Stratified False Discovery Control for Large-Scale Hypothesis Testing with Application to Genome-Wide Association Studies

Lei Sun,[1–3]* Radu V. Craiu,[3] Andrew D. Paterson[1,2] and Shelley B. Bull[1,4]

[1]*Department of Public Health Sciences, University of Toronto, Toronto, Canada*
[2]*Program in Genetics and Genomic Biology, Hospital for Sick Children, Toronto, Canada*
[3]*Department of Statistics, University of Toronto, Toronto, Canada*
[4]*Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada*

The multiplicity problem has become increasingly important in genetic studies as the capacity for high-throughput genotyping has increased. The control of False Discovery Rate (FDR) (Benjamini and Hochberg. [1995] J. R. Stat. Soc. Ser. B 57:289–300) has been adopted to address the problems of false positive control and low power inherent in high-volume genome-wide linkage and association studies. In many genetic studies, there is often a natural stratification of the m hypotheses to be tested. Given the FDR framework and the presence of such stratification, we investigate the performance of a stratified false discovery control approach (i.e. control or estimate FDR separately for each stratum) and compare it to the aggregated method (i.e. consider all hypotheses in a single stratum). Under the fixed rejection region framework (i.e. reject all hypotheses with unadjusted p-values less than a pre-specified level and then estimate FDR), we demonstrate that the aggregated FDR is a weighted average of the stratum-specific FDRs. Under the fixed FDR framework (i.e. reject as many hypotheses as possible and meanwhile control FDR at a pre-specified level), we specify a condition necessary for the expected total number of true positives under the stratified FDR method to be equal to or greater than that obtained from the aggregated FDR method. Application to a recent Genome-Wide Association (GWA) study by Maraganore et al. ([2005] Am. J. Hum. Genet. 77:685–693) illustrates the potential advantages of control or estimation of FDR by stratum. Our analyses also show that controlling FDR at a low rate, e.g. 5% or 10%, may not be feasible for some GWA studies. *Genet. Epidemiol*. 2006. © 2006 Wiley-Liss, Inc.

Key words: multiple comparisons; genome-scans; type I error; type II error; power; false discovery rate (FDR); stratified FDR

Contract grant sponsor: Natural Sciences and Engineering Research Council of Canada (NSERC); Contract grant sponsor: Canadian Institutes of Health Research (CIHR); Contract grant sponsor: National Institutes of Health subcontract; Contract grant number: N01-6-2204.

*Correspondence to: Lei Sun, Department of Public Health Sciences, 155 College Street, 6th Floor, Toronto, Ot. M5T 3M7, Canada.
E-mail: lei.sun@utoronto.ca

Received 17 January 2006; Accepted 7 May 2006
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20164

INTRODUCTION

When a large number of hypotheses are tested, it is necessary to control the occurrence of type I errors/false positives. The traditional approach is to control the Family-Wise Error Rate (FWER), which is the probability of making even one type I error. For example, the well-known genome-wide significance level of 2.2×10^{-5} (LOD score of 3.6) for linkage mapping of complex diseases using an ASP design [Lander and Kruglyak, 1995] was proposed to control FWER at 0.05. That is, statistically significant evidence is expected to occur 1 in 20 times at random in a genome scan. Although the FWER method strictly guards against false positives, the corresponding power is typically extremely low especially for a large number of tests, resulting in few or no discoveries. The seminal work of Benjamini and Hochberg [1995] provides an alternative framework by controlling the False Discovery Rate (FDR), which is the expected proportion of false discoveries among all positives. If all the hypotheses are truly null, it can be shown that FDR and FWER are identical. However, if this (unlikely) equality does not hold, then controlling FDR imposes a less stringent condition than controlling FWER so an increase in power is expected. However, one must be aware that a direct comparison of power is not appropriate because the two procedures control different levels of type I error rate. Note that the value of FWER corresponding to a rejection procedure that controls FDR at level γ is typically considerably higher than γ. For example, in a simulation study of QTL linkage analyses performed by Benjamini and Yekutieli [2005], the actual FWER is about 0.63 while FDR was controlled at 0.05. Improvements and extensions of the method of Benjamini and Hochberg [1995] have been proposed by Benjamini and Hochberg [2000], Benjamini and Yekutieli [2001], Storey [2002, 2003] and Genovese and Wasserman [2001, 2002].

In many applications, the notion of FDR is more relevant than FWER. For example, in an exploratory analysis of a dataset in which m tests are to be performed and a portion of them is suspected to be statistically significant, one is ready to accept that some of the rejections are in fact false as long as some or most of the true signals have been discovered. In this respect, controlling FDR seems to be more appropriate than FWER. The use of FDR has become common in analyses of microarray gene expression data, where m is typically large, and one is often less concerned about making a type I error. Recently, statistical analyses of genetic data have begun to adopt the FDR framework, in part due to (a) the traditional genome-wide significance criterion often leads to no findings, (b) the number of SNPs and microsatellites being genotyped has increased drastically because of reduced cost and new study designs such as Genome-Wide Association (GWA) studies, (c) multiple phenotypes, covariates and models are being tested. In addition, in the context of multi-stage analyses, the idea of using FDR as a screening tool is appealing.

In many studies, particularly in GWA studies, there is an inherent stratification of the m tests to be performed. For example, each marker might be tested for association with each of K phenotypes of interest; tests might be conducted assuming K different genetic models (e.g. additive, dominant or recessive); a group of markers might be considered high-priority candidates because they were selected from candidate genes or linkage regions, while the remaining markers are included to cover the genome and treated as a secondary group; SNPs and microsatellites could be analyzed separately; or markers could be stratified based on their allele frequencies. Given the FDR framework and the presence of such stratification, the main question we address in this report is whether it is advantageous to control or estimate FDR separately for each stratum.

MOTIVATING EXAMPLES

To motivate the subsequent methods development, we first consider a proposed genetic study of long-term complications of type 1 diabetes, the DCCT/EDIC (Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications) study [Boright et al., 2005]. The research plan incorporates a multi-stage approach, and the goal of stage one analysis is to screen about 1500 SNPs in candidate genes and identify as many truly associated SNPs as possible for follow-up studies. There are five phenotypes of interest including three retinal and two renal diabetic complications. In a project of this nature, we are faced with a multiple testing problem (here, 5×1500 for a total of 7,500 tests) and the FDR facilitates the choice and interpretation of significant results as well as the design and allocation of available genotyping resources.

