

CHOOSING THE LESSER EVIL: TRADE-OFF BETWEEN FALSE DISCOVERY RATE AND NON-DISCOVERY RATE

Radu V. Craiu and Lei Sun

University of Toronto

Abstract: The problem of multiple comparisons has become increasingly important in light of the significant surge in volume of data available to statisticians. The seminal work of Benjamini and Hochberg (1995) on the control of the false discovery rate (FDR) has brought forth an alternative way of measuring type I error rate that is often more relevant than the one based on the family-wise error rate. In this paper, we emphasize the importance of considering type II error rates in the context of multiple hypothesis testing. We propose a suitable quantity, the expected proportion of false negatives among the true alternative hypotheses, which we call non-discovery rate (NDR). We argue that NDR is a natural extension of the type II error rate of single hypothesis to multiple comparisons. The utility of NDR is emphasized through the trade-off between FDR and NDR, which is demonstrated using a few real and simulated examples. We also show analytically the equivalence between the FDR-adjusted p-value approach of Yekutieli and Benjamini (1999) and the q-value method of Storey (2002). This equivalence dissolves the dilemma encountered by many practitioners of choosing the “right” FDR controlling procedure.

Key words and phrases: False discovery rate, genome-scans, microarray data, multiple comparisons, multiple hypothesis testing, non-discovery rate, power, type I error, type II error.

1. Introduction

The advent of large dimensional data in scientific exploration underscores the need for more powerful methods to handle the multiplicity problem. In this context, once a large number m of hypothesis tests are performed, one needs to determine which, if any, of these tests have produced significant results. Traditionally, the decision is based on controlling the probability of making even one type I error, also known as the *Family-Wise Error Rate (FWER)*. However, controlling FWER for large values of m typically results in a diminished power to detect the true signal(s), although it should be noted that a clear definition of power in this context has yet to be specified.

The breakthrough paper of Benjamini and Hochberg (1995) (henceforth BH) offers a different approach in which one is interested in controlling the *False Discovery Rate (FDR)*, i.e., the fraction of erroneous rejections. If we test m hypotheses, we can summarize the findings as in Table 1, where m is assumed to be

Table 1. Summary of findings when testing m hypotheses.

	Declared non-significant	Declared significant	Total
True null hypothesis	\mathbf{U}	\mathbf{V}	m_0
Non-true null hypothesis	\mathbf{T}	\mathbf{S}	$m_1 = m - m_0$
	$m - \mathbf{R}$	\mathbf{R}	m

fixed and known, m_0 and m_1 are unknown parameters, \mathbf{R} is an observed random variable, and \mathbf{U} , \mathbf{V} , \mathbf{T} and \mathbf{S} are unobserved random variables. Given the above notation, $\text{FWER} = \Pr(\mathbf{V} \geq 1)$, and $\text{FDR} = E[\mathbf{V}/\mathbf{R}]$. To circumvent the situation in which $\mathbf{R} = 0$, FDR was alternatively defined as $E[\mathbf{V}/\mathbf{R}|\mathbf{R} > 0] \Pr(\mathbf{R} > 0)$ by BH, and as $\text{pFDR} = E[\mathbf{V}/\mathbf{R}|\mathbf{R} > 0]$ by Storey (2002). However, the distinction is not crucial in many applications because $\Pr(\mathbf{R} > 0) \approx 1$, as noted by Storey and Tibshirani (2003), among others. Here we work with the FDR alternatively defined by BH. Improvements and extensions of the BH method have been proposed by Benjamini and Hochberg (2000), Benjamini and Yekutieli (2001), Storey (2002, 2003), and Genovese and Wasserman (2001, 2002).

In the context of multiple hypothesis testing, the discussions so far have focused mostly on type I error rate, α , either in the form of FWER or FDR. However, in addition to α , of importance is also the type II error rate, β , or power, $1 - \beta$. Dudoit et al. (2003) briefly discussed three common definitions of power, namely $\Pr(\mathbf{S} \geq 1)$, $\Pr(\mathbf{S} = m_1)$ and $E[\mathbf{S}]/m_1$. The measure $E[\mathbf{S}]/m_1$ has been used to quantify power in a large number of studies (e.g., Storey, Taylor and Siegmund (2004) and Li et al. (2005)). Unfortunately, few studies offered in-depth investigation of β and power in the context of multiple hypothesis testing. In this paper, we intend to fill this gap by formally proposing a new quantity, the *Non-Discovery Rate* ($\text{NDR} = E[\mathbf{T}]/m_1$), the expected proportion of non-rejections among false null hypotheses, as one possible measure of type II error rate for multiple hypothesis testing, and investigating its properties and utilities. Of particular interest is the trade-off between NDR and FDR.

The number of false negatives \mathbf{T} was also considered by Genovese and Wasserman (2002) who defined the *False Non-discovery Rate*, $\text{FNR} = \mathbf{T}/(m - \mathbf{R})$. Although, mathematically, FNR is a suitable measure of β if FDR is used to quantify α , we note that FNR may be artificially decreased in the context of hypothesis generation as in Thomas et al. (1985), and the quantity itself is of little interest to practitioners (see Appendix for numerical examples). In addition, $1 - \text{FNR}$ is a function of true negatives but not true positives, an undesirable feature given the traditional use of $1 - \beta$ as power. (See Section 2.2 for detailed discussions.)

One of the referees has brought to our attention the *Fraction of Non-Selection* (FNS) proposed by Delongchamp et al. (2004). Our NDR measure indeed bears

considerable resemblance to FNS. However, we note that there are significant differences between the two quantities and methods. FNS was proposed specifically for the fixed rejection region approach, in the spirit of Storey (2002). That is, one rejects all tests with unadjusted p-values less than a pre-determined γ level, then estimates the corresponding FDR. (We use notation γ to distinguish it from α .) In that context, and assuming that the null p-values are Unif(0,1) distributed, they defined FNS as $\text{FNS}(\gamma) = (m_1 - (\mathbf{R} - m_0\gamma))/m_1$, where $\mathbf{R} = \#\{p_i \leq \gamma\}$. It should be noted that the threshold γ needed for FNS is not directly available if the traditional FDR control procedure is used to adjust for the multiplicity problem (i.e., controlling FDR at a pre-determined α level). In contrast, we define NDR as a measure of type II error rate in a general setting, independent of the specific multiple comparison procedure used. The chosen procedure only affects the estimation of NDR. (See Section 6 for more discussion.)

In the next section we formally define NDR and provide justification for considering this quantity as the type II error rate for multiple hypothesis testing. We emphasize that much can be gained from a clear understanding and representation of the dependence between FDR and NDR. The other main contribution of the paper is in Section 3 where we show that the FDR-adjusted p-value approach of Yekutieli and Benjamini (1999) is equivalent to Storey's q-value method (2002). This equivalence dissolves the dilemma encountered by many practitioners of choosing the "right" FDR procedure, and gives us the freedom to work with either method in estimating NDR in Section 4. We illustrate our results with a series of real and simulated data sets in Section 5. We present conclusions and discussion of further work in Section 6.

