Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Current Trends**

False discovery rate paradigms for statistical analyses of microarray gene expression data

Cheng Cheng* and Stan Pounds
Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN;
Cheng Cheng* - Cheng.Cheng@STJUDE.ORG; * Corresponding author
received December 05, 2006; accepted February 02, 2007; published online April 10, 2007

Abstract:
The microarray gene expression applications have greatly stimulated the statistical research on the massive multiple hypothesis tests problem. There is now a large body of literature in this area and basically five paradigms of massive multiple tests: control of the false discovery rate (FDR), estimation of FDR, significance threshold criteria, control of family-wise error rate (FWER) or generalized FWER (gFWER), and empirical Bayes approaches. This paper contains a technical survey of the developments of the FDR-related paradigms, emphasizing precise formulation of the problem, concepts of error measurements, and considerations in applications. The goal is not to do an exhaustive literature survey, but rather to review the current state of the field.

Keywords: multiple tests; false discovery rate; q-value; significance threshold selection; profile information criterion; microarray; gene expression

1 Background:
An important component in the analysis of a microarray gene expression experiment is to identify a list of genes that are differentially expressed under a few different biological conditions (including time course) or across several cell types (normal vs. cancer, different subtypes of a cancer, etc.), or are associated with one or more particular phenotypes of interest. This is often referred as gene expression profiling. In many studies the goal consists of identification and validation to some extent of the gene expression profile to elucidate the biological process **[1-3]**; in others the genes in the expression profile are used as biomarkers to build classifiers for a phenotype (e.g., treatment outcome). **[4]** This paper focuses on a particular statistical aspect in identifying a microarray gene expression profile – the massive multiple tests issue. Henceforth the terms "probe", "probeset" and "gene" will be used interchangeably.

Gene expression profiling usually consists of four major steps: (1) generate and normalize expression signals; (2) test each probe for its differential expression or association with the phenotype; (3) apply proper statistical significance criteria to identify the gene expression profile, that is, a specific list of genes differentially expressed or associated with the phenotype; (4) investigate functions and pathways of the genes in the expression profile, and perform some sort of validation with wet-lab experiments, external data sets, permutation test, or cross validation. Although there are a number of statistical issues in each step, those in steps (2) and (3) are the topic of this paper.

The test of a probe for differential expression or association in step (2) is carried out by testing a statistical hypothesis properly formulated for the study. For example, gene expression profiling for comparison of normal versus a type of cancer cells would test if the mean or median expression level of each probe is the same in the two cell types (the null hypothesis) vs. the opposite (the alternative hypothesis). Gene expression profiling for association with a quantitative trait would use regression modeling appropriate for the phenotype. Because a statistical hypothesis is tested for each probe, and there are typically tens of thousands of probes, such analysis creates a problem of massive multiple hypothesis tests. It is then imperative to either control or effectively assess the levels of false positive (type-I) and false negative (type-II) errors in step (3) when statistical significance criteria are considered. The microarray gene expression applications have greatly stimulated the statistical research on the massive multiple hypothesis tests problem. There is now a giant body of literature in this area and basically five paradigms of massive multiple tests: control of the false discovery rate (FDR), estimation of FDR, significance threshold criteria, control of family-wise error rate (FWER) or generalized FWER (gFWER), and empirical Bayes approaches.

The traditional approaches to controlling the FWER have proven to be too conservative in applications of microarray data analyses. Recent attention has been focused on the control of FDR. A recent non-technical review of FDR methods is described elsewhere. **[5]** The goal of this paper is to provide an overview of a few advancements of the FDR-based inference and related methodology under a unified set of notation and assumptions pertinent to microarray gene expression applications, so as to reflect the essence of the current state of the field. With a representation of multiple hypotheses tests as an estimation problem, this paper provides a technical survey of the FDR paradigms commonly used in microarray gene expression data analyses. Section 2 contains a brief review of FDR and related error measurements for massive multiple tests;

Section 3 contains a review of FDR control procedures; Section 4 contains a review of the estimation of the proportion of true null hypotheses; Section 5 contains a review of FDR estimation methods; Section 6 contains a brief review of data-driven significance threshold criteria; Sections 7 contains a brief review of sample size determination for FDR control; and some concluding remarks are made in section 8. More application-oriented readers can read Section 8 first to get a non-technical summary of the issues and the state of the field, and then read Sections 2 through 7 to obtain more technical details.

The following notation will be used throughout. *R* denotes the real line, *m* denotes the number of tests (probes), and := indicates equal by definition. $I(\cdot)$ denotes the indicator function; it takes value 1 if the statement enclosed in the parentheses is true, 0 otherwise. Convergence and convergence in probability are denoted by $\rightarrow$ and $\rightarrow_p$ respectively. A random variable is usually denoted by an upper-case letter such as *P*, *R*, *V*, etc. A cumulative distribution function (cdf) is denoted by *F*, *G* or *H*; an empirical distribution function (EDF) is indicated by a tilde, e.g., $\tilde{F}$. A population parameter is denoted by a lower-case Greek letter and a hat indicates an estimator of the parameter, e.g., $\hat{\theta}$. Asymptotic equivalence is denoted by $\cong$: $a_n \cong b_n$ as $n \rightarrow \infty$ means $\lim_{n\rightarrow\infty} a_n / b_n = 1$.

2 False discovery rate and related error measurements:

Consider testing *m* hypothesis pairs (H_{0i}, H_{Ai}), $i = 1, \ldots, m$. In most applications of microarray gene expression analyses, *m* is typically on the order of 10^5–10^6. Suppose *m* P values, $P_1, \ldots, P_m$, one for each hypothesis pair, are calculated, and a decision on whether to reject H_{0i} is to be made. Let m_0 be the number of true null hypotheses, and let $m_1 := m - m_0$ be the number of true alternative hypotheses. The outcome of testing these *m* hypotheses can be tabulated as in Table 1. **[6]**

True Hypotheses	Rejected	Not Rejected	Total
H_0	V	m_0-V	m_0
H_A	S	m_1-S	m_1
Total	R	m-R	m

Table 1: Outcome tabulation of multiple hypotheses testing

Here *V* is the number of null hypotheses erroneously rejected, *S* is the number of alternative hypotheses correctly captured, and *R* is the total number of rejections. Conceptually these quantities are random variables. Clearly only *m* is known and only *R* is observable. An important parameter is m_0, or equivalently, the *null proportion* $\pi_0 := m_0/m$. This parameter will appear frequently in the subsequent sections, and its estimation will be discussed in **Section 4.**

Multiple hypotheses tests and related error measurements can be well understood as an estimation problem, which is described below in the frequentist framework. First for two probability distributions P_1 and P_2 on R with respective cumulative distribution function (cdf) $F_1(\cdot)$ and $F_2(\cdot)$, P_1 is said *stochastically less than* P_2, written as $\mathsf{P}_1 <_{st} \mathsf{P}_2$, if $F_1(t) \geq F_2(t)$ for all $t \in \mathsf{R}$. Next define the parameter $\Theta = [\theta_1, \ldots, \theta_m]$ as $\theta_i = 1$ if H_{Ai} is true, and $\theta_i = 0$ if H_{0i} is true ($i = 1, \ldots, m$). The data consist of the P values $\{P_1, \ldots, P_m\}$, and under the assumption that each test is exact and unbiased, the population is described by the following probability model:

$$P_i \sim \mathsf{P}_{i,\theta i}$$
$$\mathsf{P}_{i,0} \text{ is } U(0,1), \text{ and } \mathsf{P}_{i,1} <_{st} U(0,1); \quad (1)$$

each distribution $\mathsf{P}_{i,l}$ has a continuously differentiable cdf $F_i(\cdot)$, $i = 1, \ldots, m$. The P values are dependent in general and have a joint distribution on $[0, 1]^m$. The marginal cdf of P_i can be written as $G_i(t) = (1 - \theta_i)t + \theta_i F_i(t)$. Note $F_i(t) \geq t$ and $G_i(t) \geq t$ for $t \in [0, 1]$.