Given the natural stratification of the data, i.e. each of the 1,500 SNPs will be tested for association with each of the five phenotypes, a relevant FDR question is whether one should control or estimate FDR separately for each phenotype (call it the stratified FDR approach) or for all phenotypes together (call it the aggregated FDR approach). Intuitively, if the underlying association structure is the same for all phenotypes, i.e. power to detect association is similar and the number of truly associated SNPs is similar among phenotypes, then the stratified FDR approach offers similar information as the aggregated FDR approach. However, if there are differences among phenotypes, then it is possible that the stratified FDR approach is superior to the aggregated one because the former incorporates auxiliary information via the phenotype indicator. For example, assume that power to detect each truly associated SNP is 90% at the 0.01 level, but the number of such SNPs varies among the phenotypes as shown in Table I(a). Then, if one is to reject all SNPs with p-values less than 0.01, assuming all tests are independent, some simple calculations provide the following results in Table I(a). That is, if FDR is estimated for all 7,500 tests, roughly half of the 152 significant results are expected to be false. However, if FDR is estimated separately for each phenotype, then one

TABLE I. The EDIC study example

	Aggregated	Phen. 1	Phen. 2	Phen. 3	Phen. 4	Phen. 5
♯SNPs	7,500	1,500	1,500	1,500	1,500	1,500
♯associated SNPs	86	50	20	10	5	1
(a) *Under the fixed rejection region framework with $\alpha = 0.01$ and $1-\beta(\alpha) = 90\%$*						
E[♯false positives]	74	14	15	15	15	15
E[♯true positives]	78	45	18	9	5	1
E[♯positives]	152	59	33	24	20	16
FDR	49%	24%	45%	63%	75%	94%
(b) *Under the fixed FDR framework with nominal FDR level at 10%*						
α used	0.0008	0.003	0.001	0.00045	0.00019	0.000023
$1-\beta(\alpha)$	67%	80%	70%	61%	52%	32%
E[♯false positives]	6	4	1.5	0.6	0.3	0.03
E[♯true positives]	58	40	14	6	2.7	0.32
E[♯positives]	64	44	15.5	6.6	3	0.35

E[♯false positives] = (♯SNPs−♯associated SNPs) × α, E[♯true positives] = ♯associated SNPs × ($1-\beta(\alpha)$), where in (a), α is pre-specified at 0.01 level and $1-\beta(\alpha)$ (power) is assumed to be 90% for $\alpha = 0.01$, and in (b), α is the largest value that satisfies equation (5) given in the text, and we assume $1-\beta(\alpha) = \Phi(\Phi^{-1}(\alpha)+3.6)$, where Φ is the cumulative probability function for the standard normal distribution.

expects 45 of the 59 rejections from phenotype 1 to be truly associated SNPs, while one is now almost certain that all 16 significant SNPs from phenotype 5 are likely to be false. The gain of information through this simple stratified FDR estimation approach is evident.

The second example arises in the context of GWA studies. Suppose one performs a GWA study with 105,000 SNPs, among which 5,000 (stratum 1) are from candidate genes or regions of linkage and the rest of the 100,000 (stratum 2) are included to systematically scan the genome for the purpose of identifying novel associations. Assume that the power to detect each truly associated SNP is 70% at the 0.001 level, and the number of associated SNPs is 100 for each of the two strata. Then, if one is to reject all SNPs with p-values less than 0.001, assuming all tests are independent, we have the results shown in Table II(a). In this case, we expect 43% of the 245 positives to be false under the aggregated approach. Among the 245 positives, 75 belong to stratum 1 with stratum-specific FDR 7%, and the remaining 170 belong to stratum 2 with stratum-specific FDR 59%.

TABLE II. The GWA study example

	Aggregated	Stratum 1	Stratum 2
♯SNPs	105,000	5,000	100,000
♯associated SNP	200	100	100
(a) *Under the fixed rejection region framework with $\alpha = 0.001$ and $1-\beta(\alpha) = 70\%$*			
E[♯false positives]	105	5	100
E[♯true positives]	140	70	70
E[♯positives]	245	75	170
FDR	43%	7%	59%
(b) *Under the fixed FDR framework with nominal FDR level at 10%*			
α used	0.00009	0.0016	0.00004
$1-\beta(\alpha)$	44%	74%	37%
E[♯false positives]	9	8	4
E[♯true positives]	88	74	37
E[♯positives]	97	83	41

E[♯false positives] = (♯SNPs−♯associated SNPs) × α, E[♯true positives] = ♯associated SNPs × ($1-\beta(\alpha)$), where in (a), α is pre-specified at 0.001 level and $1-\beta(\alpha)$ (power) is assumed to be 70% for $\alpha = 0.001$, and in (b), α is the largest value that satisfies equation (5) given in the text, and we assume $1-\beta(\alpha) = \Phi(\Phi^{-1}(\alpha)+3.6)$, where Φ is the cumulative probability function for the standard normal distribution.

The above two illustrations use the fixed rejection region method of Storey [2002, 2003]. That is, one rejects all tests with (unadjusted) p-values less than a pre-specified α level then estimates FDR among all positive results. In this case, the total number of rejections with stratification is the same as with aggregation. However, the stratum-specific FDR could be substantially closer to 0 (all positives are true) or 1 (all positives are false), depending on the proportion of null hypotheses and the power to detect true signals in that stratum.

Alternatively, one may wish to control FDR at a tolerable level and meanwhile reject as many hypotheses as possible. This can be achieved with the method of Benjamini and Hochberg [1995] and its extensions. In this case, the FDR level γ is pre-specified, but the α level used for each stratum and overall may vary to ensure the nominal FDR level. As a result, the total number of rejections may differ between the stratified and the aggregated approaches. Consider the above EDIC and GWA studies and assume power to detect each truly associated SNP follows a normal model as described in the following METHODS section. Then if FDR is chosen at 10%, we have the results shown in Table I(b) and Table II(b). It is clear from this example that the stratified FDR approach is advantageous in that the expected total number of true positives is greater than for the aggregated approach while FDR is controlled at 10% in each stratum and overall: 63 versus 58 for the EDIC example, and 111 versus 88 for the GWA example.

In the following section, we derive analytical expressions for the aggregated and stratum-specific FDR as functions of the proportion of true null hypotheses, π_0, the level used to declare significance for each test α, and the power to detect a true alternative at level α, under the assumption that all tests are independent. The independence assumption allows for analytical derivation and clear understanding of the problem but it is not crucial to the conclusion. Under the fixed rejection region framework, we demonstrate that the aggregated FDR is a weighted average of the stratum-specific FDRs, with weights proportional to the expected number of rejections. Under the fixed FDR framework, we specify the condition necessary for the expected total number of true positives under the stratified FDR method to be equal to or greater than that obtained from the aggregated FDR method. We argue that the inequality often holds for genetic studies. We also briefly describe available methods for estimation of π_0 necessary for both frameworks and of FDR for the fixed rejection region framework.