2. Joint Analysis of FDR and NDR

2.1. Definition and motivation

Definition 1. Given $m_1 > 0$, the non-discovery rate (NDR) is

$$\text{NDR} = \frac{\mathbf{E}[\mathbf{T}]}{m_1}.$$

Note that the quantity is defined given $m_1 > 0$. Similarly to FDR for which $\mathbf{R} = 0$ is a concern, one may define $\text{NDR} = \mathbf{E}[\mathbf{T}]/m_1 I(m_1 > 0)$. However, this requires a Bayesian approach rather than treating m_1 as an unknown but fixed parameter. In addition, in situations where none of the null hypotheses are likely to be false, one would probably not conduct the analysis in the first place. Thus, we limit our attention to the case where $m_1 > 0$.

The motivation of our work can be well demonstrated by the following toy example. Consider a situation in which $m = 1,000$ and $m_1 = 100$, and assume the following two strategies. Under Strategy 1, we choose $\text{FDR} = 0.05$. Suppose

that the number of rejections is $\mathbf{R} = 20$, among which one is expected to be incorrect, i.e., $\mathbf{V} = 1$. Thus, the number of false negatives \mathbf{T} is likely to be $m_1 - (\mathbf{R} - \mathbf{V}) = 81$ and $\text{NDR} = 0.81$. Under Strategy 2, we decide to increase FDR to 0.1. Suppose that $\mathbf{R} = 80$, then $E[\mathbf{V}] = 8$ and $\text{NDR} = 0.28$. A natural question is whether one should choose Strategy 1 or 2 to perform the analysis. While the answer depends on the specific objective of each study, we believe that each choice should be made while fully aware of the fact that the proportion of missed signals as measured by NDR could be unsatisfactorily large for a given FDR level, and a small increase in FDR may result in a considerable amount of decrease in NDR.

The dependence between α and β for single hypothesis testing is well documented in the literature. For example, assume that $\sqrt{n} \bar{X}/\sigma$ is used to test the mean of a normal population ($H_0 : \theta = 0$ vs. $H_1 : \theta > 0$) with known variance σ^2 , based on n i.i.d. samples. The probability of a type II error is then $\beta(\theta; \alpha) = 1 - \Phi(\Phi^{-1}(\alpha) + \sqrt{n} \theta/\sigma)$ where Φ is the cdf of $N(0, 1)$. Figure 1 illustrates the trade-off between α and β for $\theta = 1$ and 2, assuming $n = 100$ and $\sigma = 5$. The dashed lines connect α values of 0.01 and 0.05 with their corresponding β values. It can be seen that the gain in power obtained when α is relaxed from 0.01 to 0.05 is very different in the two situations. Although an increase in α comes with an increased power in the context of simple hypothesis testing, α is typically pre-specified at a small value (e.g., $\alpha = 0.05$ for social sciences and $\alpha = 0.01$ or 0.001 for natural sciences), and β is mainly discussed when design and sample size are of concern. However, such standard statistical practice is not yet available for multiple hypothesis testing utilizing FDR. First, the definition of type II error rate and power in this context is often unclear. Second, the

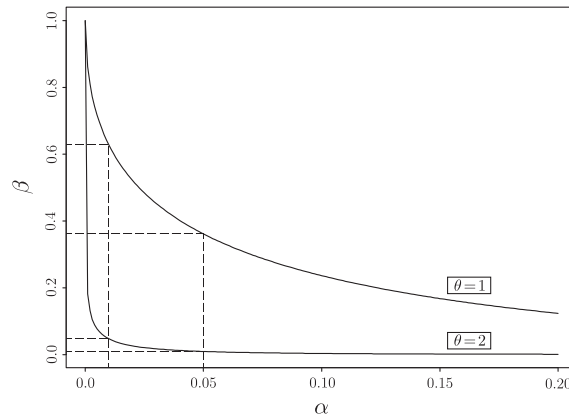


Figure 1. Illustration of the trade-off between type I and type II error rates for single hypothesis testing.

trade-off between FDR and type II error rate is not well studied. Finally, the choice of FDR level seems to be somewhat arbitrary from study to study.

2.2. NDR as type II error rate for multiple comparisons

When a large number of hypotheses are tested simultaneously, the choice of type II error rate, β , and power, $1 - \beta$, is not unique. For example, $\Pr(\mathbf{T} \leq k)$ and $\Pr(\mathbf{S} \geq m_1 - k)$, $0 \leq k \leq m_1$, could be defined as type II error rate and power, respectively, where the choice of k reflects the stringency of the criterion. However, such definitions are probably more suitable as counterparts for FWER.

To see why NDR is a good candidate as type II error rate for multiple hypothesis testing, consider m independent and identical hypotheses H_1, \dots, H_m with test statistics t_1, \dots, t_m . Let $H_i = 0$ denotes a true null hypothesis, $H_i = 1$ otherwise, and \mathcal{R} be the rejection region. It is not difficult to see that $\text{NDR} = \sum_i \Pr(t_i \notin \mathcal{R} | H_i = 1) / m_1 = \bar{\beta}$. Thus NDR is essentially the average of type II error rate of single hypothesis testing. If we further assume $\Pr(H_i = 0) \equiv \pi_0$, from the Bayesian point of view $\text{FDR} = \Pr(H_i = 0 | t_i \in \mathcal{R})$. A direct extension of this definition of α leads to the FNR of Genovese and Wasserman (2002), and $\text{E[FNR]} = \Pr(H_i = 1 | t_i \notin \mathcal{R})$. However, it is difficult to interpret $1 - \text{FNR}$ as power, because $1 - \text{E[FNR]} = \Pr(H_i = 0 | t_i \notin \mathcal{R}) = \text{E}[\mathbf{U} / (m - \mathbf{R})]$, a quantity that depends on the true negatives but not the true positives. In contrast, NDR has the traditional frequentist interpretation of β , and $1 - \text{NDR} = \sum_i \Pr(t_i \in \mathcal{R} | H_i = 1) / m_1 = \overline{1 - \beta}$, the average power. Note that $1 - \text{NDR} = \text{E}[\mathbf{S}] / m_1$ is precisely the power defined in Dudoit et al. (2003) and used in many applications. NDR can be considered a direct extension of the Per-Comparison Error Rate (PCER), where $\text{PCER} = \text{E}[\mathbf{V}] / m \leq \text{E}[\mathbf{V}] / m_0 = \sum_i \Pr(t_i \in \mathcal{R} | H_i = 0) / m_0$. Although it is mathematically more straightforward to pair FDR with FNR, and PCER with NDR, the values of FNR and PCER are of little interest to practitioners. Therefore, we choose to consider the trade-off between FDR and NDR.

