancer: [1, 3, 7–10, 13–16, 18, 23, 26–29, 31–33]. Depending on the biological question at hand, either one or two-tailed tests are appropriate. Therefore we applied our procedures to both test types. We focused on two opposing scenarios: those with a small number (less than 2% of m, for the BH95 procedure with $q = 0.05$) of discoveries, and those with a large number (more than 10% of m). The 33 sorted sets of p_i values are plotted, versus i/m, in Fig. 6, separately for the four types of comparisons that were made (one/two-tailed test, low/high number of discoveries).

As can be seen in Fig. 6, for each type of comparison the sorted p-value curve has a typical shape. In the case of a large number of discoveries, Fig. 6a and c, the curve is more convex (and flatter near zero) than in the case of a small number of discoveries, Fig. 6b and d. Another clear difference is between the two-tailed (Fig. 6a and b) and the one-tailed (Fig. 6c and d) sorted p-value curves. In
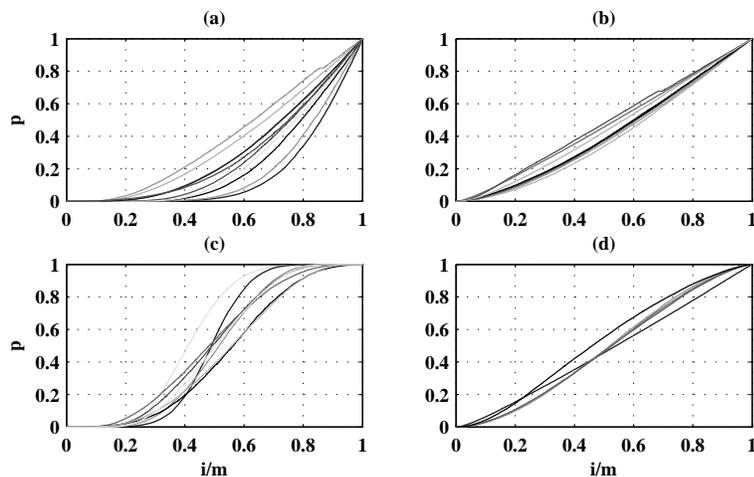
FIG 6. *Sorted p-value vectors from 33 expression datasets of various cancer-related comparisons: (a) - two tailed tests with large numbers of discoveries, (b) - two tailed tests with small numbers of discoveries, (c) - one tailed tests with large numbers of discoveries, (d) - one tailed tests with small numbers of discoveries.*

the case of two-tailed tests, the entire curve is convex, while for one-tailed tests the right side of the curve is concave; the reason is that in the latter case there are very often some hypotheses that are shifted, with respect to the null hypothesis, in the direction opposite to the one tested for by the one-sided test (for example, if one looks for up-regulated genes, there are typically also many down regulated genes, which produce very high p-values). For detailed treatment of FDR estimation in the case of one tailed tests see [19].

We compare here the performance of five procedures: the BH95,BKY,STS, IBHsum and IBHlog (both IBH in the step-down mode). For each of the improved procedures we determined the ratio between the number of rejected hypotheses it yielded and the number of hypotheses rejected by BH95. We present in Table 2 the mean value of this figure of merit and its standard deviation, calculated for the datasets of each of the types of comparisons mentioned above, at $q = 0.05$ and $q = 0.1$.

Inspection of Table 2 reveals that for types (a),(b) - of two tailed tests, irrespective of the number of discoveries and FDR level, STS and both IBH procedures give significantly higher improvement over BH95 than the BKY procedure, with STS performing slightly better than IBHlog, followed by IBHsum. For the one-tailed test with large numbers of discoveries (type (c)) the mean improvement of BKY is the highest while STS and IBHsum are quite similar. IBHlog fails dramatically in this case due to the abundance of p-values close to one, giving an over-estimation of m_0. For type (d), one tailed tests with a small number of discoveries, IBHsum is slightly better than STS and both yield a significantly higher improvement than BKY. In all four types and for all values of FDR, the standard deviations of V/R^+ of the STS method are significantly higher than those of BKY and the IBHsum procedures. Furthermore, as shown in Sec. 5.1 (see Fig. 4b), in the case of positively dependent test statistics the STS procedure loses control of the FDR in a much more drastic manner than our IBH procedures. Since we expect that correlations between the expression profiles of different genes will be present in most data, the STS method may produce unreliable values of the figure of merit presented here.