METHODS

NOTATION

Table III summarizes the underlying events when a large number of hypotheses are examined. Note that m is the total number of tests, fixed in advance; m_0 is the number of true null hypotheses and m_1 is the number of true alternative hypotheses, and both m_0 and m_1 are unknown parameters. **R** is the observed number of positives for a given rejection procedure; **U**, **V**, **T** and **S** are all unobserved random variables and are, respectively, the number of true negatives, false positives, false negatives and true positives. Without loss of generality, we assume that the first m_0 hypotheses are true nulls and the rest are true alternatives, and let p_i, $i=1,\ldots,m$ be the unadjusted p-values for the m tests. In addition, let $\pi_0=m_0/m$ and $\pi_1=m_1/m$, and α be the level to declare significance for a single test based on p_i and $1-\beta_i(\alpha)$ be the corresponding power. Note that power for each of the m_1 alternatives may differ at the given α level, while power $=\alpha$ for the m_0 true nulls. If there are K strata among the m tests, we then use superscript $^{(k)}$ to denote the corresponding m, m_0, m_1, π_0, π_1, α and $\beta_i(\alpha)$ for each stratum k, $k=1,\ldots,K$.

TABLE III. Summary of events for multiple hypothesis testing

	Declared non-significant	Declared significant	Total counts
Truth: H_0	**U**	**V**	m_0
Truth: H_1	**T**	**S**	m_1
Total	$m-$**R**	**R**	m

FRAMEWORK I: FIXED REJECTION REGION

We first investigate the stratified approach under the fixed rejection region framework. This framework rejects all tests with p-values less than a chosen α level, then estimates FDR among all the positives. Given the above notation and the independence assumption, some simple calculations show that

$$\mathrm{E}[\mathbf{V}] = \sum_{i=1}^{m_0} \Pr(p_i \le \alpha) = m_0\,\alpha = m\,\pi_0\,\alpha$$

$$\mathrm{E}[\mathbf{S}] = \sum_{i=m_0+1}^{m} \Pr(p_i \le \alpha) = \sum_{i=m_0+1}^{m} (1-\beta_i(\alpha)) = m_1(1-\overline{\beta(\alpha)}) = m(1-\pi_0)(1-\overline{\beta(\alpha)})$$

where $1-\overline{\beta(\alpha)} = 1-\sum_{i=m_0+1}^{m}\beta_i(\alpha)/m_1$, the average power to detect m_1 alternatives, each at level α. Storey [2003] has shown that $\mathrm{E}[\mathbf{V}/\mathbf{R}]=\mathrm{E}[\mathbf{V}]/\mathrm{E}[\mathbf{R}]$ for independent tests, thus the aggregated FDR is

$$\mathrm{FDR} = \mathrm{E}\left[\frac{\mathbf{V}}{\mathbf{R}}\right] = \frac{\mathrm{E}[\mathbf{V}]}{\mathrm{E}[\mathbf{R}]} = \frac{m\,\pi_0\,\alpha}{m\,\pi_0\,\alpha + m\,(1-\pi_0)\,(1-\overline{\beta(\alpha)})} = \frac{\pi_0\,\alpha}{\pi_0\,\alpha + (1-\pi_0)(1-\overline{\beta(\alpha)})} = \frac{1}{1+(1/\pi_0-1)(1-\overline{\beta(\alpha)})/\alpha}. \quad (1)$$

From the above expression, it is apparent that FDR increases as π_0 increases. FDR also increases if power decreases for a given α level. However, FDR does not necessarily increase as α increases, unless the rate of change in power is slower than that in α as measured by $\{\alpha_1(1-\overline{\beta(\alpha_2)})\}/\{\alpha_2(1-\overline{\beta(\alpha_1)})\}$. Assume for example $\pi_0 = 0.9$, if power is 10% and 50%, respectively, at α levels of 0.001 and 0.005 for each of the m_1 alternatives, then the two corresponding FDRs are both 8.25% if one rejects all tests with p-values less than 0.001 or 0.005. If power is 10% and 80% instead, then using 0.005 leads to a smaller FDR of 5.33%. However if power is 20% and 50%, then using 0.001 leads to a smaller FDR of 4.32%. This non-monotonicity was also observed in the following application to Maraganore et al. [2005].

Suppose the m tests can be naturally stratified into K groups, among the $\mathbf{R} = \sum_k \mathbf{R}^{(k)}$ rejections, the stratified FDR approach would estimate FDR separately for each $\mathbf{R}^{(k)}$. The true FDR for the kth stratum is

$$\mathrm{FDR}^{(k)} = \mathrm{E}\left[\frac{\mathbf{V}^{(k)}}{\mathbf{R}^{(k)}}\right] = \frac{\mathrm{E}[\mathbf{V}^{(k)}]}{\mathrm{E}[\mathbf{R}^{(k)}]} = \frac{m^{(k)}\pi_0^{(k)}\alpha}{m^{(k)}\pi_0^{(k)}\alpha + m^{(k)}\cdot(1-\pi_0^{(k)})(1-\overline{\beta(\alpha)}^{(k)})}. \tag{2}$$

The aggregated FDR can be re-expressed in the following way:

$$\mathrm{FDR} = \mathrm{E}\left[\frac{\sum_k \mathbf{V}^{(k)}}{\sum_j \mathbf{R}^{(j)}}\right] = \frac{\sum_k \mathrm{E}[\mathbf{V}^{(k)}]}{\sum_j \mathrm{E}[\mathbf{R}^{(j)}]} = \sum_k w^{(k)}\mathrm{FDR}^{(k)} \tag{3}$$

where $w^{(k)} = \mathrm{E}[R^{(k)}]/\sum_j \mathrm{E}[R^{(j)}]$, the weighting factor for each stratum.

From equations (2) and (3), we see that the aggregated FDR is a weighted average of the K stratum-specific FDRs, with weights proportional to the expected number of rejections in each stratum. If strata are homogeneous in that the proportion of the null hypotheses is the same, $\pi_0^{(k)} \equiv \pi_0$, and the average power to detect the alternatives is the same, $\overline{\beta(\alpha)}^{(k)} \equiv \overline{\beta(\alpha)}$, then $\mathrm{FDR}^{(k)} = \mathrm{FDR} = \pi_0\alpha/(\pi_0\alpha + (1-\pi_0)(1-\overline{\beta(\alpha)}))$ regardless of the number of tests in each stratum. However, if $\pi_0^{(k)}$ or $\overline{\beta(\alpha)}^{(k)}$ vary among strata, the stratum-specific FDR would differ from the aggregated FDR. In particular, the stratum-specific FDR is bigger than the aggregated FDR if a stratum has a larger proportion of null hypotheses or lower power to detect alternatives. Similarly, smaller $\pi_0^{(k)}$ or higher $1-\overline{\beta(\alpha)}^{(k)}$ leads to smaller FDR. Note that, in the above EDIC and GWA examples, we assumed that $\overline{\beta(\alpha)}^{(k)} \equiv \overline{\beta(\alpha)}$ but $\pi_0^{(k)}$ varied among strata. Indeed, the stratum with the smallest $\pi_0^{(k)}$ has the smallest FDR. In the GWA example, if the power to detect associated SNPs from candidate genes or linkage regions is in fact higher than the others, e.g. 90% in stratum 1 and 40% in stratum 2 both at the 0.001 level, then the advantage of the stratified FDR approach is even more clear: the expected number of total rejections is 235 with FDR 45%, among which 95 are from stratum 1 with FDR 5% and 140 from stratum 2 with FDR 71%.