A test procedure is an estimator of Θ: $\hat{\Theta} = \hat{\Theta}(P_1, \ldots, P_m) = [\hat{\theta}_1, \ldots, \hat{\theta}_m] \in \{0, 1\}^m$, where $\hat{\theta}_i = 1$ indicates rejecting H_{0i} in favor of H_{Ai}, $i = 1, \ldots, m$. With this notation, the random variables in Table 1 can be expressed as

$$V = V_\Theta(\hat{\Theta}) = \sum_{i=1}^{m} (1-\theta_i)\hat{\theta}_i$$
$$S = S_\Theta(\hat{\Theta}) = \sum_{i=1}^{m} \theta_i \hat{\theta}_i \quad (2)$$
$$R = R(\hat{\Theta}) = \sum_{i=1}^{m} \hat{\theta}_i .$$

A natural and perhaps the simplest procedure is the "hard-thresholding" (HT) estimator $\hat{\Theta} = \hat{\Theta}(\alpha)$ defined as

$$HT(\alpha): \ \hat{\theta}_i = 1 \text{ iff } P_i \leq \alpha , \quad (3)$$

Where $\alpha \in (0, 1)$ is a significance threshold common to all tests. Clearly for this procedure the distributions of the random variables *V*, *S*, and *R* all depend on α.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Current Trends**

2.1 False discovery rate

At least one *family-wise type-I error* is committed if $V > 0$, and procedures for multiple hypothesis testing have traditionally been produced for solely controlling the family-wise type-I error probability $\Pr(V > 0)$. It is well-known that such procedures are often lack of statistical power. In an effort to develop more powerful procedures, **[6]** approached the multiple testing problem from a different perspective and introduced the concept of *false discovery rate* (FDR), which is, loosely speaking, the expected value of the ratio *V/R*. Rigorously, the FDR is defined as $FDR = E\left[V/R \mid R > 0\right]\Pr(R > 0)$. Note that if no alternative hypothesis is true, i.e., $m_0 = m$, then $V = R$ and $E[V/R|R>0] = 1$ with probability one; therefore $FDR = \Pr(V > 0)$, the family-wise type-I error probability.

$$FDR_{\Theta}(\hat{\Theta}) = E\left[\frac{\sum_{i=1}^{m}\hat{\theta}_i(1-\theta_i)}{\sum_{i=1}^{m}\hat{\theta}_i + \Pi_{i=1}^{m}(1-\hat{\theta}_i)}\right] \qquad (4)$$

Benjamini and Hochberg (1995) aim at determining an α based on the P values so that the FDR of the *HT*(α) procedure (3) is controlled below a *pre-specified* level. **[7]**

2.2 Positive FDR and *q*-value

For more discovery-oriented applications the FDR level is often not specified *a priori*, but rather determined after one sees the data (P values), and it is often determined in a way allowing for some "discovery" (rejecting one or more null hypotheses). Hence the *positive false discovery rate* (pFDR; **[7, 8]**, defined as $pFDR := E[V/R|R>0]$, is a more appropriate error measurement. Storey (2002) develops estimators of FDR and pFDR and introduces the concept of *q*-value in a Bayesian framework. **[7]** Assuming that each θ_i is a Bernoulli random variable with $\Pr(\theta_i = 1) = \Pr(H_{0i}) = 1 - \pi_0$ (prior probability), all test statistics have the same null distribution, all test statistics have the same alternative distribution, and all tests are performed with identical rejection regions **[7]**, the pFDR of the *HT*(α) procedure is $pFDR(\alpha) = \pi_0\alpha/\Pr(P \leq \alpha)$, where *P* is the random P value resulted from any test. Storey (2002) uses the phrase "identical tests" to describe the set of assumptions. **[7]**

To understand the *q*-value, first consider the P value. Suppose there are *m* two-sample Student-t tests with a common degrees of freedom *d* and observed statistics $t_1, \ldots, t_m$. For a single test, say the *i*th test, the P value is $P_i = \Pr_{H_{oi}}\left(\left|T_d\right| \geq \left|t_i\right|\right)$, where T_d is a random variable following the *t* distribution with *d* degrees freedom. If a threshold $t^* > 0$ is applied to make the decision whether to reject the null hypothesis, i.e., reject the *i*th null if and only if $|t_i| \geq t^*$ or equivalently, $|t_i|$ is in the rejection region $[t^*,\infty)$, then the P value at $|t_i|$ is $P_i = \inf_{t^* \geq |t_i|}\{\Pr H_{0i}\ (|T_d| \geq t^*)\}$, that is, the minimum probability over all the rejection regions less stringent than $|t_i|$ under the *i*th null hypothesis. Note the P value is defined for a single test. The *q*-value is defined for all *m* tests as a whole, using pFDR in lieu of the probability distribution under the null hypothesis. Storey (2002) gives a general definition of the *q*-value **[7]**; for the *HT*(α) procedure (3) the *q*-value at α is defined as $q(\alpha) := \inf_{\gamma \geq \alpha}\{pFDR(\gamma)\}$, and $q(\alpha) = \inf_{\gamma \geq \alpha}\{\pi_0\gamma/\Pr(P \leq \gamma)\}$ under the Bayesian model. So the *q*-value at α is the minimum pFDR over all the rejection regions less stringent than α. Thus the *q*-value is an error measurement related to the positive FDR, but it is neither the pFDR nor the FDR. The *q*-value can only be meaningfully interpreted in the Bayesian framework. **[7]** Storey (2003) shows that in the Bayesian framework the *q*-value $q(\alpha)$ can be interpreted as the posterior probability of the null hypothesis given $P \leq \alpha$. **[9]** Estimation of the pFDR and *q*-value will be reviewed in **Section 5.1.**

2.3 Erroneous rejection ratio

As discussed by Benjamini and Hochberg (1995, 2000), the FDR criterion has many desirable properties not possessed by other intuitive alternative criteria for multiple tests. **[6, 10]** However, methodological and theoretical developments and extensions of the FDR approach require to assume certain weak dependence conditions **[9, 11, 12]** or positive dependence structure **[13]** among the test statistics. These conditions may be too strong for genome-wide tests of gene expression–phenotype associations, in which a substantial proportion of the tests can be strongly dependent. **[14]** In such applications it may not be even reasonable to assume that the tests of the true null hypotheses are independent, an assumption often used in FDR research. Without these assumptions however, the FDR becomes difficult to handle analytically. Cheng (2006) defines an analytically simple error measurement in the same spirit of FDR **[15]**, called the *erroneous rejection ratio* (ERR): With notation given in Equation (2),

$$ERR_{\Theta}(\hat{\Theta}) := \frac{E[V_{\Theta}(\hat{\Theta})]}{E[R(\hat{\Theta})]}\Pr(R(\Theta) > 0). \qquad (5)$$

Just like FDR, when all null hypotheses are true $ERR = \Pr(R(\Theta) > 0)$, which is the family-wise type-I error probability because now $V_{\Theta}(\hat{\Theta}) = R(\hat{\Theta})$ with probability one. An advantage of ERR is that it can be handled under arbitrary dependent relationships among the tests; this will be elaborated later. Denote by $V(\alpha)$ and $R(\alpha)$ respectively the *V* and *R* random variables in Table 1 and by $ERR(\alpha)$ the ERR of the *HT*(α) procedure. Then

$$ERR(\alpha) = \frac{E[V(\alpha)]}{E[R(\alpha)]}\Pr(R(\alpha) > 0).$$

Let $FDR(\alpha) := E[V(\alpha) / R(\alpha)|R(\alpha) > 0]\ \Pr(R(\alpha) > 0)$. $ERR(\alpha)$ is essentially $FDR(\alpha)$. Under the hierarchical (or random effect) model employed in several papers **[7, 8, 9, 12, 16],** $FDR(\alpha) = ERR(\alpha)$ for all $\alpha \in (0, 1]$, following from Lemma 2.1 of Genovese and Wasserman (2004). **[12]** More generally $ERR/FDR = \{E[V]/E[R]\}/E[V/R|R > 0]$ provided

$\Pr(R > 0) > 0$. Asymptotically as $m \to \infty$, if $\Pr(R > 0) \to 1$ then $E[V/R|R > 0] \cong E[V/R]$; if furthermore $E[V/R] \cong E[V]/E[R]$, then $ERR / FDR \to 1$. The last condition is approximately satisfied for the $HT(\alpha)$ procedure if α is close to zero **[8]**, which is often true in microarray applications.