In summary, our IBH procedures constitute in all cases a significant improvement over the original BH95; in all but one of the comparison types the improvement is significantly better than that of the BKY method. Comparison with STS yields mixed results, but the edge of STS over IBH in two of the four comparison types is overshadowed by the fact that STS does not provide a reliable bound

q	BKY	STS	IBHsum	IBHlog
a. Two tailed, large number of discoveries (10 studies)				
0.05	1.110 (0.043)	1.239 (0.138)	1.200 (0.110)	1.222 (0.130)
0.1	1.155 (0.057)	1.258 (0.117)	1.213 (0.087)	1.237 (0.102)
b. Two tailed, small number of discoveries (10 studies)				
0.05	1.003 (0.003)	1.316 (0.197)	1.231 (0.140)	1.291 (0.179)
0.1	1.017 (0.027)	1.308 (0.161)	1.230 (0.117)	1.275 (0.137)
c. One tailed, large number of discoveries (8 studies)				
0.05	1.049 (0.019)	1.011 (0.033)	1.014 (0.026)	0.108 (0.306)
0.1	1.062 (0.026)	1.012 (0.0340	1.014 (0.024)	0.108 (0.305)
d. One tailed, small number of discoveries (5 studies)				
0.05	0.998 (0.020)	1.027 (0.052)	1.025 (0.017)	0.882 (0.123)
0.1	1.004 (0.031)	1.028 (0.079)	1.031 (0.022)	0.888 (0.120)

TABLE 2

Comparison of the improvement in power (ratio between numbers of rejected hypotheses with respect to the BH95 procedure: R/R_{BH95}) of several methods: BKY [5], STS [25] IBHsum and IBHlog in the step-down version. Mean values and standard deviations (in parentheses) are given for each of the four types of comparisons.

for datasets with positive correlations between probe sets, while IBH remains reliable.

7. Discussion. We addressed the problem of controlling the False Discovery Rate in the case of a large number of comparisons, or hypotheses to be tested simultaneously. Providing a reliable and possibly tight bound on the FDR is an issue of major importance for analysis of high-throughput biological data, such as obtained using gene expression microarrays. We presented here two estimators of m_0, the number of true null hypotheses. We proved that both estimators can be used for FDR estimation and, more importantly, for FDR control. Thus, we added two procedures to the rather limited repertoire of improved FDR procedures for which control of the FDR is known to hold. Our proof of control relies on a general theorem, which provides a bound on the FDR for improved procedure using any estimator $\hat{m}_0(p_1, ...p_m)$ provided a condition of monotonicity is satisfied, and one is able to bound the reciprocal mean of the estimator. In addition, we proved a novel result, that FDR procedures satisfy a monotonicity property under some very plausible assumptions. As a corollary of this theorem, we show that any bound on the FDR that was proved for the step up procedure, holds also for the more conservative step down procedure as well. Our proofs of control hold only for the independent case. For the dependent case, results for control are even more scarce, and limited to certain specific types of dependency. We therefore studied the behavior of our procedures, compared to others known from the literature, under dependency, using simulations. In addition to studying behavior under dependency, our simulations also enabled us to understand the distribution of the fraction of false hypothesis, and in particular the probability of violating the bound for a particular given realization. Further research on this aspect of comparing procedures is needed and we expect it to provide interesting new insights and measures for comparisons of different procedures. We finally applied our procedures, as well as several others, to a large number of cancer-related expression datasets. For both real and simulated data, our new procedures provided more rejections (separating genes) than the similar list of Benjamini and Hochberg and the very recently introduced improved bound of BKY [5], for a fixed desired value of the FDR. In some cases the improved bound of STS [25]

gives more rejection than our method, but as we have shown on synthetic data, when there are positive correlations, STS loses control of the FDR in a much more pronounced way than our procedure. To summarize: a researcher may either obtain a desired number of differentially expressed genes at a lower FDR, or get a longer list of such genes at the desired FDR level, at no added computational cost. We recommend using our IBHlog procedure for two-tailed tests, and IBHsum procedure for one-tailed test, to increase discovery power while controlling FDR levels.