FRAMEWORK II: FIXED FDR

We now consider the stratified FDR approach under the fixed FDR framework. This framework pre-specifies FDR at level γ then finds a rejection procedure that rejects as many tests as possible but controls the proportion of false discoveries at γ. (We use γ to denote the FDR level in distinction from α.) There have been a number of improvements and extensions of the original procedure of Benjamini and Hochberg [1995], including the FDR-adjusted p-value approach [Yekutieli and Benjamini, 1999] and the q-value method [Storey, 2002, 2003]. Here we use the q-value approach since it has been shown to be equivalent to the adaptive FDR-adjusted p-value approach [Craiu and Sun, 2006]. Roughly speaking, the q-value of an observed test statistic associated with a hypothesis H_i is the minimum possible FDR for calling H_i significant. Consequently, controlling FDR at level γ is equivalent to calling all tests with q-values $\leq\gamma$ significant. The estimates of the q-values can be obtained by the use of the following recursive formula:

$$\hat{q}_{(i)} = \min\left\{\frac{\hat{\pi}_0\, m\, p_{(i)}}{i}, \hat{q}_{(i+1)}\right\} \tag{4}$$

where $p_{(1)} \leq \cdots \leq p_{(m)}$ is the ordered sequence of the m available p-values, $\hat{q}_{(m)} = \hat{\pi}_0\, p_{(m)}$, and $\hat{\pi}_0$ is an estimate of π_0.

Under the fixed FDR framework, the question of interest is whether the stratified FDR approach rejects in total more hypotheses while it still controls FDR at the nominal level γ for each stratum, therefore identifying more true positives. To obtain the expected number of true positives, E[**S**], we investigate the connection between the fixed FDR framework and the fixed rejection region framework and use it to derive E[**S**]. We try to find the largest α and $\alpha^{(k)}$, $k = 1,\ldots,K$, such that the corresponding FDR is controlled at γ.

(Note that α is the level used to call a single test significant based on its unadjusted p-value.) Once α and $\alpha^{(k)}$ are obtained, we can then use the results from the fixed rejection region approach above. However, we note that in this case FDR is pre-specified at a given level, but the α level used may vary among strata and differ from the over all value. To control FDR at level γ for a set of tests, based on equation (1), α has the following expression:

$$\alpha = \frac{(1-\pi_0)\,\gamma\,(1-\overline{\beta(\alpha)})}{\pi_0\,(1-\gamma)}. \qquad (5)$$

The expression for $\alpha^{(k)}$ is identical to equation (5) with the addition of superscript $^{(k)}$ for π_0 and $\overline{\beta(\alpha)}$. The determination of α however depends on the average power function $1-\overline{\beta(\alpha)}$, in addition to the values of π_0 and γ.

The specific form of the power function depends on the test statistic used as well as the "distance" between the null and alternative models. For example, assume one uses the sample average of size n to test if the mean of a normally distributed population (with known variance σ^2) equals zero or greater, then $1-\beta(\alpha;\mu) = \Phi(\Phi^{-1}(\alpha) + \sqrt{n}\mu/\sigma)$, where $\mu > 0$ is the true mean and Φ is the cumulative probability function for the standard normal distribution. In the EDIC and GWA examples, the pairs of values for α and $1-\beta(\alpha)$ given in Table I(b) and Table II(b) were determined assuming the above normal model with $n = 100$, $\mu = 1.8$ and $\sigma = 5$. In that case, $\gamma = 0.1$ and the α used is the largest value subject to equation (5) for the respective π_0 value. It is clear from equation (5) that α decreases as π_0 increases, given γ and a power function. That is, when the proportion of the null hypotheses increases, a more stringent significance criterion is required to control FDR at the nominal level. This can be also seen from equation (4) where the q-value increases as π_0 increases.

Let α be the level obtained for the aggregated data, and $\alpha^{(k)}$ for the kth stratum as described, all controlling FDR at level γ. Then the expected total number of true positives is

$$E[\mathbf{S}] = m_1(1-\overline{\beta(\alpha)})$$

and

$$E[\sum_k \mathbf{S}^{(k)}] = \sum_k m_1^{(k)}(1-\overline{\beta(\alpha^{(k)})}^{(k)})$$

respectively, for aggregated and stratified approaches, and

$$E[\sum_k \mathbf{S}^{(k)}] \geq E[\mathbf{S}] \Longleftrightarrow \sum_k \frac{m_1^{(k)}}{m_1}(1-\overline{\beta(\alpha^{(k)})}^{(k)}) \geq (1-\overline{\beta(\alpha)}).$$

Therefore, $E[\sum_k \mathbf{S}^{(k)}] \geq E[\mathbf{S}]$ if the weighted average of power across strata is equal to or greater than the aggregated one, with weights proportional to the number of true alternatives in each stratum. Due to the complex interplay between α, π_0 and γ and the unknown average power function $\{1-\overline{\beta(\alpha)}\}$, it is difficult to show that this inequality always holds. However, by the use of equation (5), we have $\{1-\overline{\beta(\alpha)}\} = \{\alpha\,\pi_0(1-\gamma)\}/\{(1-\pi_0)\gamma\}$ and $\{1-\overline{\beta(\alpha^{(k)})}^{(k)}\} = \{\alpha^{(k)}\pi_0^{(k)}(1-\gamma)\}/\{(1-\pi_0^{(k)})\gamma\}$. Thus, if we replace $\{1-\overline{\beta(\alpha)}\}$ and $\{1-\overline{\beta(\alpha^{(k)})}^{(k)}\}$ above, we have the following:

$$E[\sum_k \mathbf{S}^{(k)}] \geq E[\mathbf{S}] \Longleftrightarrow \sum_k \frac{m_0^{(k)}}{m_0}\alpha^{(k)} \geq \alpha. \qquad (6)$$