2.3. Trade-off between FDR and NDR

An astute reader will not be surprised that, similar to the trade-off between α and β in the context of single hypothesis, there is one between FDR and NDR for multiple comparisons. To illustrate this, assume that the p-values are from $(\text{Unif}[0, 1])^\theta$, where $\theta = 1$ corresponds to the true null hypotheses and $\theta > 1$ corresponds to the false null hypotheses. Figure 2 shows three different types of dependencies between FDR and NDR corresponding to $(\theta = 2, \pi_0 = 0.9)$, $(\theta = 3, \pi_0 = 0.7)$ and $(\theta = 4, \pi_0 = 0.5)$, where π_0 is the proportion of true

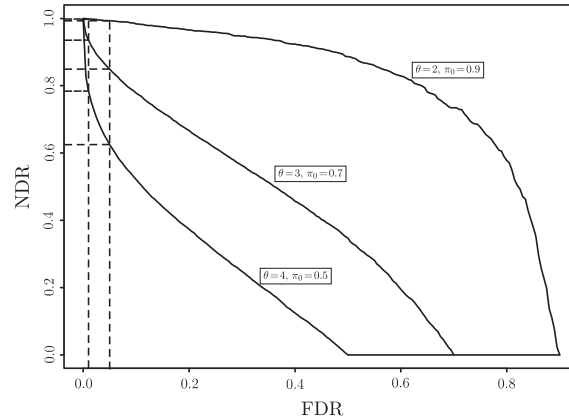


Figure 2. Illustration of dependence between FDR and NDR for multiple hypothesis testing.

null hypotheses. Dashed lines connect FDR values of 0.01 and 0.05 with their corresponding NDR values. As shown clearly by the graph, a) for some cases, FDR at 0.01 or 0.05 level may not be suitable if the objective is to screen for as many true signals as possible for follow up studies, b) a slight increase of FDR may result in various amounts of decrease in NDR for different alternatives. Evidently, a better understanding of the relationship between NDR and FDR can assist with the choice of FDR as well as sample size calculation.

Unlike the α level for a single test, there is no “golden standard” (e.g., 0.05) for FDR and it could be argued that FDR is chosen based on the specific objective of each study. For example, in the context of exploratory microarray analyses, a larger value of FDR might be preferred in order to minimize NDR and identify as many interesting/promising genes as possible. If the goal is to perform (often costly) functional studies on significant genes, a more stringent FDR level might be required to minimize the number of false positives and reduce the unnecessary spending.

In the context of FNS control, Delongchamp et al. (2004) considered the trade-off between power ($=1-\text{FNS}$) and FDR which is also of interest. In fact, the curve of $1-\text{NDR}$ vs. FDR can be interpreted as a Receiver Operating Characteristics (ROC) curve, where $1-\text{NDR}$ is the sensitivity and FDR is $1-\text{specificity}$. The methodology developed for ROC curves could be used to evaluate different designs and analytic methods in the context of false discovery control. However, in order to emphasize the main message of the paper which is the trade-off between the two types of errors, we continue the discussion around the dependency between NDR and FDR.

2.4. Cost-effectiveness measure based on FDR and NDR

The decision to use a particular level of FDR depends on the cost of introducing false positives and the gain of discovering true positives. In this respect, some measures of cost-effectiveness or efficiency might be proposed. For example,

$$e_{\text{ratio}} = w_{\text{ratio}} \frac{\frac{\text{NDR}_2}{\text{FDR}_1}}{\frac{\text{NDR}_1}{\text{FDR}_2}},$$

$$e_{\text{slope}} = w_{\text{slope}} \frac{\text{NDR}_2 - \text{NDR}_1}{\text{FDR}_2 - \text{FDR}_1},$$

where w_{ratio} and w_{slope} are weighting factors, and $e_{\text{ratio}} > 1$ or $e_{\text{slope}} > -1$ indicates that FDR_1 is more cost-effective than FDR_2 . If e_{ratio} is used with $w_{\text{ratio}} = 1$, increasing FDR from 0.01 to 0.1 requires a 10-fold decrease in NDR to make the two choices of FDR equally efficient, which may not be reasonable. Alternatively, $w_{\text{ratio}} < 1$ could be used to give more weight to the effectiveness of a decrease in NDR. Compared to e_{ratio} , e_{slope} has a natural interpretation when $w_{\text{slope}} = 1$: it can be viewed as the slope of the NDR versus FDR curve. In that case, if the slope of the curve at a chosen FDR value is less than -1, using a less stringent FDR level could be beneficial. The final choice of e_{ratio} or e_{slope} , or some other measure of efficiency, is likely to be study specific and of interest for future research. In Section 5 we demonstrate patterns observed in three datasets collected for distinctive genetic and genomic studies.

3. Equivalence between FDR-adjusted p-values and q-values

The estimation of NDR depends on the specific procedure for FDR control. Currently, there are two main approaches. The first, due to Yekutieli and Benjamini (1999), is based on the FDR-adjusted p-value. The second is Storey's (2002) q-value method. We show in the following section that the two approaches are in fact equivalent, a result that was also shown numerically by Black (2004) through simulation studies.

In the original BH procedure, if $p_{(1)} \leq \dots \leq p_{(m)}$ is the ordered sequence of m available p-values, we search for the largest k such that $p_{(k)} \leq (k/m) \alpha$ and reject all $H_{(j)}$, $j \leq k$. Benjamini and Hochberg (2000), Benjamini and Yekutieli (2001), Genovese and Wasserman (2002) and Finner and Roters (2002) have all shown that this procedure is conservative in that $\text{FDR} = \alpha \pi_0$, where $\pi_0 = m_0/m$. (Evidently, $\text{FDR} \leq \pi_0$ because $\alpha \leq 1$.) A natural modification is given by the adaptive BH procedure (Benjamini and Hochberg (2000)) in which we look for

the largest k such that $p_{(k)} \leq (k/m)(\alpha/\pi_0)$ and reject all $H_{(j)}$, $j \leq k$. This leads to $\text{FDR} = (\alpha/\pi_0) \pi_0 = \alpha$.

Equivalently, the BH procedure can be performed by means of the FDR adjusted p-value (Yekutieli and Benjamini (1999)). The FDR adjusted p-value corresponding to $p_{(i)}$ is

$$p_{(i)}^{\text{FDR}} = \min \left\{ \frac{mp_{(j)}}{j} : j \geq i \right\} = \min \left\{ \frac{mp_{(i)}}{i}, p_{(i+1)}^{\text{FDR}} \right\}, \quad (3.1)$$

with $p_{(m)}^{\text{FDR}} = p_{(m)}$. In order to maintain $\text{FDR} \leq \alpha$ one rejects all hypotheses with $p_{(i)}^{\text{FDR}} \leq \alpha$. Obviously, an adaptive BH method can be implemented by rejecting all hypotheses with $p_{(i)}^{\text{FDR}} \leq \alpha/\hat{\pi}_0$, resulting in $\mathbf{R} = \#\{p_{(i)}^{\text{FDR}} \leq \alpha/\hat{\pi}_0\}$. The superiority of the adaptive BH procedure is apparent in that the number of rejections \mathbf{R} is at least as large as in the original BH procedure while still controlling FDR at level α . However, we emphasize that both the adaptive BH procedure and Storey's approach below require a good approximation of π_0 .