References.

[1] A. Andersson, C. Ritz, D. Lindgren, P. Edén, C. Lassen, J. Heldrup, T. Olofsson, J. RA de, M. Fontes, A. Porwit-Macdonald, M. Behrendtz, M. Höglund, B. Johansson, and T. Fioretos. Microarray-based classification of a consecutive series of 121 childhood acute leukemias: Prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia*, 21:1198–1203, 2007.

[2] T. Aven and U. Jensen. *Stochastic Models in Reliability*. Springer, 1999.

[3] Katia Basso, Adam A. Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human B cells. *Nat Genet*, 37:382–390, 2005.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J.R. Statist. Soc. B*, 57(1):289–300, 1995.

[5] Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedure that control the false dicovery rate. *Biometrica*, 93:491–507, 2006.

[6] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1168, 2001.

[7] Michael Bittner. A window on the dynamics of biological switches. *Nat Biotechnol*, 23:183–184, 2005.

[8] Lars Bullinger, Konstanze Döhner, Eric Bair, Stefan Fröhling, Richard F. Schlenk, Robert Tibshirani, Hartmut Döhner, and Jonathan R. Pollack. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*, 350:1605–1616, 2004.

[9] Y. L. Choi, K. Tsukasaki, M. C. O'Neill, Y. Yamada, Y. Onimaru, K. Matsumoto, J. Ohashi, Y. Yamashita, S. Tsutsumi, R. Kaneda, S. Takada, H. Aburatani, S. Kamihira, T. Nakamura, M. Tomonaga, and H. Mano. A genomic analysis of adult T-cell leukemia. *Oncogene*, 26:1245–1255, 2007.

[10] Dondapati Chowdary, Jessica Lathrop, Joanne Skelton, Kathleen Curtin, Thomas Briggs, Yi Zhang, Jack Yu, Yixin Wang, and Abhijit Mazumder. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *J Mol Diagn*, 8:31–39, 2006.

[11] Yulia Gavrilov, Yoav Benjamini, and Sanat K. Sarkar. An adaptive step-down procedure with proven fdr control under independence. *The Annals of Statistics*, 37(2):619–629, 2009.

[12] C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol*, 64:499517, 2002.

[13] Esther Graudens, Virginie Boulanger, Cindy Mollard, Régine Mariage-Samson, Xavier Barlet, Guilaine Grémy, Christine Couillault, Malika Lajémi, Dominique Piatier-Tonneau, Patrick Zaborski, Eric Eveno, Charles Auffray, and Sandrine Imbeaud. Deciphering cellular states of innate tumor drug responses. *Genome Biol*, 7:R19–R19, 2006.

[14] K. Koinuma, Y. Yamashita, W. Liu, H. Hatanaka, K. Kurashina, T. Wada, S. Takada, R. Kaneda, Y. L. Choi, S-I Fujiwara, Y. Miyakura, H. Nagai, and H. Mano. Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene*, 25:139–146, 2006.

[15] P. Laiho, A. Kokko, S. Vanharanta, R. Salovaara, H. Sammalkorpi, H. Järvinen, J-P Mecklin, T. J. Karttunen, K. Tuppurainen, V. Davalos, S. Schwartz, D. Arango, M. J. Mäkinen, and L. A. Aaltonen. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene*, 26:312–320, 2007.

[16] Lance D. Miller, Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T. Liu, and Jonas Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A*, 102:13550–13555, 2005.

[17] Art B. Owen. Variance of the number of false discoveries. *Journal Of The Royal Statistical Society Series B*, 67(3):411–426, 2005.

[18] Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, Edison T. Liu, Lance Miller, Hans Nordgren, Alexander Ploner, Kerstin Sandelin, Peter M.

Shaw, Johanna Smeds, Lambert Skoog, Sara Wedrén, and Jonas Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Res*, 7:R953–R964, 2005.

[19] Stan Pounds and Cheng Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22:1979–1987, 2006.

[20] Stanley B. Pounds. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform*, 7:25–36, 2006.