Similar to the fixed rejection region framework, if the strata are homogeneous, $\pi_0^{(k)} \equiv \pi_0$ and $\overline{\beta(\alpha)}^{(k)} \equiv \overline{\beta(\alpha)}$, then $\alpha^{(k)}$ used for each stratum is the same as α and $\sum_k (m_0^{(k)}/m_0)\alpha^{(k)} = \alpha$. Therefore, $E[\sum_k \mathbf{S}^{(k)}] = E[\mathbf{S}]$. However, if $\pi_0^{(k)}$ or the power functions differ, the above inequality may or may not hold, depending on the specification of the strata and the underlying power function. If for example we assume the $\pi_0^{(k)}$ vary but the power function is the same, as in the EDIC and GWA examples, then $\alpha^{(k^*)} > \alpha$ if $\pi_0^{(k^*)} < \pi_0$ for stratum k^*. Since the rate of change in α is typically much faster than that in π_0, we expect $\sum_k (m_0^{(k)}/m_0)\alpha^{(k)} > \alpha$. If we assume that the power functions vary but $\pi_0^{(k)}$ is the same, then $\alpha^{(k^*)} > \alpha$ if power for stratum k^* is stochastically greater than the overall power. Thus, in this case we also expect $\sum_k (m_0^{(k)}/m_0)\alpha^{(k)} > \alpha$.

Under some extreme situations, it is possible that stratification may lead to fewer rejections, For instance, in the case when $\pi_0^{(k^*)} < \gamma$ for stratum k^* (i.e. the proportion of the null hypotheses in a stratum is smaller than the pre-specified FDR level), rejecting all tests controls FDR at a level ($\pi_0^{(k^*)}$) smaller than the pre-specified one (γ). This "un-used" portion of FDR (i.e. $\gamma - \pi_0^{(k^*)}$) however could be fully utilized under aggregation resulting in more rejections. For example, assume that $m_1 = 5{,}000$, $m_1^{(1)} = 4{,}750$ and $m_1^{(2)} = 250$, $m_0 =$

5,000, $m_0^{(1)} = 250$ and $m_0^{(2)} = 4750$ ($\pi_0 = 0.5$, $\pi_0^{(1)} = 0.05$ and $\pi_0^{(2)} = 0.95$), power to detect each alternative is $\Phi(\Phi^{-1}(\alpha) + 3.6)$, and FDR is fixed at $\gamma = 20\%$. In that case, $\alpha = 0.25$, $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 0.012$ (corresponding power is 99.8%, 100% and 91%), and the expected true positives are respectively 4,990, 4,750 and 227. Indeed, stratification results in 13 fewer true positives. However, it could be argued that this is not a fair comparison in that FDR is controlled at 5% for stratum 1, much lower than the given 20%. In addition, such an extreme case, where the proportion of null hypotheses is lower than the pre-specified FDR level, is neither practically plausible (for $\pi_0 > 0.5$ is expected for most if not all cases of large-scale hypothesis testing, with $\pi_0 > 0.9$ for GWA studies), nor meaningful (for one would reject all hypotheses in that case, while the usual goal in large-scale hypothesis testing is to narrow down all the hypotheses to a smaller set of interesting ones).

ESTIMATION OF π_0 AND FDR

The above analytical derivation demonstrates the expected performance of FDR control and estimation procedures. For a given set of hypotheses, π_0 needs to be estimated if the fixed FDR framework is used, and both π_0 and FDR must be estimated if the fixed rejection framework is adopted.

Several estimators of π_0 have been proposed, among which

$$\hat{\pi}_0(\lambda) = \#\{p_i > \lambda\}/(m\,(1-\lambda))$$

with $\lambda = 0.5$ is most commonly used. It has been noted that $\hat{\pi}_0(\lambda)$ is biased upward. Although the bias decreases as λ increases, the variance of $\hat{\pi}_0(\lambda)$ increases which makes the estimate less reliable. To balance between bias and variance, Storey and Tibshirani [2003] proposed a more sophisticated estimation method: first calculate the above $\hat{\pi}_0(\lambda)$ for a range of λ, $\lambda = 0,0.01,0.02,\ldots,0.95$, then fit the natural cubic spline $\hat{f}(\lambda)$ of $\hat{\pi}_0(\lambda)$ on λ with 3 d.f., and finally set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(\lambda = 1)$.

Under the fixed rejection region framework, the following estimator could be used to estimate FDR among the **R** positives [Storey and Tibshirani, 2003]:

$$\widehat{\mathrm{FDR}}(\alpha) = (m\,\hat{\pi}_0\,\alpha)/\#\{p_i \leq \alpha\}$$

where $\mathbf{R} = \{\# p_i \leq \alpha\}$, and $m_0\,\hat{\pi}_0\,\alpha$ is an estimate of the number of false positives using the fact that the null (independent) p-values follow the Unif (0,1) distribution. To bound the FDR estimate below 1, one can use

$$\widehat{\mathrm{FDR}}(\alpha) = \min\{(m\,\hat{\pi}_0\,\alpha)/\#\{p_i \leq \alpha\}, 1\}. \qquad (7)$$

APPLICATION

Maraganore et al. [2005] recently reported a two-stage GWA study of Parkinson's disease. 198,345 SNPs, uniformly spaced across the genome, were analyzed in stage 1 using a discordant sib-pair design. A total of 443 case-unaffected sibling pairs were analyzed using the sibling transmission/disequilibrium test (sTDT) method [Schaid and Rowland, 1998], adjusting the analyses for age and sex. After exclusion of SNPs with fewer than nine discordant pairs, Maraganore et al. [2005] identified 1,862 SNPs with p-values ≤ 0.01. These positive SNPs were then followed up in stage 2 using a different study design with 332 matched case-unrelated control pairs and an additional 300 SNPs for genomic control. For the purpose of this paper, we focused on their stage 1 and obtained detailed information about the stage 1 SNPs, including allele frequencies in controls and p-values, from their text file 1 (online only, the American Journal of Human Genetics). Note that Maraganore et al. [2005] essentially used a fixed rejection region approach with $\alpha = 0.01$ although they did not estimate the corresponding FDR.