Recently Storey (2002, 2003) has proposed the notion of q -value as the FDR counterpart of the p-value. Roughly speaking, the q -value of an observed test statistic associated with hypothesis H_i is the minimum possible FDR for calling H_i significant. If we declare all hypotheses with q -values less or equal to α significant, then $\text{FDR} \leq \alpha$ for large m . Using the same notation as in Storey and Tibshirani (2003), the q -value can be estimated using

$$\hat{q}_{(i)} = \min \left\{ \frac{\hat{\pi}_0 mp_{(i)}}{i}, \hat{q}_{(i+1)} \right\},$$

where $\hat{q}_{(m)} = \hat{\pi}_0 p_{(m)}$. It is not hard to see that $\hat{q}_{(i)} = \hat{\pi}_0 p_{(i)}^{\text{FDR}}$, therefore the number of rejections based on q -values, $\mathbf{R}_q = \#\{\hat{q}_{(i)} \leq \alpha\} = \#\{p_{(i)}^{\text{FDR}} \leq \alpha/\hat{\pi}_0\} = \mathbf{R}$.

4. Estimation of NDR

To estimate NDR, we choose to use the adaptive BH procedure based on the FDR adjusted p-value. Given the results in Section 3, the estimator works equally well for the q -value approach.

Storey (2003) has shown that, under certain assumptions including independence between tests,

$$\mathbb{E} \left[\frac{\mathbf{V}}{\mathbf{R}} \mid \mathbf{R} > 0 \right] = \frac{\mathbb{E}[\mathbf{V}]}{\mathbb{E}[\mathbf{R}]}. \quad (4.1)$$

For general cases, Storey and Tibshirani (2003) argued that (4.1) holds approximately for large m . In the following we assume that (4.1) holds. For a chosen

FDR level α such that $0 \leq \alpha \leq \pi_0$ and $E[(\mathbf{V}/\mathbf{R})|\mathbf{R} > \mathbf{0}] \Pr(\mathbf{R} > \mathbf{0}) = \alpha$,

$$\text{NDR} = \frac{E[\mathbf{T}]}{m_1} = 1 - \frac{E[\mathbf{R} - \mathbf{V}]}{m_1} = 1 - \frac{\left(1 - \frac{\alpha}{\Pr(\mathbf{R} > \mathbf{0})}\right)E[\mathbf{R}]}{(1 - \pi_0)m}. \tag{4.2}$$

A simple estimate of NDR may be obtained by replacing $E[\mathbf{R}]$ with its observed value \mathbf{R} , where $\mathbf{R} = \#\{p_{(i)}^{\text{FDR}} \leq \alpha/\hat{\pi}_0\}$, and π_0 with $\hat{\pi}_0$. We obtain

$$\widehat{\text{NDR}} = \left\{ 1 - \frac{(1 - \alpha)\mathbf{R}}{(1 - \hat{\pi}_0)m} \right\} I(\hat{\pi}_0 < 1). \tag{4.3}$$

Note that the event $\{\mathbf{R} = 0\}$ is of little concern here because a) $\Pr(\mathbf{R} > \mathbf{0}) \approx 1$ in practice, as argued by Storey and Tibshirani (2003) among others, and b) when $\mathbf{R} = 0$, NDR obviously should be 1 which is the case based on $\widehat{\text{NDR}}$ above. Because of the small variance of $\hat{\pi}_0$ (on the order of $o(1/m)$ using the $\hat{\pi}_0(\lambda)$ estimator below and assuming tests are independent), $\text{cov}(\mathbf{R}, \hat{\pi}_0)$ is negligible. Therefore (for simplicity we omit the indicator $I(\hat{\pi}_0 < 1)$ in the following expression),

$$\begin{aligned} E[\widehat{\text{NDR}}] &\approx 1 - \frac{(1 - \alpha)E[\mathbf{R}]}{(1 - E[\hat{\pi}_0])m} = 1 - \frac{(1 - \alpha)E[\mathbf{R}]}{(1 - \pi_0 - c)m} \\ &= 1 - \frac{(1 - \alpha)E[\mathbf{R}]}{m} \left\{ \frac{1}{1 - \pi_0 - c} - \frac{1}{1 - \pi_0} + \frac{1}{1 - \pi_0} \right\} \\ &= 1 - \frac{(1 - \alpha)E[\mathbf{R}]}{(1 - \pi_0)m} - \frac{(1 - \alpha)E[\mathbf{R}]}{(1 - \pi_0)m} \frac{c}{1 - \pi_0 - c}, \end{aligned} \tag{4.4}$$

where c is the bias of $\hat{\pi}_0$. Equation (4.4) suggests that the bias of $\widehat{\text{NDR}}$ mainly depends on the bias of $\hat{\pi}_0$. Consider the following commonly used π_0 estimator

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m(1 - \lambda)}. \tag{4.5}$$

With $\lambda = 0.5$, it is easy to show that the (upward) bias of the estimator is $2(1 - \pi_0)\epsilon$, where ϵ is the probability an alternative p-value is greater than 0.5. Let F_0 and F_1 be the cdf of the test statistic under the null and alternative hypotheses, $\epsilon = F_1[F_0^{-1}(0.5)]$, a small value as long as F_1 and F_0 are not close. For example, assume $F_0(t) = \Phi_{0,1}(t)$ and $F_1(t) = \Phi_{\mu_1,1}(t)$, $\mu_1 > 0$, where $\Phi_{\mu,\sigma^2}(t)$ is the cdf of $N(\mu, \sigma^2)$, as in Cox and Wong (2004). Then $c = 2(1 - \pi_0)\Phi_{0,1}(-\mu_1)$, a value that decreases fast as μ_1 increases. However, in situations where π_0 is small and the “distance” between the null and alternative populations is also small, the bias of $\hat{\pi}_0$ could be considerable (Black (2004)) which may lead to a non-negligible downward bias in $\widehat{\text{NDR}}$, particularly when the chosen FDR level is

high. We thus must acknowledge the central role of accurate estimation of π_0 in the performance of the method proposed here. Specifying good estimators for π_0 goes beyond the scope of this paper, and we refer readers to the work of Storey (2002), Storey and Tibshirani (2003), Storey, Taylor and Siegmund (2004), and Langaas, Lindqvist and Ferkingstad (2005).

5. Examples and Simulation Study

The following examples demonstrate the various relationships that may exist between FDR and NDR. For practical reasons, one might be also interested in knowing $E[\mathbf{R}]/m$, the proportion of rejections among all the m tests at the given FDR level. We call this quantity *Proportion Of Rejection* (POR). A natural unbiased estimator is $\widehat{\text{POR}} = \mathbf{R}/m$. Some simple algebra can also show that $\text{POR} \approx (1 - \pi_0)(1 - \text{NDR})/(1 - \text{FDR})$. For clarification, we summarize here the procedure used to produce the plots and tables in this section.