Similar to pFDR is the *positive ERR*, $pERR := E[V] / E[R]$. It is well-defined provided $\Pr(R > 0) > 0$. The relationship between pERR and pFDR is the same as that between ERR and FDR described above.

It is instructive to examine each component of $ERR(\alpha)$. Let $P_{1:m}$ be the smallest P value. First, under model (1)

$$E[V(\alpha)] = \sum_{i=1}^{m} (1-\theta_i) \Pr(\hat{\theta}_i = 1) = m_0 \alpha$$

$$E[R(\alpha)] = \sum_{i=1}^{m} \Pr(\hat{\theta}_i = 1) = m_0 \alpha + \sum_{j:\theta_j=1} F_j(\alpha)$$

$$\Pr(R(\alpha) > 0) = \Pr(P_{1:m} \le \alpha).$$

Define

$$F_m(t) := m^{-1} \sum_{i=1}^{m} G_i(t) = \pi_0 t + (1-\pi_0) H_m(t),$$

$$H_m(t) := m_1^{-1} \sum_{j:\theta_j=1} F_j(t),$$

$t \in \mathsf{R}$. Then

$$ERR(\alpha) = \frac{\pi_0 \alpha}{F_m(\alpha)} \Pr(P_{1:m} \le \alpha). \qquad (6)$$

Note the functions $F_m(\cdot)$ and $H_m(\cdot)$ both are cdf's with $F_m(0) = H_m(0) = 0$ and $F_m(1) = H_m(1) = 1$. $F_m(\cdot)$ is the average of all P value individual (marginal) cdf's. It describes the ensemble behavior of all P values, hence will be called the *ensemble P value cdf*. $H_m(\cdot)$ is the average of the P value marginal cdf's corresponding to the true alternative hypotheses, and describes the ensemble behavior of the P values corresponding to the true alternative hypotheses; hence will be called the *ensemble P value alternative cdf*. Next, these functions are linked to the actual data (i.e., observed P values) by the Empirical Distribution Function (EDF) of the P values defined as $\widetilde{F}_m(t) := m^{-1} \sum_{i=1}^{m} I(P_i \le t)$, $t \in \mathsf{R}$. Simple calculations show that under model (1)

$$E[\widetilde{F}_m(t)] = F_m(t) = \pi_0 t + (1-\pi_0) H_m(t), \; t \in [0,1]. \qquad (7)$$

This link provides opportunities to develop estimators of the FDR and data-driven significance criteria which will be reviewed in **Sections 4, 5, and 6.**

The false positive error behavior of a given multiple test procedure can be investigated in terms of either FDR (pFDR) or ERR (pERR). The ratio $pERR(\alpha) = E[V(\alpha)] / E[R(\alpha)]$ can be handled easily under arbitrary dependence among the tests because $E[V]$ and $E[R]$ are simply means of sums of indicator random variables. Cheng (2006) **[15]** develops a data-driven significance threshold criterion to determine an α for the hard-thresholding $HT(\alpha)$ procedure (3) so that its ERR and pERR are guaranteed to diminish asymptotically as the number of tests m goes to infinity, for arbitrarily dependent tests; see **Section 6.**

2.4 Other error measurements

The expected number of type-II errors (false negatives) is $E[m_1 - S]$. For the $HT(\alpha)$ procedure, under model (1) $E[m_1 - S] = m_1 - \sum_{i=1}^{m} I(\theta_i = 1) G_i(\alpha) = m_1 - m_1 Hm(\alpha)$. The *false negative proportion* is $m^{-1}E[m_1 - S] = (1 - \pi_0)(1 - H_m(\alpha))$. This quantity will be further considered in Section 6.2.

Symmetric to FDR, the false non-discovery rate (FNR) can be defined as $FNR = E[(m_1 - S) / (m - R) \mid R < m]$. **[11]**

Lehmann and Ramano (2005) introduced the *generalized family-wise error rate* (gFWER) which is $\Pr(V > k)$ for a specified k. **[17]** The traditional FWER corresponds to $k = 0$. In a series of papers van der Laan and colleagues develop resampling and augmentation procedures of controlling gFWER and the probability $\Pr(V/R > k)$ for a specified k.

3 FDR control:

3.1 The linear step-up procedure

Let $P_{1:m} \le P_{2:m} \le \cdots \le P_{m:m}$ be the order statistics of the P values, and let $\pi_0 = m_0/m$. Assuming that the P values corresponding to the true null hypotheses are independent, Benjamini and Hochberg (1995) prove that for any specified $q^* \in (0, 1)$, rejecting all the null hypotheses corresponding to $P_{1:m}, \ldots, P_{k^*:m}$ with $k^* = \max\{k : P_{k:m} / (k/m) \le q^*\}$ controls the FDR at the level $\pi_0 q^*$, i.e., $FDR_\Theta(\hat{\Theta}(P_{k^*:m})) \le \pi_0 q^* \le q^*$ in the notation given in Section 2. **[6]** Note this procedure is equivalent to applying the data-driven threshold $\alpha = P_{k^*:m}$ to all P values in (3), i.e., $HT(P_{k^*:m})$.

3.2 Adaptive FDR control

Recognizing the potential of constructing less conservative FDR control by the above procedure, Benjamini and Hochberg (2000) propose an estimator of m_0, $\hat{m}_0$, (hence an estimator of π_0, $\hat{\pi}_0 = \hat{m}_0 / m$), and replace k / m by $k / \hat{m}_0$ in determining k^*. **[10]** They call this procedure

adaptive FDR control. The estimator $\hat{\pi}_0 = \hat{m}_0 / m$ will be discussed in **Section 4.** A recent development of adaptive FDR control can be found in Benjamini *et al.* (2006). **[18]**

3.3 Another adaptive FDR control

Storey (2002) **[7]** considers the FDR estimator $\widehat{\text{FDR}}(\alpha) := \frac{\hat{\pi}_0(\lambda)\alpha}{\max\{R(\alpha),1\}/m}$ for a P value cut-off α, where $\hat{\pi}_0(\lambda)$ is an estimator of π_0 (See **Section 4.1**) and $R(\alpha)$ is the number of P values less than or equal to α. FDR control can be performed by "inverting" this estimator: for a given FDR level q^*, find the largest possible $\hat{\alpha}$ such that $\widehat{\text{FDR}}(\hat{\alpha}) \leq q^*$, and reject all the null hypotheses with $P \leq \hat{\alpha}$. This operation can be represented in a "q-value style". Let $q_i := \inf_{j \geq i}\{\widehat{\text{FDR}}(P_{j:m})\}$, $i = 1, \ldots, m$; then reject all the null hypotheses for which $q_i \leq q^*$. Storey *et al.* (2004) **[8]** show that using a slightly modified version of $\widehat{\text{FDR}}(\cdot)$ this procedure guarantees to control the FDR under q^* if the P values corresponding to the true null hypotheses are independent.

3.4 Dependent Tests

Storey *et al.* (2004) **[8]** show that if the P values are weakly dependent in the sense of being dependent in general but satisfying certain ergodicity conditions as $m \to \infty$, then the procedure is conservative in the limit in the sense that $\lim_{m\to\infty} \hat{\alpha} < \lim_{m\to\infty} \alpha^*$ where α* is the largest possible α such that the actual $FDR(\alpha) \leq q^*$.