Figure 1(a) shows the histogram of 197,222 p-values that were available. The minimum p-value is 1.3×10^{-5} and the maximum is 1. The estimate of π_0 is 0.9764 using the method of Storey and Tibshirani [2003] as described above. There were 1,906 SNPs with p-values ≤ 0.01. The number is slightly larger than the 1,862 of Maraganore et al. [2005] because they excluded SNPs with fewer than nine discordant pairs, using information that was not available to us. The estimate of the aggregated FDR based on equation (7) turns out to be 100%. Although it is unlikely that all the 1,906 positives are false (sampling variation and inaccuracy in the estimate of π_0 could lead to an upward biased estimate of FDR), the message is clear: we expect only a few, if any, of the 1,906 positives to be true discoveries. This is not surprising because of the large proportion of null hypotheses (i.e. $\pi_0 = 0.9764$) and the possible underlying low power to detect any associated SNP. In fact, the smallest estimated q-value is 0.7684, and even with a much more stringent

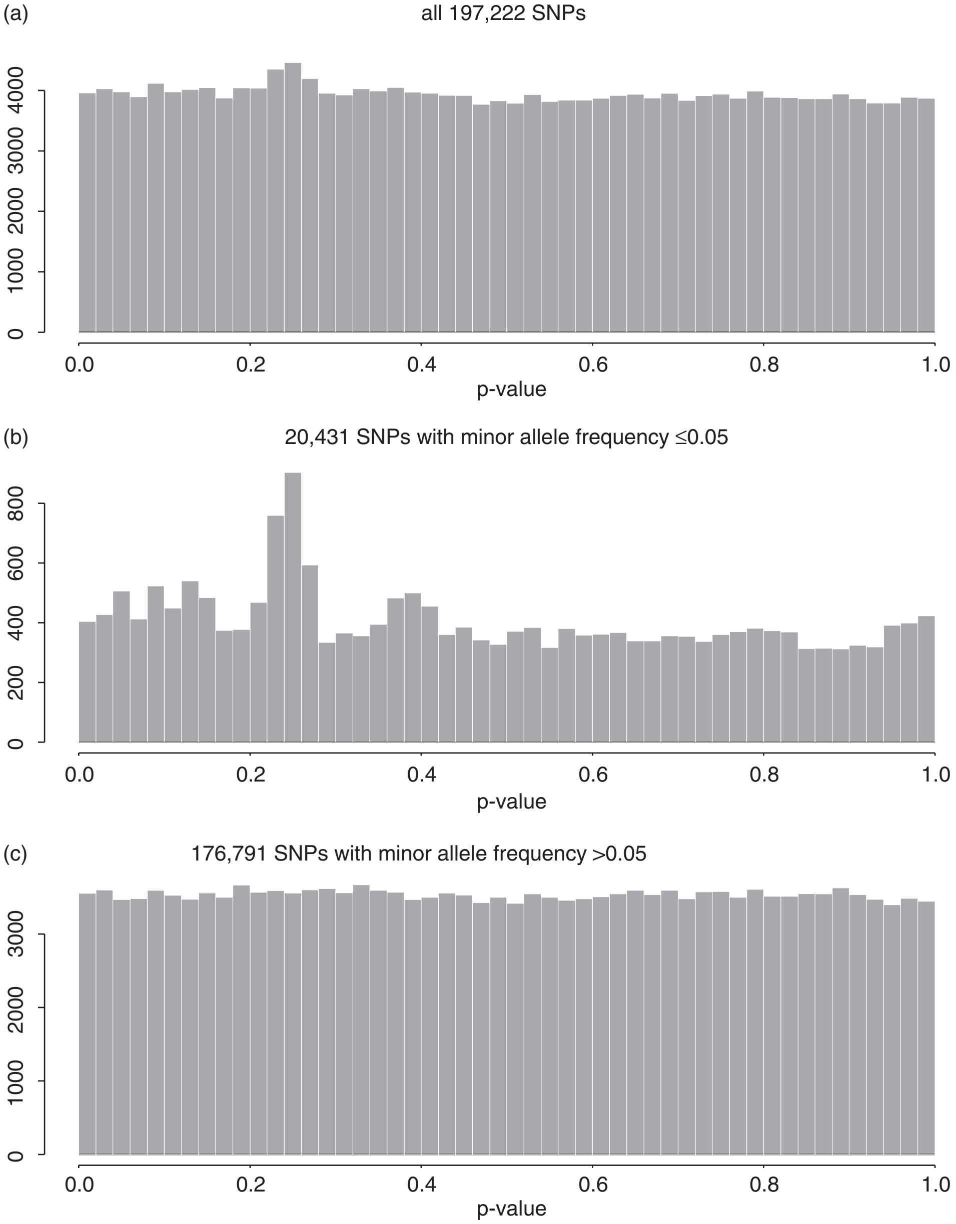

Fig. 1. Histogram of *p*-values for SNPs tested by Maraganore et al. [2005]: (a) all 197,222 SNPs; (b) 20,431 SNPs with minor allele frequency ≤ 0.05; and (c) 176,791 SNPs with minor allele frequency >0.05.

criterion of $\alpha = 0.0001$, FDR is estimated to be 80% among the 24 positives detected without stratification.

There has been some discussion concerning whether common or rare variants are responsible for susceptibility to complex diseases [e.g. Pritchard, 2001]. In addition, it is known that allele frequency plays an important role in power determination of an associated SNP. Therefore, we stratified all SNPs into two strata based on the allele frequency distribution in controls. Stratum 1 contains 20,431 SNPs with minor allele frequency ≤ 0.05 and stratum 2 includes the remaining 176,791 SNPs. Figures 1(b) and (c) show the histograms of the corresponding *p*-values. The minimum and maximum *p*-values are 4.97×10^{-5} and 0.99981 for stratum 1, and 1.3×10^{-5} and 1 for stratum 2. The estimates of $\pi_0^{(k)}$ are 0.8640 and 0.9894, respectively, for strata 1 and 2. It is clear that the two strata are not homogeneous in that

TABLE IV. Application to Maraganore et al. [2005] under the fixed rejection region framework with $\alpha = 0.01$, 0.001 and 0.0001

	Aggregated	Stratum 1	Stratum 2
♯SNPs	197,222	20,431	176,791
$\hat{\pi}_0$	0.9764	0.8640	0.9894
minimal q-value	0.76	0.50	0.79
$\alpha = 0.01$			
♯rejections	1,906	195	1,711
$\widehat{FDR}$	100%	91%	100%
$\alpha = 0.001$			
♯rejections	216	15	201
$\widehat{FDR}$	89%	100%	87%
$\alpha = 0.0001$			
♯rejections	24	3	21
$\widehat{FDR}$	80%	59%	83%

the proportion of truly associated SNPs in strata 1 appears to be substantially larger than that in strata 2. This seems to contradict the "common disease, common variant" hypothesis, although we note that the SNPs of Maraganore et al. [2005] were selected based on their physical locations (uniformly spaced across the genome) rather than on linkage disequilibrium (LD) patterns.