Step 1: Estimate π_0 , e.g., using (4.5) with $\lambda = 0.5$.

Step 2: Choose $FDR = \alpha$, e.g., $\alpha \in (0, \hat{\pi}_0)$, on a grid of 0.01.

Step 3: Derive \mathbf{R} , e.g., using the above adaptive BH procedure.

Step 4: Estimate NDR using (4.3), and POR as above.

5.1. Microarray data

Our first illustration uses the data from Example 1 of Storey and Tibshirani (2003) which contains $m = 3,170$ p-values calculated from a study of microarray gene expression data (p-values were obtained from Dr. Storey's website at <http://faculty.washington.edu/~jstorey/>). In this case we have also used Storey and Tibshirani's estimate of $\hat{\pi}_0 = 0.67$.

Figure 3 presents the relationships between NDR and FDR (left panel), and between POR and FDR (right panel). The dashed lines connect $FDR = 0.01, 0.1$ and 0.2 with their corresponding NDR and POR estimates. The results indicate that increasing FDR from 0.1 to 0.2 can work well for this dataset since it reduces NDR from 0.73 to 0.45. If one uses $w_{\text{slope}} = 1$, then $e_{\text{slope}} = -2.8$ which favors $FDR = 0.2$. The slope of the curve at $FDR = 0.1$ is -2.44 , which is obtained by fitting a linear regression model in a local interval of $FDR, (0.1 - 0.02, 0.1 + 0.02)$. For extreme situations where the pattern is non-linear, more sophisticated models such as cubic splines may be required to approximate the slope at a given point. In the right hand panel, it can be seen that roughly 10% and 23% of the 3,170 tests would be rejected (317 and 729 out of 3,170), respectively, for an FDR of 0.1 and 0.2.

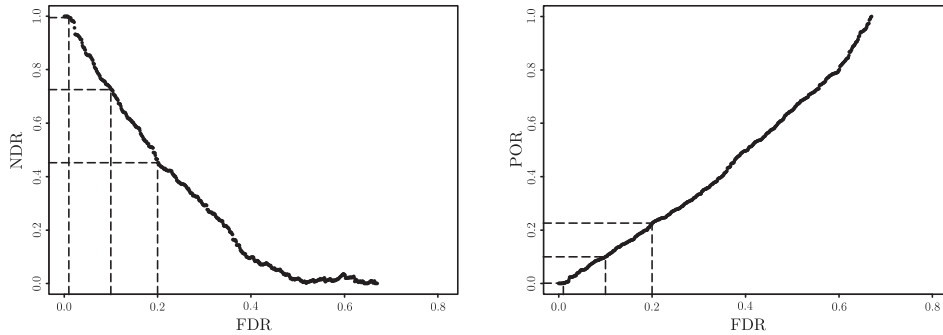


Figure 3. NDR vs. FDR (left panel), and POR vs. FDR (right panel) for the microarray data example.

5.2. Pedigree error detection using genomewide genetic marker data

The following two examples concern pedigree error detection in the context of genome-scans for localizing disease susceptibility genes. Both datasets were distributed as part of the biennial Genetic Analysis Workshops (GAW). The COGA dataset was collected for the study of genetics of alcoholism (GAW11), and its potentially misspecified relationships among relative pairs were analyzed by McPeck and Sun (2000). The CSGA dataset was used for an asthma study (GAW12), and its pedigree errors were analyzed by Sun, Abney and McPeck (2001). Given a set of collected families, the null hypothesis for a particular relative pair is the relationship type indicated by the given pedigree, e.g., a sib pair. Genome-wide genetic marker data are used to perform the corresponding hypothesis test. Figure 4 shows the histogram of the 5381 p-values from the COGA dataset and the 3276 p-values from the CSGA dataset. The estimates of π_0 are 0.68 and 0.81 for the COGA and CSGA datasets, respectively.

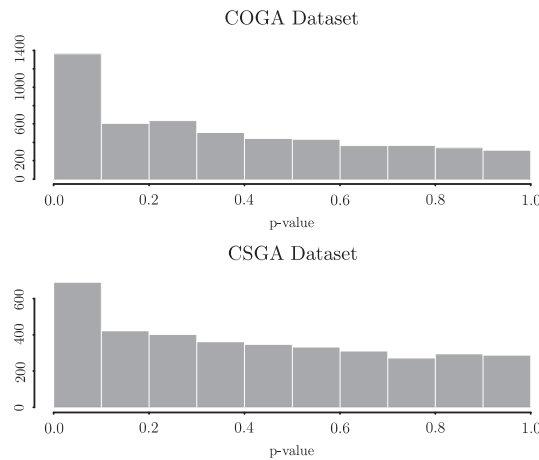


Figure 4. Histogram of the p-values for the two pedigree error examples.

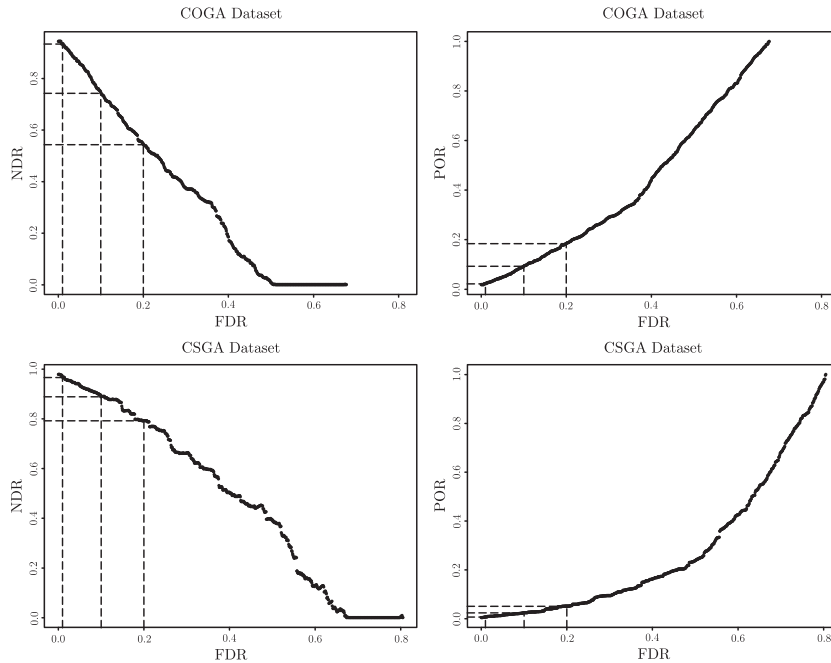


Figure 5. NDR vs. FDR (left panel) and POR vs. FDR (right panel) for the two pedigree error examples.