Yekutieli and Benjamini (1999) **[19]** develop a resampling-based approach to FDR control for correlated tests. Qiu *et al.* (2005) **[20]** also describe the use of resampling to assess the stability of gene selection in microarray analysis. Benjamini and Yekutieli (2001) **[13]** show that the Benjamini and Hochberg (1995) **[6]** procedure controls the FDR if the test statistics satisfy the "positive regression dependence" condition. They also introduce a very conservative, but universal procedure that guarantees the FDR control for arbitrary P values (dependent or independent, discrete or continuous): control the FDR at the level $q^{**} = \left(\sum_{i=1}^{m}(1/i)\right)^{-1} q^*$ with the Benjamini and Hochberg (1995) **[6]** procedure guarantees to control the FDR at level q^* regardless dependence and/or discreteness of the P values.

In a series of papers, van der Laan and colleagues **[21, 22**] and Dutdoit and colleagues **[23]** developed procedures to control the gFWER for arbitrarily dependent tests.

4 Estimation of the null proportion:

Recall from Equation (8) that the EDF of the P values $\widetilde{F}_m(t)$ has expected value $E[\widetilde{F}_m(t)] = F_m(t)$ for every t; that is, $\widetilde{F}_m(\cdot)$ is an unbiased estimator of the P value ensemble cdf $F_m(\cdot)$. Cheng *et al.* (2004) **[24]** observe that if the tests $\hat{\theta}_i$ ($i = 1, \ldots, m$) are not too much correlated asymptotically in the sense $\sum_{i=j} Cov(\hat{\theta}_i, \hat{\theta}_j) = o(m^2)$ as $m \to \infty$, $\widetilde{F}_m(\cdot)$ is "asymptotically consistent" for $F_m(\cdot)$ in the sense $\left|\widetilde{F}_m(t) - F_m(t)\right| \to_p 0$ for every $t \in$ IR. These results provide heuristics for the estimation of π_0, the estimation of FDR, and data-adaptive determination of α for the $HT(\alpha)$ procedure. Estimation of π_0 is reviewed in this section.

As noted in the previous sections, the proportion of the true null hypotheses π_0 is an important parameter in FDR-related procedures. Consider first the P value ensemble cdf $F_m(\cdot)$. Because for any $t \in (0, 1)$ $\pi_0 = [H_m(t) - F_m(t)] / [H_m(t) - t]$, a plausible estimator of π_0 is $\hat{\pi}_0 = \frac{\Lambda - \widetilde{F}_m(t_0)}{\Lambda - t_0}$ for properly chosen Λ and t_0. The inverse function of $F_m(\cdot)$, defined as $Q_m(u) := F_m^{-1}(u) := \inf\{t : F_m(t) \geq u\}$, is the *P value ensemble quantile function*. The sample version is the *empirical quantile function* (EQF) defined as $Q_m(u) := \widetilde{F}_m^{-1}(u) := \inf\{x : \widetilde{F}_m(x) \geq u\}$. Then $\pi_0 = [H_m(Q_m(u)) - u] / [H_m(Q_m(u)) - Q_m(u)]$, for $u \in (0, 1)$, and with Λ_1 and u_0 properly chosen, $\hat{\pi}_0 = \frac{\Lambda_1 - u_0}{\Lambda_1 - Q_m(u_0)}$ is a plausible estimator. Many of the estimators take either of the above two basic representation with some modifications.

Clearly it is necessary to have $\Lambda_1 \geq u_0$ in order to have a meaningful estimator. Because $Q_m(u_0) \leq u_0$ by the stochastic order assumption [cf. (1)], choosing Λ_1 too close to $u0$ will produce an estimator much biased downward. A heuristic is that if u_0 is so chosen that all P values corresponding to the alternative hypotheses concentrate in the interval $[0, Q_m(u_0)]$ then $H_m(Q_m(u_0)) = 1$; thus setting $\Lambda_1 = 1$. A similar heuristic leads to setting $\Lambda = 1$.

4.1 Slope estimator

Taking a graphical approach Schweder and Spjøtvoll (1982) **[25]** consider the slope from the point $(\lambda, \widetilde{F}_m(\lambda))$ to the point (1,1), and an estimator of m_0 as $\hat{m}_0 = m(1 - \widetilde{F}_m(\lambda)) / (1 - \lambda)$ for a properly chosen λ; hence a corresponding estimator of π_0 is $\hat{\pi}_0(\lambda) = \hat{m}_0 / m = (1 - \widetilde{F}_m(\lambda)) / (1 - \lambda)$. Storey's (2002) **[7]** estimator is exactly this one. Additionally, Storey (2002) **[7]** observes that λ is a tuning parameter that dictates the bias and variance of the estimator, and proposes computing $\hat{\pi}_0$ on a grid of λ

values, smoothing them by a spline function, and taking the smoothed $\hat{\pi}_0$ at a λ close to 1, (e.g. 0.95) as the final estimator. Storey *et al.* (2003) **[8]** propose a bootstrap procedure to estimate the mean-squared error (MSE) and pick the λ that gives the minimal estimated MSE; a simulation study in Cheng (2006) **[15]** and investigation in Langaas *et al.* (2005) **[26]** show that this estimator tends to be biased downward.

4.2 Quantile slope estimator

Approaching to the problem from the quantile perspective Benjamini and Hochberg (2000) **[10]** propose $\hat{m}_0 = \min\{1 + m + 1(m+1-j)/(1-P_{j:m}), m\}$ for a properly chosen j; hence $\hat{\pi}_0 = \hat{m}_0/m$. The index j is determined by examining the slopes $S_i=(1-P_{i:m})/(m+1-i)$, $i = 1, \ldots, m$, and is taken to be the smallest index such that $S_j < S_{j-1}$. Then $\hat{m}_0 = \min\{1+1/S_j, m\}$. Cheng (2006) **[15]** shows that as m gets large the event $\{S_j < S_{j-1}\}$ tends to occur early (i.e., at small j) with high probability; therefore the estimator tends to be increasingly conservative (i.e., biased upward) as the number of tests m increases. The conservativeness is also demonstrated by the simulation study in Cheng (2006). **[15]**

4.3 Quantile slope estimator by quantile modeling

Cheng (2006) **[15]** develops an improvement of Benjamini and Hochberg's' (2000) **[10]** estimator by considering a shape requirement on the P value ensemble quantile function $Q_m(\cdot)$. Heuristically, the stochastic order requirement in model (1) implies that $F_m(\cdot)$ is approximately concave and hence $Q_m(\cdot)$ is approximately convex. When there is a substantial proportion of true null and true alternative hypotheses, there is a "bend point" $\tau_m \in (0, 1)$ such that $Q_m(\cdot)$ assumes roughly a nonlinear shape on the interval $[0, \tau_m]$, primarily dictated by the distributions of the P values corresponding to the true alternative hypotheses, and $Q_m(\cdot)$ is essentially linear on the interval $[\tau_m, 1]$, primarily dictated by the $U(0, 1)$ distribution of the null P values. The estimation of π_0 can benefit from properly capturing this shape characteristic using a model. Cheng (2006) **[15]** considers a two-piece function approximation (model) for $Q_m(\cdot)$. In an interval $[0, \tau_m]$ $Q_m(u)$ is approximated by a polynomial of the form $\eta u^\gamma + \delta u$ with $\gamma \geq 1$, $\eta \geq 1$, and $0 \leq \delta \leq 1$; on the interval $[\tau_m, 1]$ it is approximated by a linear function $\beta_0 + \beta_1 u$ with $\beta_0 \leq 0$ and $\beta_1 \geq 1$. The two pieces are joint smoothly at τ_m by the constraints $\eta\tau_m^\gamma + \delta\tau_m = \beta_0 + \beta_1\tau_m$ (continuity) and $\eta\gamma\tau_m^{\gamma-1} + \delta = \beta_1$ (differentiability). For identifiability it is further required that $\gamma = \eta = 1$ and $\delta = 0$ if and only if $\tau_m = 0$. These parameters are determined by minimizing the integrated absolute difference (L^1 distance) between $Q_m(u)$ and

$$Q_m^*(u) := I(0 \leq u \leq \tau_m)(\eta u^\gamma + \delta u) + I(\tau_m \leq u \leq 1)(\beta_0 + \beta_1 u),$$

subject to the above constraints. Cheng (2006) develops a procedure to estimate these parameters from the P value EQF $\widetilde{Q}_m(\cdot)$. The estimator of π_0 is the reciprocal of the estimator of β_1: $\hat{\pi}_0 := 1/\hat{\beta}_1$.