We now estimate FDR for each stratum under the fixed rejection region approach. Among the total 1,906 positives, 195 were from stratum 1 and 1,711 from stratum 2. The estimated FDR for stratum 1 is 91% while the estimated FDR for strata 2 is 100% (Table IV). An astute reader may notice the discrepancy between the stratified and aggregated approaches in that the estimated total number of true positives is $195 \times (1–91\%)+0 \approx 18$ for the former and 0 for the latter. However, the 18 possible true positives leads to $1–18/1906 = 99\%$ aggregated FDR for which an estimate of 100% is not unlikely considering sampling variation. Nevertheless, there is a need for more accurate estimation of π_0 and FDR. The advantage of stratification is not clearly demonstrated by this example because the FDR estimate ≈ 1. In that case, stratification does not change the fact that all R rejections are expected to be false. Results for $\alpha = 0.0001$ (Table IV) however illustrate the benefit of stratification: among the 24 rejections with aggregated FDR of 80%, 3 belong to stratum 1 with a much lower stratum-specific FDR of 59%. In addition, the minimal attainable FDR is 76% under aggregation (minimal q-value is 0.76), while it is at a much lower level of 50% for stratum 1 but at only slightly higher level of 79% for stratum 2.

DISCUSSION

We have proposed a stratified false discovery control approach for genetic studies in which a large number of hypotheses have some inherent stratification among them. The proposed method is a simple way to incorporate available auxiliary information where the auxiliary variable is the stratum indicator I. In an ideal situation, if $I = 1$ for a true null hypothesis and $I = 2$ for a true alternative hypothesis, then for a given rejection procedure that rejects $\mathbf{R}$ $(= \mathbf{R}^{(1)} + \mathbf{R}^{(2)})$ hypotheses, FDR would be 1 for stratum 1 (i.e. all $\mathbf{R}^{(1)}$ are false positives) and 0 for stratum 2 (i.e. all $\mathbf{R}^{(2)}$ are true positives), while the aggregated FDR could be 50% (i.e. we expect half of the $\mathbf{R}$ positives to be false but cannot conclude which ones, a lack of specificity). If we split the hypotheses randomly (equivalent to an uninformative stratification variable), then we will gain nothing and possibly lose some efficiency because the smaller number of hypotheses in each stratum may reduce the precision in the estimation of π_0. Using the $\hat{\pi}_0(\lambda = 0.5)$ estimator above, and under the independence assumption, one can show that the bias $(= 2(1 - \pi_0)\varepsilon \leq 2\varepsilon$, where $\varepsilon = \Pr(p \geq 0.5|H_1)$, the probability that an alternative p-value is greater than 0.5) is not affected by the stratification, but the variance of the estimator is in the order of $o(1/m^{(k)})$. In the context of large-scale hypothesis testing, the increase in the variance is negligible as long as there are at least hundreds of tests in each stratum.

The accuracy of the auxiliary information is central to the improvements of the stratified false discovery control method. However, as demonstrated by both analytical results and application, substantial gains can be achieved by the use of stratification even if the stratification variable is far from the ideal I above, as long as the proportion of the null hypotheses and/or the power to detect the alternatives differs among strata. In practice, such variables are often inherent in the study design and data analyses, for instance the phenotype in the EDIC example above, the SNPs in candidate genes/linkage regions versus the remaining ones in the GWA study example above, or *cis* versus *trans* regulators (SNPs close to the gene of interest versus not) in the association analyses of SNP-to-gene expression data. Although these variables will not perfectly separate the noise from the signals, the resulting strata are likely to have different proportions of true signals and/or power to detect those signals.

In the application to Maraganore et al. [2005], we recognize that SNPs were selected based on their physical locations (uniformly spaced across the genome), and allele frequency was not part of the study design. Therefore, using allele frequency as the stratifying variable is less than ideal and more as means to demonstrate the proposed method. However, the choice was also motivated by the current discussions concerning whether common or rare variants are responsible for susceptibility to complex diseases and by the known effect of allele frequency on power of the analyses [e.g. Pritchard, 2001]. The MAF of 5% was chosen because it is commonly used as the cutoff point for rare variants. Although it is possible to adopt another threshold, yielding a different stratification, we note that choosing the stratifying variable that gives the ''optimal'' result creates another level of multiple hypothesis testing and leads to bias. In addition, we emphasize that the stratifying variable must be a covariate that is independent of the observed p-values.

The auxiliary information required for our method is rather general. In situations where previous genome-wide linkage results are available for GWA studies, Roeder et al. [2006] recently proposed a weighted p-value approach that treats the linkage results as prior or auxiliary information. The method applies the traditional FDR control procedure to weighted p-values of association tests, p_i/w_i, where p_i is the original p-value of an association test at marker i, and w_i is the weighting factor proportional to the linkage result at that marker. They showed that if linkage results are informative, the weighted method improves power considerably, and if linkage results are uninformative, the loss in power is small. As pointed out by Roeder et al. [2006], in addition to having adequate sample size, one does need to assume that detectable loci by linkage design are the same as those by association. Nevertheless, the weighted p-value approach is promising for GWA studies, and it emphasizes the fact that ''multiple testing inherent in broad genomic searches diminishes power to detect association, even for genes falling in regions of the genome favored a priori'' [Roeder et al., 2006. The stratified FDR approach and the weighted p-value method are complementary formulations that incorporate prior information to increase the power to detect signals in a targeted set of hypotheses. In general, one can combine the two methods by using the weighted p-value approach within the stratified FDR framework. For example, if linkage results were available for the GWA study of Maraganore et al. [2005], one could apply the weighted p-value method within each of the two strata. It remains an important research question to identify the optimal method that effectively utilizes all available auxiliary information.

In the application to Maraganore et al. [2005], we demonstrated the stratified method under the fixed rejection region framework. For illustration, we can also consider the fixed FDR framework. Since an FDR less than 77% is not attainable for aggregated data, with FDR $>50\%$ for stratum 1 and $>79\%$ for stratum 2, we choose $\gamma = 0.8$, although in practice FDR as high as 80% level is probably undesirable. In this case, q-values were first calculated based on equation (4) for all SNPs under aggregation or stratification. SNPs with q-values ≤ 0.8 were then rejected, a procedure that controls FDR at the 80% level [Storey, 2002, 2003]. As a result, 27 tests were rejected in the aggregated approach, while 13,151 tests were rejected in stratum 1 and 24 were rejected in stratum 2. Given limited resources, it might be a concern that extra $13{,}151+24-27=13{,}148$ SNPs must be further analyzed in stage 2. However, the stratified FDR approach also identifies $13{,}148 \times (1–80\%) = 2{,}630$ more potentially associated SNPs in stage 1 which was designed as a screening stage. Recently, Craiu and Sun [2006] have proposed a quantity called non-discovery rate (NDR), which is the proportion of false negatives among true alternatives, as a measure of type II error rate for multiple hypothesis testing (hence 1 - NDR as power, which is the proportion of true positives among true alternatives). They emphasized the importance of controlling type II error in addition to type I error and the advantages of jointly analyzing FDR and NDR. This is of particular interest for GWA studies in which the appropriate FDR and NDR levels are unclear and is the subject of on-going research. If resources are limited, the stratified FDR approach also allows the flexibility to choose different FDR levels for different stratum. For example, in the application to Maraganore et al. [2005], one could choose FDR at 65% for stratum 1 which leads to eight rejections.