The top panel of Figure 5 (COGA dataset) shows trends similar to those from the previous microarray example. Increasing FDR from 0.1 to 0.2 seems to be beneficial: NDR decreases from 74% to 54% with $e_{\text{slope}} = -2$. The bottom panel in Figure 5 (CSGA dataset) presents a different image. Indeed, even with $\text{FDR} = 0.2$, NDR remains high (about 0.8) and only 20% of the truly misspecified relative pairs could be detected. Increasing FDR from 0.1 to 0.2 leads to $e_{\text{slope}} = -0.9$. In this case, $\hat{\pi}_0 = 0.81$, so one might think that with a smaller $\pi_1 = m_1/m$ it would be easier to identify all the m_1 true alternatives. However, since the noise in the data is larger, it is in fact more difficult to detect the true signals. The following simulation study further shows that, given the same level of FDR, NDR increases as π_0 increases.

5.3. A simulation study

We performed a simulation study to investigate the bias and variance of the NDR and POR estimates. The simulation model considered is similar to that of Storey (2002) and Black (2004). We generated $m = 5,000$ independent data points from a normal distribution with mean μ and known variance $\sigma^2 = 1$. For each set of data, $m\pi_0$ observations were simulated from $\mu_0 = 0$ and the remaining ones from $\mu_1 = 1, 1.5, 2, 2.5$ and 3, with π_0 ranging from 0.6 to 0.9 on a grid

of 0.1. We considered FDR at level 0.01, 0.05, and from 0.1 to 0.5 on a grid of 0.1. For each replication, to estimate π_0 , NDR and POR, we used the procedure described at the beginning of this section. We repeated the above 1,000 times. The true NDR was estimated through another set of 1,000 simulations in which \mathbf{T} was tracked.

Figure 6 demonstrates the relationships among FDR, NDR and POR for the models considered, for a particular simulation realization. The graph clearly shows that NDR increases as π_0 increases. Not surprisingly, for a given π_0 , the “distance” between the null and the alternative hypotheses plays a significant role in determining NDR. For example, when $\pi_0 = 0.7$ and $\mu_1 = 3$, by allowing FDR to be 0.2 we could identify most of the true signals. In contrast, if $\mu_1 = 1$, NDR does not change almost at all even if one increases FDR from, say, 0.01 to 0.2. The above remarks are also clearly reflected by the differences in slopes of the corresponding NDR versus FDR curves.

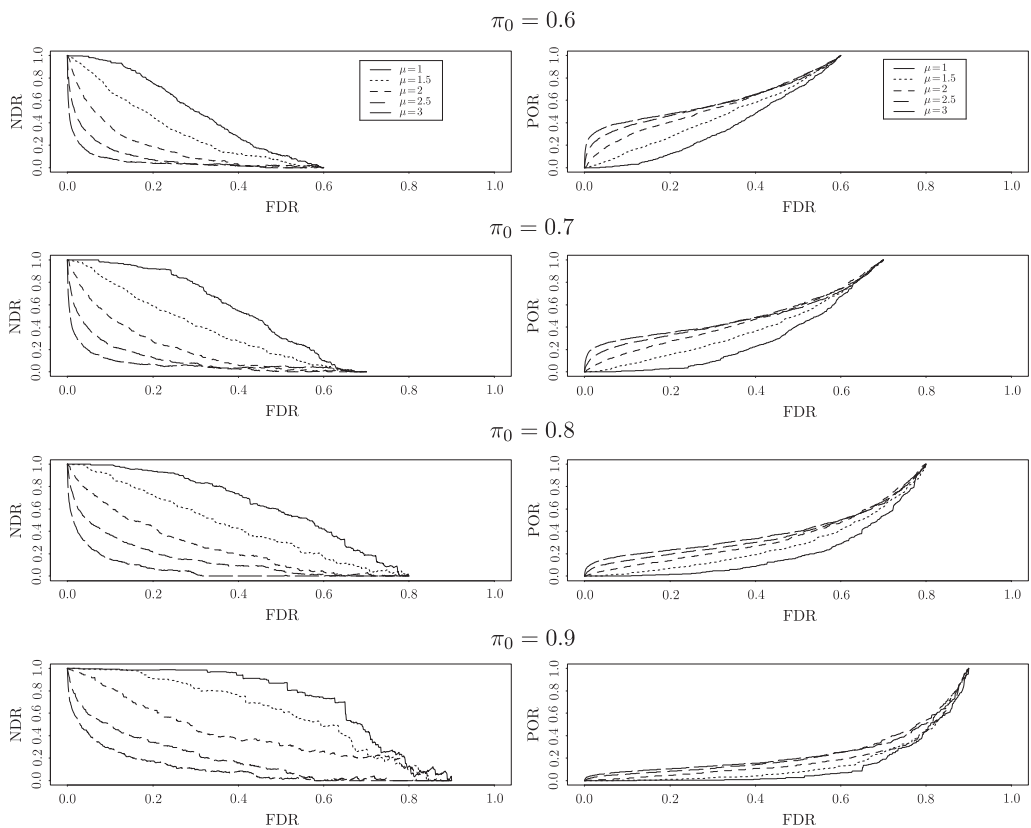


Figure 6. NDR vs. FDR (left panel) and POR vs. FDR (right panel) for the simulation model.

Table 2 summarizes the results over the 1,000 replicates and gives the sample average of the NDR estimates and their standard error (SE). (Results for $\mu_1 = 1.5$ and 2.5 are not shown because of their similarities to the others.) The estimate of NDR tends to be downward biased because of the upward bias of $\hat{\pi}_0$. In most cases, the biases are very small. The worse case scenario is the situation when π_0 and μ_1 are small and FDR is large (the largest bias is -0.18 when $\pi_0 = 0.6$, $\mu_1 = 1$, and $FDR = 0.5$). When π_0 and μ_1 are small, the estimate of π_0 tends to be less accurate, resulting in a larger bias c . In that case, if the chosen FDR level is very high, the proportion of rejections increases considerably and the bias of NDR becomes non-negligible because it is proportional to $E[\mathbf{R}]/m$. For example, when $\pi_0 = 0.6$ and $\mu_1 = 1$, $c \approx 0.1$. If $FDR = 0.5$, roughly half of the tests are rejected, and the downward bias of $\widehat{\text{NDR}}$ is about 0.2, while the true NDR is about 0.3. This might be an extreme case in practice, since FDR is unlikely to be chosen at the 0.5 level.