A simulation study by Cheng (2006) **[15]** indicate that in a reasonably wide range of scenarios this estimator is slightly biased upward (i.e., conservative); the upward bias is usually less than the downward bias of the bootstrap estimator of Storey *et al.* (2003), **[8]** and is much less than the upward bias of Benjamini and Hochberg (2000) **[10]** estimator. In this regard this quantile slope estimator outperforms the other two estimators, as well as in terms of the mean square error.

4.4 Monotone convex and smooth density estimators

Note that under model (1) the probability density function (pdf) of $F_m(\cdot)$, the P value ensemble pdf, is

$$f_m(t) := \frac{d}{dt}F_m(t) = \pi_0 + (1-\pi_0)h_m(t),\ t \in [0,1],$$

where $h_m(t) := \frac{d}{dt}H_m(t)$, the P value ensemble alternative pdf. Note $\pi_0 \approx f_m(1)$ if $h_m(1) \approx 0$; this is achievable under the heuristic that essentially all the P values corresponding to the true alternative hypotheses concentrate in an interval away from 1. Langaas *et al.* (2005) **[26]** consider estimating π_0 by requiring $F_m(\cdot)$ be strictly concave and thus $f_m(\cdot)$ be monotone and convex. They propose to estimate $f_m(\cdot)$ by the nonparametric maximum likelihood estimator $\hat{f}_m^*(\cdot)$ under the constraint of monotonicity and convexity, and to estimate π_0 by $\hat{\pi}_0 := \hat{f}_m^*(1)$. The simulation study therein indicates this estimator performs very well in a range of scenarios.

Cheng *et al.* (2004) **[24]** consider a spline function estimator $\hat{F}_m(\cdot)$ of $F_m(\cdot)$. $\hat{F}_m(\cdot)$ is a B-spline function constructed by smoothing the P value EDF $\widetilde{F}_m(\cdot)$. The spline knots are placed in a way that gives little smoothing in the vicinity of 0 but a large amount of smoothing in the right tail. An estimator of $f_m(\cdot)$ is the derivative function $\hat{f}_m(1) := \frac{d}{dt}\hat{F}_m(t), t \in [0,1]$. Then an estimator of π_0 is given by $\hat{\pi}_0 := \hat{f}_m(1)$. The simulation study in Cheng *et al.* (2004) **[24]** indicate that this estimator is slightly upward biased (conservative) in a range of scenarios as long as the true π_0 is not too close to 1.

4.5 Mixture model estimators

Allison *et al.* (2002) **[27]** and Pounds and Morris (2003) **[28]** describe methods that estimate the FDR via P value modeling. These methods also estimate π_0. Allison *et al.*

(2002) **[27]** describe a method that models the P values as arising from a mixture distribution with one *U*(0, 1) component and potentially several beta components. The model is fit by maximum likelihood estimation and the bootstrap is used to determine the number of beta components that are used in the model. Allison *et al.* (2002) **[27]** note that it is often unnecessary in practice to include more than one beta component in the model. Pounds and Morris (2003) **[28]** give a detailed description of the use of a specific model with one beta component. Assuming null p-values follow a U(0,1) distribution, Pounds and Morris (2003) **[28]** show that π_0 must be less than or equal to the minimum of the ensemble P value pdf. Thus, they propose to estimate π_0 by the minimum of the pdf of the mixture model fit to the p-values. Allison *et al.* (2002) **[27]** estimate π_0 by the mixing weight for the uniform component of the fitted model. It is theoretically possible that the mixing weight of the uniform component could be substantially smaller than the minimum of the fitted pdf. In this case, the mixing weight estimator understates the proportion of the fitted density that could be attributed to a uniform (0,1) distribution.

4.6 Moment estimator

Pounds and Cheng (2006) **[29]** describe a simple moment-based estimator of π_0. Let $\overline{P} = m^{-1}\sum_{i=1}^{m} P_i$. Assuming that $E[P_i] \geq 1/2$ if $\theta_i = 0$ (i.e., H_{0i} is true), it follows that $E\left[\overline{P}\right] \geq 2\pi_0$. This observation motivates $\hat{\pi}_0 = \min(1, 2\overline{P})$ as an estimator of π_0. This estimator has several advantages over those described above. It is very simple to compute, and it does not rely on continuity or model assumptions for the P values. However, it is considerably more conservative than the other estimators when the assumptions of those estimators hold.

5 FDR estimation:

As mentioned in **Section 2.2,** for discovery-oriented applications the FDR level often is not specified *a priori*, but rather determined *a posteriori*, and it is often determined in a way allowing for some "discovery" (rejecting one or more null hypotheses). Hence an alternative to FDR control is estimation of FDR and pFDR.

5.1 EDF-based estimators

Recall for **Section 3.3** that Storey (2002) **[7]** considers estimating the FDR for a fixed P value cut-off α by

$\widehat{\mathrm{FDR}}(\alpha) := \dfrac{\hat{\pi}_0(\lambda)\alpha}{\max\{R(\alpha),1\}/m}$, where $\hat{\pi}_0(\lambda)$ is an estimator of π_0 (See Section 4.1) and $R(\alpha)$ is the number of P values less than or equal to α. In term of the P value EDF,

$\widehat{\mathrm{FDR}}(\alpha) := \dfrac{\hat{\pi}_0(\lambda)\alpha}{\max\{\widetilde{F}_m(\alpha),1/m\}}$. Storey *et al.* (2003) **[8]** show that this estimator is biased upward and asymptotically conservative in the sense that with probability 1 $\lim_{m\to\infty} \inf_{t\geq\alpha}\{\widehat{\mathrm{FDR}}(t) - FDR(t)\} \geq 0$ for each α > 0. Storey (2002) considers an estimator of the pFDR given by

$p\widehat{\mathrm{FDR}}(\alpha) := \dfrac{\hat{\pi}_0(\lambda)\alpha}{\max\{R(\alpha),1\}/m][1-(1-\alpha)^m]}$. **[7]** In term of the P value EDF,

$p\widehat{\mathrm{FDR}}(\alpha) = \dfrac{\hat{\pi}_0(\lambda)\alpha}{\max\{\widetilde{F}_m(\alpha),1/m\}[1-(1-\alpha)^m]}$.

Hence $\lim_{\alpha\to 0} p\widehat{\mathrm{FDR}}(\alpha) = \lim_{\alpha\to 1} p\widehat{\mathrm{FDR}}(\alpha) = \hat{\pi}_0(\lambda)$ for any fixed $m > 1$, and in general $p\widehat{\mathrm{FDR}}(\alpha)$ is not monotone in α. Storey (2002) **[7]** establishes mean-squared error properties of this estimator and its asymptotic conservativeness that with probability 1, $\lim_{m\to\infty} p\widehat{\mathrm{FDR}}(\alpha) \geq pFDR(\alpha)$. It not difficult to see that with the multiplier $1-(1-\alpha)^m$ in its denominator this estimator may tend to have large variance (thus be unstable) for small α.