The application to Maraganore et al. [2005] also shows that (a) an FDR level of 5% or 10% may not be reasonable or achievable for some GWA studies, (b) under the fixed rejection region framework, FDR level may decrease as α increases for certain range of α values. The low power associated with controlling FWER has been well

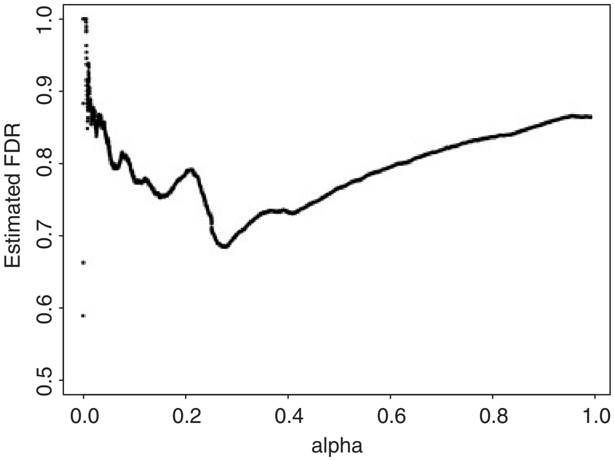

Fig. 2. Application to Maraganore et al. [2005]—Estimated FDR versus the α level under the fixed rejection region framework for the 176,791 SNPs in stratum 1.

acknowledged, and FDR was proposed as an alternative for a less stringent control of type I error rate. However, our analysis shows that when (i) the underlying signals are weak, (ii) the proportion of null hypotheses is close to 1 and (iii) there are large number of hypotheses to be tested, as is likely to be the case for GWA studies, one might have to accept a large value of FDR so that some of the true signals can be detected. This is consistent with the statement of Roeder et al. [2006] cited above. We have also observed that under the fixed rejection region framework, FDR does not necessarily have a monotone relationship with α. This appears to be the case for the stratum 1 of Maraganore et al. [2005] in which a greater density occurs around 0.25 in Figure 1(b) indicating a cluster of alternative hypotheses. (However, we note that correlation among null hypotheses could also cause such a cluster.) Figure 2 shows the corresponding estimated FDR level for α in the range of (0.0001, 0.99). Indeed, although the estimated FDR decreases as α increases from 0.0001 to 0.001, for $0.001<\alpha<0.25$ the estimated FDR seems to have a decreasing trend as α increases. A more in-depth investigation of such phenomena is of interest.

Current FDR methods work well for independent tests and tests with positive regression dependency (PRD) [Benjamini and Yekutieli, 2001]. Li et al. [2005] have investigated the impact of general dependency on FDR control. In their simulations of a microarray gene expression study, they demonstrated that the actual FDR could be twice the nominal level when the dependence structure among tests was generated under realistic assumptions, and if the proportion of null genes was greater than 90%. Unfortunately, $\pi_0>0.9$ is likely to be the case for genetic studies in which the proportion of truly linked or associated markers is small. However, Sabatti et al. [2003] showed analytically that PRD holds for linkage tests and their simulation studies demonstrated that FDR control is at the nominal level for association mapping from case-control data. In practice, when the PRD assumption is in doubt and $\hat{\pi}_0>0.9$, one may use a crude adjustment by using half of the nominal FDR level as suggested by Li et al. [2005]. However, further research to refine this correction is necessary. The recent work of Efron [2005] suggests that empirical null distributions [Efron, 2004] could be used as a more correlation-resistant FDR control technique.

ACKNOWLEDGMENTS

We sincerely thank Demetrius Maraganore and his colleagues for providing the detailed information about the SNPs in Maraganore et al. [2005] as on-line resources at the American Journal of Human Genetics, which made the application possible. We also thank two anonymous reviewers for their careful review of the manuscript and various helpful comments. This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.S., S.B.B. and R.V.C., and grants from the Canadian Institutes of Health Research (CIHR) to L.S., S.B.B. and A.D.P. S.B.B. holds a Senior Investigator Award from the CIHR, and A.D.P. holds a Canada Research Chair in Genetics of Complex Diseases. The EDIC study is supported by National Institutes of Health subcontract (N01-6-2204).

REFERENCES

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300.

Benjamini Y, Hochberg Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Statist 25:60–83.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188.

Benjamini Y, Yekutieli D. 2005. Quantitative trait loci analysis using the false discovery rate. Genetics 171:783–970.

Boright AP, Paterson AD, Mirea L, Bull SB, Mowjoodi A, Scherer SW, Zinman B, DCCT/EDIC Research Group. 2005. Genetic variation at the ACE gene is associated with persistent microalbuminuria and severe nephropathy in type 1 diabetes: the DCCT/EDIC Genetics Study. Diabetes 54:1238–1244.

Craiu RV, Sun L. 2006. Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. Stat Sinica (to appear).

Efron B. 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99:96–104.

Efron B. 2005. Correlation and large-scale simultaneous significance testing. Technical Report, Department of Statistics, Stanford University.

Genovese CR, Wasserman L. 2001. A large-sample approach to controlling false discovery rates. Technical Report, Department of Statistics, Carnegie Mellon University.

Genovese CR, Wasserman L. 2002. Operating characteristics and extensions of the false discovery rate procedure. J R Stat Soc Ser B 64:499–517.

Genovese CR, Roeder K, Wasserman L. 2005. False discovery control with p-value weighting. Technical Report, Department of Statistics, Carnegie Mellon University.

Lander E, Kruglyak L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247.

Li SS, Bigler J, Lampe JW, Potter JD, Feng Z. 2005. FDR-controlling testing procedures and sample size determination for microarrays. Stat Med 24:2267–2280.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocaa WA, Pant PVK et al. 2005. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685–693.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Roeder K, Bacanu SA, Wasserman L, Devlin B. 2006. Using linkage genome scans to improve power of association genome scans. Am J Hum Genet 78:243–252.

Sabatti C, Service S, Freimer N. 2003. False discovery rates in linkage and association genome screens for complex disorders. Genet 164:829–833.

Schaid DJ, Rowland C. 1998. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506.

Storey JD. 2002. A direct approach to false discovery rates. J R Stat Soc Ser B 64:479–498.

Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31:2013–2035.

Storey JD, Tibshirani R. 2003. Statistical significance for genomwide studies. Proc Natl Acad Sci USA 100: 9440–9445.

Yekutieli D, Benjamini Y. 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plann Infer 82:171–196.