Table 2. Simulation results for the estimate of NDR

Parameters μ_1, FDR	$\widehat{\text{NDR}}(\text{Bias, SE})$			
	$\pi_0 = 0.9$	$\pi_0 = 0.8$	$\pi_0 = 0.7$	$\pi_0 = 0.6$
$\mu_1 = 1.0$				
$FDR = 0.01$	1.00 (0.00, 0.001)	1.00 (0.00, 0.001)	1.00 (0.00, 0.001)	1.00 (0.00, 0.001)
$FDR = 0.05$	1.00 (0.00, 0.003)	1.00 (0.00, 0.003)	1.00 (0.00, 0.004)	0.99 (0.00, 0.006)
$FDR = 0.10$	1.00 (0.00, 0.007)	0.99 (0.00, 0.008)	0.98 (-0.01, 0.011)	0.96 (-0.01, 0.017)
$FDR = 0.20$	0.99 (-0.01, 0.014)	0.96 (-0.01, 0.022)	0.90 (-0.03, 0.031)	0.81 (-0.06, 0.034)
$FDR = 0.30$	0.97 (-0.01, 0.027)	0.89 (-0.03, 0.040)	0.75 (-0.07, 0.045)	0.57 (-0.11, 0.042)
$FDR = 0.40$	0.93 (-0.03, 0.046)	0.76 (-0.07, 0.059)	0.53 (-0.12, 0.052)	0.32 (-0.16, 0.037)
$FDR = 0.50$	0.86 (-0.05, 0.074)	0.57 (-0.12, 0.069)	0.30 (-0.16, 0.050)	0.10 (-0.18, 0.031)
$\mu_1 = 2.0$				
$FDR = 0.01$	0.98 (0.00, 0.014)	0.95 (0.00, 0.015)	0.91 (0.00, 0.014)	0.87 (-0.01, 0.015)
$FDR = 0.05$	0.88 (-0.01, 0.036)	0.76 (-0.01, 0.029)	0.65 (-0.02, 0.023)	0.55 (-0.02, 0.019)
$FDR = 0.10$	0.77 (-0.02, 0.049)	0.60 (-0.02, 0.035)	0.47 (-0.02, 0.027)	0.36 (-0.03, 0.020)
$FDR = 0.20$	0.59 (-0.03, 0.069)	0.40 (-0.03, 0.042)	0.26 (-0.03, 0.029)	0.16 (-0.03, 0.019)
$FDR = 0.30$	0.45 (-0.04, 0.083)	0.25 (-0.03, 0.045)	0.13 (-0.03, 0.027)	0.06 (-0.03, 0.015)
$FDR = 0.40$	0.34 (-0.04, 0.094)	0.15 (-0.04, 0.044)	0.06 (-0.03, 0.022)	0.01 (-0.03, 0.008)
$FDR = 0.50$	0.23 (-0.05, 0.094)	0.08 (-0.03, 0.037)	0.01 (-0.03, 0.012)	0.00 (-0.01, 0.003)
$\mu_1 = 3.0$				
$FDR = 0.01$	0.63 (0.00, 0.053)	0.51 (0.00, 0.036)	0.42 (0.00, 0.027)	0.35 (0.00, 0.020)
$FDR = 0.05$	0.37 (-0.01, 0.084)	0.25 (0.00, 0.048)	0.18 (0.00, 0.032)	0.13 (0.00, 0.022)
$FDR = 0.10$	0.25 (-0.01, 0.094)	0.15 (0.00, 0.051)	0.10 (0.00, 0.033)	0.06 (0.00, 0.021)
$FDR = 0.20$	0.15 (0.00, 0.093)	0.07 (0.00, 0.047)	0.04 (0.00, 0.028)	0.02 (0.00, 0.016)
$FDR = 0.30$	0.11 (0.00, 0.085)	0.04 (0.00, 0.039)	0.02 (0.00, 0.020)	0.01 (0.00, 0.010)
$FDR = 0.40$	0.08 (0.00, 0.076)	0.03 (0.00, 0.030)	0.01 (0.00, 0.013)	0.00 (0.00, 0.004)
$FDR = 0.50$	0.06 (0.00, 0.066)	0.02 (0.00, 0.023)	0.00 (0.00, 0.006)	0.00 (0.00, 0.005)

6. Conclusions and Future Work

In this paper, we proposed the use of the quantity NDR, the expected proportion of non-rejections among the false null hypotheses, which can be viewed as a natural extension of the type II error rate for multiple hypothesis testing. (The concept of NDR certainly relies on the assumption that there are true alternatives, i.e., $m_1 > 0$.) We also proposed a simple estimator for NDR and investigated its accuracy through simulation studies. Although the observed bias of our estimator was small, we note that its performance depends highly on the accuracy of the π_0 estimation, particularly when $\pi_0 \approx 1$ or when tests are not independent of each other. Alternatively, the bias of $\hat{\pi}_0$ could be potentially incorporated directly in the estimation of NDR. This is of future research interest.

NDR and its trade-off relationship to FDR can be utilized in many exploratory studies in which the problem of multiple comparisons is of concern, yet an “optimal” level of FDR to be controlled is unknown. The NDR measure is also useful at the stage of study design, in particular, for the determination of adequate sample size for a required level of accuracy. In this context, FDR is the type I error rate to be controlled, and NDR is the type II error rate to be minimized, so power is considered to be $1 - \text{NDR}$. For example, in the simulation study when $\mu_1 = 1$ and $\pi_0 = 0.6$, there is almost no power ($1 - \text{NDR} = 1 - 0.96 = 0.04$) for FDR at the 0.1 level. However, if a power of 80% is desired while FDR needs to be maintained at the 0.1 level, then a simple simulation study shows that a sample 4 times that of the original one is required for each test. Müller et al. (2004), and Tsai et al. (2004) also considered the sample size calculation for multiple hypothesis testing, but in the context of FNR and/or the number of false negatives.

It has been shown that the current FDR controlling procedures work well for independent tests and tests with Positive Regression Dependency (PRD). However, the effect of general dependence has not been well studied. In their recent simulation studies of microarray data, Li et al. (2005) have demonstrated that the actual FDR could be twice the nominal level when the dependent structure among tests was mimicked under realistic assumptions, and if the proportion of null genes is greater than 90%. Unfortunately, this is likely to be the case for most microarray analyses and genome-wide genetic studies. In addition, an estimator of π_0 such as the one given in (4.5) is sensitive to the assumption of independence. Accurate π_0 estimation and FDR control under general conditions is still an open question. The recent work of Efron (2005) suggests that empirical null distributions (Efron (2004)) could be used as a more robust technique to control FDR in the presence of correlation. Based on the current estimator of NDR using (4.3), the downward bias of FDR would lead to a downward bias of NDR. Li et al. (2005) have recommended adjusting the nominal FDR level by

half when $\pi_0 > 0.9$. In that case, one crude adjustment for the NDR estimate is to replace the nominal FDR level α by 2α in (4.3).

In contrast to the FNS of Delongchamp et al. (2004) that was defined exclusively for the fixed rejection region procedure, our NDR is defined as a general measure of type II error rate regardless of the specific FDR procedure. The chosen method however does affect the estimation of NDR. The estimator considered in (4.3) was developed for the fixed FDR procedure, but can be easily modified to obtain an NDR estimate under the fixed rejection framework by replacing the nominal FDR value α with the estimated FDR level. Discussions on the connection between the fixed rejection and fixed FDR methods can be found in Storey, Taylor and Siegmund (2004) and Sun et al. (2006).

Our study of NDR, as well as most work on FDR, has focused on the mean of the estimators. The variance is another quantity of interest, especially in the context of constructing confidence intervals. The variances of $\widehat{\text{NDR}}$ and $\widehat{\text{POR}}$ are both proportional to the variance of \mathbf{R} , which depends on the structure of both null and alternative models. The recent paper of Owen (2005) shows possible pathways of exploring the variance of \mathbf{R} and is of particular importance to further development of the work presented here.