The "empirical" *q*-values are defined as $\hat{q}_i := \hat{q}(P_{i:m}) := \min_{j\geq i}\{p\widehat{\mathrm{FDR}}(P_{j:m})\}$, $i = 1, \ldots, m$. **[7]** Clearly $\hat{q}_1 \leq \cdots \leq \hat{q}_m$. Storey *et al.* (2003) **[8]** consider the more general *q*-value estimator $\hat{q}(\alpha) := \inf_{s\geq\alpha}\{p\widehat{\mathrm{FDR}}(s)\}$ for $q(\alpha)$ defined in **Section 2.2,** and show its conservativeness that $\lim_{m\to\infty} \inf_{t\geq\alpha}\{\hat{q}(t) - q(t)\} \geq 0$ with probability 1 for each $\alpha > 0$ under a specific Bayesian model (see section 2.2) and certain ergodicity conditions.

5.2 Smooth ensemble cdf and pdf estimator

Cheng *et al.* (2004) **[24]** consider an estimator of the FDR of the *HT*(α) procedure (3) by $\widehat{\mathrm{FDR}}(\alpha) = \dfrac{\hat{\pi}_0\alpha}{\hat{F}_m(\alpha)}$, where $\hat{\pi}_0$ and $\hat{F}_m(\cdot)$ are respectively the estimators of π_0 and the P value ensemble cdf $F_m(\cdot)$, derived from a spline smoothing of the P value EDF $\widetilde{F}_m(\cdot)$; see **Section 4.4.** Cheng *et al.* (2004) **[24]** consider using this estimator to provide an FDR estimate at a P value cut-off threshold $\hat{\alpha}$ generated by a data-driven significance criterion **(see Section 6)**. Simulation results therein indicate that the estimator is able to provide a reasonably conservative (upward biased) FDR estimate at the data-driven significance threshold in a wide range of scenarios.

Pounds and Cheng (2004) **[30]** propose an estimator of the P value ensemble pdf $f_m(\cdot)$ by properly transforming and smoothing a histogram constructed from the spacings defined by the ordered P values. An estimator $\hat{f}_m(\cdot)$ is constructed by back-transforming and normalizing the smooth function, an estimator of π_0 by

$\hat{\pi}_0 := \min\{\hat{f}_m(P_i),\ i = 1, \ldots, m\}$, an estimator $\hat{F}_m(\cdot)$ of $F_m(\cdot)$ by the trapezoid rule of integration applied to $\hat{f}_m(\cdot)$, and an FDR estimator by plugging the estimators into the above formula. Simulation results therein indicate that this estimator performs well in estimating the cFDR (or pFDR). For estimating pFDR it is much more stable (i.e., having less variance) than Storey's (2002) **[7]** estimator at the α values close to zero, which are often used in microarray applications.

5.3 Mixture model estimator

For the mixture models discussed in section 4.5, the FDR estimate is determined by substituting the fitted model's π_0 estimate and cdf into $\pi_0\alpha / F_m(\alpha)$. For the specific model of Pounds and Morris (2003), **[28]** the FDR estimate monotonically increases as α increases.

5.4 Robust estimator

As previously described, most of the available FDR estimation methods assume that $G_i(t) = t$ when $\theta_i = 0$ (i.e., H_{0i} is true). Pounds and Cheng (2006) **[29]** noted that this critical assumption is violated by discrete P values and P values from testing one-sided hypotheses. In particular, any test $\hat{\theta}_i$ that is one-sided or based on a discrete test statistic may have $G_i(t) < t$ for some t when $\theta_i = 0$. This violation can have severe and undesirable consequences for methods that estimate π_0 as part of their calculations. Pounds and Cheng (2005, 2006) **[31, 29]** describe these consequences in greater detail. Thus, Pounds and Cheng (2006) **[29]** develop a robust FDR estimator. The robust FDR estimator is conservative provided that $\Pr\left(\overline{P} \le 1/2\right) \approx 1$ and $G_i(t) \le t$ for $\theta_i = 0$, even when applied to one-sided tests or discrete P values. The method borrows ideas from least trimmed squares **[32]** and rank regression **[33]** to smooth raw FDR estimates obtained from the P value EDF. For one-sided tests, a folding transformation is used to make p-values essentially two-sided for purposes of estimating π_0 and then other calculations are performed on the original one-sided p-values.

5.5 Estimation of local FDR or empirical Bayes posterior

With estimators $\hat{\pi}_0$ and $\hat{f}_m(\cdot)$, an estimator of the empirical Bayes posterior probability (EBP or local FDR) **[16, 34]** of the null hypothesis H_{0i} conditional on $P_i = p_i$ is given by $\hat{\pi}_0 / \hat{f}_m(p_i)$. Efron (2004) **[34]** advocates to estimate the null ensemble density function of the test statistics from the empirical distribution and cautions against the use of random permutations. In the P value domain, this means to estimate the P value ensemble distribution under the "grand null" $H_0^* = \cap_{i=1}^m H_{0i}$ in lieu of assuming the $U(0, 1)$ distribution as in model (1). In a similar spirit, Datta and Datta (2005) **[35]** proposed an empirical Bayes method that first transforms p-values using the quantile-function of the standard normal distribution and then apply kernel density estimation methods to the transformed P values to obtain an EBP.

6 Data driven significance criteria:

6.1 Profile information criteria

Abramovich *et al.* (2000) **[36]** consider theoretically thresholding estimators of a sequence of Normal distribution means, where the threshold is determined by a lack of fit criterion (l^p distance) penalized by FDR. They show that the estimators are asymptotically minimax. Regarding massive multiple tests as the estimation problem described in Section 2, Cheng *et al.* (2004) **[24]** develop criteria to determine the significance threshold α for the $HT(\alpha)$ procedure (3). The *profile information* (I_p) criterion consists of a lack-of-fit term of the P value ensemble quantile function from $U(0, 1)$ penalized by the expected number of false discoveries under model (1). Empirically, the lack-of-fit term is defined by $\tilde{D}(\alpha) = \sqrt{m}\left\{\int_0^\alpha [t - \tilde{Q}_m(t)]_+^2 dt\right\}^{1/2}, \alpha \in (0,1],$

where $\tilde{Q}_m(\cdot)$ is the P value EQF (cf. Section 4) and $[x]_+$ denotes the positive part of x, i.e., $[x]_+ = \max\{x, 0\}$. So $\tilde{D}(\alpha)$ measures how far are the P value sample quantiles below the diagonal line on the interval $(0, \alpha]$. Empirically the profile information criterion I_p is given by

$$\tilde{I}_p(\alpha) = [\tilde{D}(\alpha)]^{-1} + \lambda(m, \hat{\pi}_0)\, m\, \hat{\pi}_0\, \alpha, \qquad \alpha \in (0, 1)$$

Here $m\hat{\pi}_0\alpha$ is an estimate of the expected number of false positives, $\lambda(m, \hat{\pi}_0)$ is a penalty factor, and $[\tilde{D}(\alpha)]^{-1}$ measures the deviation of the P values from the $U(0, 1)$ distribution. The more concentrated are the P values towardzero, the larger is $[\tilde{D}(\alpha)]$ and thus the smaller is $[\tilde{D}(\alpha)]^{-1}$; therefore one minimizes $\tilde{I}_p(\alpha)$ with respect to α. So the data-driven "optimal" significance threshold is the $\hat{\alpha}^*$ that minimizes $\tilde{I}_p(\alpha)$; and the $HT(\hat{\alpha}^*)$ procedure rejects H_{i0} if $Pi \le \hat{\alpha}^*$. Cheng (2006) **[15]** extends I_p by introducing the *adaptive profile information* (API) criterion based on the quantile model $Q_m^*(u) = I(0 \le u \le \tau)(nu^\gamma + \delta u) + I(\tau_m \le u \le 1)(\beta_0 + \beta_{1u})$ (cf. Section 4.3). API is defined as

$$API(\alpha) := \left[\int_0^{\alpha} (t - Q_m^*(t))^{\gamma} \, dt \right]^{-1/\gamma} + \lambda(m, \pi_0, \delta) m \pi_0 \alpha,$$

$$\alpha \in (0, 1)$$

Here the major modification is on the lack-of-fit term: the L^2 norm is replace by the L^{γ} norm. Recall that $\gamma \geq 1$ is a parameter reflecting how far the P value quantiles are below the $U(0, 1)$ quantiles in the vicinity of zero. The L^{γ} norm emphasis this deviation and makes the criterion more adaptive to the P value behavior around zero. Cheng (2006) **[15]** considers an approximation of the lack-of-fit term that simplifies both theoretical development and computation in practice, and proposes a procedure to estimate the parameters in API. The data-driven optimal significance threshold $\hat{\alpha}^*$ is the α that minimizes an approximate API with estimated parameters in $Q_m^*(\cdot)$.