[21] REINER A. FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal*, 49(1):107–126, 2007.

[22] Daniel R. Rhodes, Shanker Kalyana-Sundaram, Vasudeva Mahavisno, Radhika Varambally, Jianjun Yu, Benjamin B. Briggs, Terrence R. Barrette, Matthew J. Anstet, Colleen Kincead-Beal, Prakash Kulkarni, Sooryanaryana Varambally, Debashis Ghosh, and Arul M. Chinnaiyan. Oncomine 3.0: Genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9:166–180, 2007.

[23] Mary E. Ross, Xiaodong Zhou, Guangchun Song, Sheila A. Shurtleff, Kevin Girtman, W. Kent Williams, Hsi-Che Liu, Rami Mahfouz, Susana C. Raimondi, Noel Lenny, Anami Patel, and James R. Downing. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102:2951–2959, 2003.

[24] John D. Storey. A direct approach to false discovery rate. *J. Roy. Statist. Soc. B*, 64:479–498, 2002.

[25] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J.R. Statist. Soc. B*, 66(1):187–205, 2004.

[26] Peter J. M. Valk, Roel G. W. Verhaak, M. Antoinette Beijen, Claudia A. J. Erpelinck, Sahar Barjesteh van Waalwijk van Doorn-Khosrovani, Judith M. Boer, H. Berna Beverloo, Michael J. Moorhouse, Peter J. van der Spek, Bob Löwenberg, and Ruud Delwel. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*, 350:1617–1628, 2004.

[27] Marc J. van de Vijver, Yudong D. He, Laura J. van't Veer, Hongyue Dai, Augustinus A. M. Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts, Matthew J. Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T. Rutgers, Stephen H. Friend, and René Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347:1999–2009, 2002.

[28] Yixin Wang, Jan G. M. Klijn, Yi Zhang, Anieta M. Sieuwerts, Maxime P. Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E. Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M. J. J. Berns, David Atkins, and John A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.

[29] Toshiaki Watanabe, Takashi Kobunai, Etsuko Toda, Yoko Yamamoto, Takamitsu Kanazawa, Yoshihiro Kazama, Junichiro Tanaka, Toshiaki Tanaka, Tsuyosi Konishi, Yoshihiro Okayama, Yoshikazu Sugimoto, Toshinori Oka, Shin Sasaki, Tetsuichiro Muto, and Hirokazu Nagawa. Distal colorectal cancers with microsatellite instability (MSI) display distinct gene expression profiles that are different from proximal MSI cancers. *Cancer Res*, 66:9804–9808, 2006.

[30] D. Yekutieli and Y. Benjamini. Resampling based false discovery rate controlling multiple testing procedure for correlated test statistics. *J Statist. Plann. Inference*, 82:171–196, 1999.

[31] Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.

[32] Hongjuan Zhao, Anita Langerød, Youngran Ji, Kent W. Nowels, Jahn M. Nesland, Rob Tibshirani, Ida K. Bukholm, Rolf KA resen, David Botstein, Anne-Lise Børresen-Dale, and Stefanie S. Jeffrey. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol Biol Cell*, 15:2523–2536, 2004.

[33] Tong-Tong Zou, Florin M. Selaru, Yan Xu, Valentina Shustova, Jing Yin, Yuriko Mori, David Shibata, Fumiaki Sato, Suma Wang, Andreea Olaru, Elena Deacu, Thomas C. Liu, John M. Abraham, and Stephen J. Meltzer. Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene*, 21:4855–4862, 2002.

DEPARTMENT OF PHYSICS OF COMPLEX SYSTEMS,
THE WEIZMANN INSTITUTE OF SCIENCE, REHOVOT, ISRAEL
E-MAIL: amit.zeisel@weizmann.ac.il

BROAD INSTITUTE OF MIT AND HARVARD,
CAMBRIDGE, MA, USA
E-MAIL: orzuk@broadinstitute.org

DEPARTMENT OF PHYSICS OF COMPLEX SYSTEMS,
THE WEIZMANN INSTITUTE OF SCIENCE, REHOVOT, ISRAEL
E-MAIL: eytan.domany@weizmann.ac.il