Appendix

In this Appendix, we provide numerical examples that demonstrate the differences between NDR and FNR. FNR was proposed to compare the performance of different FDR controlling procedures (Genovese and Wasserman (2002)), and has a clear connection with FDR as discussed above. For a given dataset, there is also a trade-off between FDR and FNR, as shown in Table A and Table B. However, because the value of FNR depends on the number of null hypotheses, the value of FNR could be artificially decreased as demonstrated by the comparison between the two tables. Therefore, it is difficult to use FNR as a measure of type II error rate across datasets. In contrast, NDR and $1 - \text{NDR}$ are quantities of particular interest to practitioners.

For illustration, we assume $\sqrt{n} \bar{X}/\sigma$ is used to test the mean of a normal population ($H_0 : \theta = 0$ vs. $H_1 : \theta > 0$) with known variance $\sigma^2 = 5^2$, based on $n = 100$ i.i.d. samples. The power to detect a single hypothesis at level α is then $\Phi(\Phi^{-1}(\alpha) + \sqrt{n} \theta/\sigma)$ where Φ is the cdf of $N(0, 1)$. We first assume that there are 1,000 hypotheses among which 100 are from the alternative population, and 50 have $\mu_1 = 1$ while the remaining 50 have $\mu_1 = 1.5$. To control FDR at 5% and 10%, one can reject all hypotheses with (unadjusted) p-values ≤ 0.0022 and 0.0062 respectively (Sun et al. (2006)). Table A summarizes the results.

Suppose that an additional set of 1,000 hypotheses would be included. It is likely that the second set contains fewer true hypotheses. For example, in an

Table A: NDR and FNR when there are 900 null and 100 alternative hypotheses.

	Control FDR at 5%		Control FDR at 10%		Total
	Declared non-significant	Declared significant	Declared non-significant	Declared significant	
Truth: H_0	898	2	894.5	5.5	900
Truth: H_1	62	38	50	50	100
Total	960	40	944.5	55.5	1000
	NDR = 0.62		NDR = 0.50		
	FNR = 0.06		FNR = 0.05		

Table B: NDR and FNR when there are 1890 null and 110 alternative hypotheses.

	Control FDR at 5%		Control FDR at 10%		Total
	Declared non-significant	Declared significant	Declared non-significant	Declared significant	
Truth: H_0	1888.3	1.7	1885.2	4.8	1890
Truth: H_1	78	32	66.5	43.5	110
Total	1966.3	33.7	1951.7	48.3	2000
	NDR = 0.71		NDR = 0.60		
	FNR = 0.04		FNR = 0.03		

exploratory analysis of gene-expression data a large number of secondary genes might be added to the set containing high priority genes; in genome-wide linkage and association studies, a large number of genetic markers are included to cover the genome in addition to the ones selected from targeted regions. Assume that there are in fact only 10 alternatives among which 5 have $\mu_1 = 1$ and the remaining 5 have $\mu_1 = 1.5$. Because the proportion of the noise as measured by π_0 is greater, it would be more difficult to identify the alternatives. In other words, a more stringent criterion is required to control FDR at the same level. (Controlling FDR at 5% and 10% is equivalent to rejecting all hypotheses with unadjusted p-values ≤ 0.0009 and 0.0026 respectively.) This is reflected by the NDR measure which increases from 0.62 to 0.71 for FDR at the 5% level, and from 0.5 to 0.6 for FDR at 10%. In contrast, FNR decreases from Table A to Table B, a result of increased m_0 rather than a true reduction of false negatives.

Acknowledgement

We sincerely thank David Andrews, Dan Nicolae, Nancy Reid, two anonymous referees and an associate editor for their helpful comments and suggestions as well as for bringing to our attention some of the existent literature. The

COGA data were provided by the Collaborative Study on the Genetics of Alcoholism (U10AA008401). The CSGA data were provided by the investigators of the Collaborative Study on the Genetics of Asthma. We also would like to thank the Genetic Analysis Workshop Advisory Committee for making the data available. This work was supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) grants to each author.

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.
- Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Ed. Behav. Statist.* **25**, 60-83.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Amer. Statist.* **29**, 1165-1188.
- Black, M. A. (2004). A note on the adaptive control of false discovery rates. *J. Roy. Statist. Soc. Ser. B* **66**, 297-304.
- Cox, D. R. and Wong, M. Y. (2004). A simple procedure for the selection of significant effects. *J. Roy. Statist. Soc. Ser. B* **66**, 395-400.
- Delongchamp, R. R., Bowyer, J. F., Chen, J. J. and Kodell, R. L. (2004). Multiple-testing strategy for analyzing cdna array data on gene expression. *Biometrics* **60**, 774-782.
- Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71-103.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96-104.
- Efron, B. (2005). Correlation and large-scale simultaneous significance testing. Manuscript *Department of Statistics, Stanford University*.
- Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.* **30**, 220-238.
- Genovese, C. and Wasserman, L. (2001). A large-sample approach to controlling false discovery rates. *Tech. Report, Department of Statistics, Carnegie Mellon University*.
- Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64**, 499-517.
- Langaas, M., Lindqvist, B. H. and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. Roy. Statist. Soc. Ser. B* **67**, 555-572.
- Li, S. S., Bigler, J., Lampe, J. W., Potter, J. D. and Feng, Z. (2005). Fdr-controlling testing procedures and sample size determination for microarrays. *Statist. Medicine* **24**, 2267-2280.
- McPeck, M. S. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**, 1076-1094.
- Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99**, 990-1001.
- Owen, A. B. (2005). Variance of the number of false discoveries. *J. Roy. Statist. Soc. Ser. B* **67**, 411-426.

- Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **64**, 479-498.
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013-2035.
- Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. Ser. B* **66**, 187-205.
- Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440-9445.
- Sun, L., Abney, M. and McPeck, M. S. (2001). Detection of misspecified relationships in inbred and outbred pedigrees. *Genet. Epidemiol.* **21**, S36-S41.
- Sun, L., Craiu, R., Paterson, A. and Bull, S. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet. Epidemiol.* **30**, 519-530.
- Thomas, D. C., Siemiatycki, J. and Dewar, R. (1985). The problem of multiple inference in studies designed to generate hypotheses. *Am. J. Epidemiol.* **122**, 1080-1095.
- Tsai, C., Wang, S., Chen, D. and Chen, J. J. (2004). Sample size for gene expression microarray experiments. *Bioinformatics* **21**, 1502-1508.
- Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference* **82**, 171-196.

Department of Statistics, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada.

E-mail: craiu@utstat.toronto.edu

Departments of Public Health Sciences and Statistics, University of Toronto, Program in Genetics and Genomic Biology, Hospital for Sick Children Research Institute, 155 College Street, Toronto, ON M5T 3M7, Canada.

E-mail: sun@utstat.toronto.edu

(Received December 2005; accepted June 2006)