A key issue is the choice of the penalty factor λ. Cheng *et al.* (2004) **[24]** and Cheng (2006) **[15]** consider a few conservative choices and show for $\pi_0 < 1$ the pERR of the $HT(\hat{\alpha}^*)$ procedure (3) diminishes to zero as $m \to \infty$ regardless the dependence among the P values; and for $\pi_0 \leq 1$ the ERR diminishes to zero as $m \to \infty$ if the P values posses certain dependence structure. The simulation studies therein indicate that these choices perform well when there is substantial power to reject the false null hypothesis in a number of individual tests, and they tend to be conservative when the power is low. Moreover, in a range of scenarios API moderately outperforms I_p.

6.2 Total error proportion

Pounds and Morris (2003) **[28]** observe that given a threshold α, the area under the P value density function can be partitioned into four distinct regions corresponding to the four hypothesis testing outcomes resulted from the $HT(\alpha)$ procedure (3). More specifically, the area to the left of α corresponds to rejections and the area below π_0 can be attributed to the $U(0, 1)$ distribution. Thus, assuming that the null distribution of the P values is $U(0, 1)$, the area left of α and below π_0 corresponds to Type I errors, the area left of α and above π_0 corresponds to correct rejections, the area above π_0 and right of α corresponds to Type II errors, and the area below π_0 and right of α corresponds to correct non-rejections. In particular, under model (1) the expected proportion of tests resulting in a Type I error is given by $FP(\alpha) = \pi_0 \alpha$. Additionally, the expected proportion of tests resulting in a Type II error is given by $FN(\alpha) = (1 - \pi_0)(1 - Hm(\alpha))$. The *total error proportion* is the sum $TE(\alpha) = FP(\alpha) + FN(\alpha)$, which is the expected proportion of tests resulting in a Type I or Type II error. Cheng *et al.* (2004) **[24]** use the term "total error criterion" and Genovese and Wasserman (2002) **[11]** use the term "total misclassification risk" to describe the total error proportion.

In practice, an estimate of the total error proportion can be used as a criterion to guide the selection of α. An estimate of TE(α) can be obtained by substituting estimates for the terms in FP and FN. Then, the value of α that minimizes this TE estimate can be easily determined. The TE estimators can be nonparametric **[24]** or parametric with the mixture models. **[27, 28]** Let $\hat{\alpha}_{TE}$ be the α so obtained.

Using $\hat{\alpha}_{TE}$ to declare significance has some useful operating characteristics. First, if the estimate of $F_m(\cdot), \hat{F}_m(t) = t$ for all t (indicating an all null case), then $\hat{\alpha}_{TE} = 0$ (no rejections are made). Additionally, $\hat{\alpha}_{TE}$ corresponds to a 50% empirical Bayes probability that the null hypothesis is true. **[28]**

7 Sample size determination for FDR control:

Several methods have been proposed to perform power and sample size calculations for a microarray study that will use FDR-type measures of significance in the final analysis. **[37, 38]** However, most of hese methods are designed only for two-group designs, such as studies that compare tumor expression to normal expression. Pounds and Cheng (2005) **[39]** describe a general method to perform power and sample-size calculations for tudies that will use the FDR to determine significance in the final analysis. For $i = 1, \ldots, m$, their method assumes hat cumulative distribution function of P_i, $G_i(\cdot; \delta_i, n)$ can be computed given the sample size n and an effect size i. Their method uses the *average power* (AP) as a measure of statistical power. The average power is defined as the arithmetic average of the powers of the tests with a true alternative. Under model (1), the average power is simply $AP(\alpha) := H_m(\alpha; \Delta, n)$.

The sample size determination procedure uses the *anticipated false discovery rate* (aFDR), $aFDR(\alpha, \Delta, n) := \hat{E}(\hat{\pi}_0; \Delta, n)\alpha / F_m(\alpha; \Delta, n)$ to perform its calculations. The ensemble P value cdf $F_m(\cdot; \Delta, n)$ is either postulated or estimated from preliminary data. The method is designed to determine the sample size necessary to achieve an average power of γ while keeping the aFDR below τ. The values of γ and τ must be chosen by the user. The method proceeds iteratively. With an initial sample size n_0 and a specified value or estimate for Δ, the procedure first finds α^* such than $AP(\alpha^*) = \gamma$. Then, it computes $aFDR(\alpha^*)$. If $aFDR(\alpha^*) \leq \tau$, then the procedure reports that n_0 is an adequate sample size to achieve average power γ while keeping the aFDR below τ. Otherwise, it increments n and repeats the calculations. The process is iterated until a maximum sample size is reached or the conditions for the aFDR and AP are satisfied. Pounds and Cheng (2005) **[39]** also describe a method to estimate necessary parameters from pilot data. They observed that the parameter-estimation and sample-size calculation method performed well in traditional simulation studies and in resampling-based simulation studies performed using real data.

8 Conclusion:

We have reviewed a few massive multiple hypothesis test paradigms related to FDR. This is by no means an exhaustive survey; other variations on the theme can be easily found in the literature, but the essence of the current state of the field has been well reflected.

For applications, it is not advisable to totally ignore the mathematical development of a concept. For example, an empirical *q*-value is often misinterpreted as an (frequentist) estimate of the FDR at the corresponding P value; whereas in fact it is not. A *q*-value is meaningful only under a specific Bayesian framework regarding the hypotheses random, then it is the probability of the corresponding null hypothesis given the data (observed P values); and the empirical *q*-value is an estimate of this probability. In their theoretical development, Storey (2002) **[7]** and Storey et al. (2003) **[8]** did not demonstrate that the empirical *q*-values can be used as (frequentist) FDR control quantities, but did demonstrate that the empirical *q*-values are conservative estimators of the population *q*-values (cf. Sections 2.2 and 5.1). Additionally, there has been empirical evidence that regarding the empirical *q*-values as FDR estimates gives downward-biased estimators; see the simulation study in Pounds and Cheng (2004). **[30]**

There are numerous methods for FDR control and FDR estimation. Thus, selecting a reasonable procedure for a specific application can be challenging. Pounds (2006) **[5]** notes that the choice can be simplified by a few basic application-specific considerations: whether FDR estimation or control is preferred, whether the p-values are onesided or discrete, and the correlation among p-values. In studies designed with adequate power for a pre-specified FDR coltrol level (Section 7), FDR-control procedures (Section 3) should be used because in these settings they typically offer greater power than do FDR-estimation methods. For undesigned (retrospective or observational) studies however, it is not always clear what an appropriate FDR control level should be. FDR-estimation methods (Sections 5 and 4) are preferred because interpreting the results of FDR-control procedures as FDR estimates can under-represent the actual prevalence of false positives. The data-driven significance threshold criteria (Section 6) can provide a rough guideline for the P value cut-off or FDR level to consider, and obtaining estimates of the FDR and pFDR at the data-driven significance threshold should be a part of the analysis. The sidedness, discreteness, and correlation of P values are important considerations to guide the selection of a method. Several methods have been shown to maintain their desirable statistical properties under mild or limited dependence among tests. Other methods have been developed to address strong or extensive dependence between the tests. Some methods implicitly assume that P values are continuously distributed in estimating π_0; these methods perform very poorly when applied to discrete P values. **[29]** Additionally, Pounds and Cheng (2006) observed that one-sided P values may be stochastically greater than uniform under the null hypothesis, thus violating the assumption that P values are uniformly distributed under the null. **[29]** This violation can cause such methods to perform in unpredictable and undesirable ways. Thus, Pounds and Cheng (2006) **[29]** developed an FDR-estimation method specifically for applications involving P values that are discrete or one-sided.

Certainly, there are still a number of open questions in the field. An important question is whether the correlation structure of microarray data satisfies the conditions required for the procedures to maintain their stated statistical properties. In our view, the answer to this question is likely to be specific to the particular application and methods under consideration. Yakovlev and colleagues **[20, 40]** have used resampling and permutation techniques to study the performance of several FDR procedures for the analysis of a data set of gene expression in pediatric acute lymphoblastic leukemia. **[41]** They found that gene-gene correlations may induce a high degree of variability in the number of rejections of many FDR procedures. However, by applying similar techniques to a data set of expression pediatric acute myeloid leukemia **[42]**, we observed that our robust FDR estimation method performs quite well. **[29]** Subsequently, we believe it would be useful to develop tests that determine whether a data set provides significant evidence of departure from the assumptions of specific methods. Such tests could be helpful for determining when computationally-intensive resampling methods (**[19, 20]**) are required.

Most research has focused on controlling or estimating the expected value of the ratio *V/R* or similar quantities. Future work should also attempt to estimate or control the variance of *V/R*; Owen (2005) **[43]** has done some initial work on this topic. As previously mentioned, some empirical studies of real microarray data sets have found that the variance in the number of rejections determined by multiple-testing procedures can be quite large. **[20]** This observation indicates that the interpretation of analysis results should be tempered by consideration of the variability of the FDR estimation or control procedures. Thus, it would be useful to develop procedures that also consider the variance of *V/R*. Additionally, incorporating variance considerations into the procedures may lead to interval estimates for the FDR. Storey (2002) **[7]** has mentioned that bootstrapping the P values is potentially one way to construct such an interval estimate.

It is also important to compare procedures performances against one another. So far, little effort has been invested in learning which methods are best suited for settings. Considering the biological and technical complexity of microarray data, it is unlikely that the assumptions of any method will strictly hold for any application. Certainly, no method will be superior across all applications. Thus, it is important to identify which procedures are best suited for use in certain sets of conditions. This research would likely involve a lengthy series of traditional simulation studies

and simulation-like studies performed by resampling, perturbing, or permuting numerous real data sets.

Acknowledgment:
This research is supported in part by the NIH/NIGMS Pharmacogenetics Research Network and Database (U01 GM61393, U01GM61374, http://pharmgkb.org/) from the National Institutes of Health (CC), Cancer Center Support Grant P30 CA-21765 (CC, SP), and the American Lebanese and Syrian Associated Charities (ALSAC).

References:

[01] L. Bullinger, *et al., N Engl J Med.,* 350:1605 (2004) [PMID: 15084693]
[02] A. Holleman, *et al., N Engl J Med.,* 351:533 (2004) [PMID: 15295046]
[03] C. Flotho, *et al., Blood.,* 108:1050 (2006) [PMID: 16627760]
[04] R. Simon, *J Clin Oncol.,* 23:7332 (2005) [PMID: 16145063]
[05] S. Pounds, *Briefings in Bioinformatics*., 7:25 (2006)
[06] Y. Benjamini & Y. Hochberg, *J R Stat Soc – B.,* 57:289 (1995)
[07] J. D. Storey, *J R Stat Soc – B.*, 64:479 (2002)
[08] J. D. Storey, *et al.*, *J R Stat Soc – B.*, 66:187 (2004)
[09] J. D. Storey, *Ann Stat.,* 31:2103 (2003)
[10] Y. Benjamini & Y. Hochberg, *J Educ Behav Stat.,* 25:60 (2000)
[11] C. Genovese & L. Wasserman, *J R Stat Soc – B.,* 64:499 (2002)
[12] C. Genovese & L. Wasserman, *Ann Stat.,* 32:1035 (2004)
[13] Y. Benjamini & D. Yekutieli, *Ann Stat.,* 29:1165 (2001)
[14] L. Klebanov, *et al., Stat Appl Gen Mol Biol.,* 5:7 (2006) [PMID: 16646871]
[15] C. Cheng, *Optimality: 2nd Erich L. Lehmann Symp – IMS Lect Notes Monogr.,* 49:51 (2006)
[16] B. Efron, *et al., J Am Stat Assoc.,* 96:1151 (2001)
[17] E. Lehmann & J. Ramano, *Ann Stat.,* 33:1138 (2005)
[18] Y. Benjamini, *et al., Biometrika.,* 93:491 (2006)
[19] D. Yekutieli & Y. Benjamini, *J Stat Plan Inf.,* 82:171 (1999)
[20] X. Qiu, *et al., BMC Bioinformatics*., 6:120 (2005) [PMID: 15904488]
[21] M. Van der Laan, *et al., Stat Appl Gen Mol Biol.,* 3:14 (2004) [PMID: 16646792]
[22] M. J. van der Laan, *et al., Stat Appl Gen Mol Biol.,* 3:15 (2004) [PMID: 16646793]
[23] S. Dudoit, *et al., Stat Appl Gen Mol Biol.,* 3:13 (2004) [PMID: 16646791]
[24] C. Cheng, *et al., Stat Appl Gen Mol Biol.,* 3:36 (2004) [PMID: 16646816]
[25] T. Schweder & E. Spjøtvoll, *Biometrika*., 69:493 (1982)
[26] M. Langaas & E. Ferkingstad, *J R Stat Soc – B.,* 67:555 (2005)
[27] D. B. Allison, *et al., Comput Stat Data Anal.,* 39:1 (2002)
[28] S. Pounds & S. Morris, *Bioinformatics*., 19:1236 (2003) [PMID: 12835267]
[29] S. Pounds & C. Cheng, *Bioinformatics*., 22:1979 (2006) [PMID: 16777905]
[30] S. Pounds & C. Cheng, *Bioinformatics*., 20:1737 (2004) [PMID: 14988112]
[31] S. Pounds & C. Cheng, *J Comput Biol.,* 12:482 (2005) [PMID: 15882143]
[32] P. J. Rousseeuw, *J Am Stat Assoc.,* 79:871 (1984)
[33] R. L. Iman & W. J. Conover, *Technometrics.,* 21:49 (1979)
[34] B. Efron, *J Am Stat Assoc.,* 99:96 (2004)
[35] S. Datta & S. Datta, *Bioinformatics,* 21:1987 (2005) [PMID: 15691856]
[36] F. Abramovich*et, et al., Tech Rep 2000-19* Dept. Statistics, Stanford University (2000)
[37] S. H. Jung, *Bioinformatics.,* 21:3097 (2005) [PMID: 15845654]
[38] W. Pan, *et al., Genome Biol.,* 3:5 (2004)
[39] S. Pounds & C. Cheng, *Bioinformatics*, 21:4263 (2005) [PMID: 16204346]
[40] L. Klebanov & A. Yakovlev, *Stat Appl Gen Mol Biol.,* 5:9 (2006) [PMID: 16646873]
[41] E. J. Yeoh, *et al., Cancer Cell*, 1:133 (2002) [PMID: 12086872]
[42] M. E. Ross, *et al., Blood*, 104:3679 (2004) [PMID: 15226186]
[43] A. B. Owen, *J Royal Stat Soc – B.,* 67:411 (2005)

Edited by Susmita Datta
Citation: Cheng & Pounds, Bioinformation 1(10): 436-446 (2007)