False-negative-rate based approach selecting top single-nucleotide polymorphisms in the first stage of a two-stage genome-wide association study

Zhuying Huang[*,†], Jian Wang[*,†], Chih-Chieh Wu, Richard S. Houlston, Melissa L. Bondy and Sanjay Shete[†,‡]

Genome-wide association (GWA) studies, where hundreds of thousands of single-nucleotide polymorphisms (SNPs) are tested simultaneously, are becoming popular for identifying disease loci for common diseases. Most commonly, a GWA study involves two stages: the first stage includes testing the association between all SNPs and the disease and the second stage includes replication of SNPs selected from the first stage to validate associations in an independent sample. The first stage is considered to be more fundamental since the second stage is contingent on the results of the first stage. Selection of SNPs from stage one for genotyping in stage two is typically based on an arbitrary threshold or controlling type I errors. These strategies can be inefficient and have the potential to exclude genotyping of disease-associated SNPs in stage two. We propose an approach for selecting top SNPs that uses a strategy based on the false-negative rate (FNR). Using the FNR approach, we proposed the number of SNPs that should be selected based on the observed p-values and a pre-specified multi-testing power in the first stage. We applied our method to simulated data and a GWA study of glioma (a rare form of brain tumor) data. Results from simulation and the glioma GWA indicate that the proposed approach provides an FNR-based way to select SNPs using pre-specified power.

AMS 2000 SUBJECT CLASSIFICATIONS: Primary 60K35, 60K35; secondary 60K35.
KEYWORDS AND PHRASES: False negative rate, Single-nucleotide polymorphism, Two-stage genome-wide association study.

1. INTRODUCTION

Genome-wide association (GWA) studies have been shown to be a powerful approach to identify common variants for many complex diseases [1, 6, 9, 13, 22]. GWA studies are designed to identify common, low-penetrance disease

alleles without prior knowledge of their location and function [9]. Over the past few years, GWA studies have been applied to many different complex diseases and have identified a large number of genetic variants, such as those associated with coronary heart disease [6, 12, 19, 23], type 2 diabetes [25, 27, 30, 38], lung cancer [2, 15, 35], prostate cancer [10, 34], colorectal cancer [14, 36], melanoma [4], and glioma [28, 37].

In a GWA study typically hundreds of thousands of single-nucleotide polymorphisms (SNPs), which are the most common form of genetic variants, are genotyped using high-throughput technologies. Such analyses are costly and time consuming. Also, the large number of tests performed leads to a high proportion of false-positives. Most GWA studies are therefore based on multistage designs, in order to reduce the number of false-positive results, minimize the amount of genotyping performed, and retain power [13, 29]. Generally, for multistage designs, the investigator performs a genome-wide scan on an initial group of case and control participants and then replicates a much smaller number of associated SNPs in a second or third group of cases and controls. For example, in a type 2 diabetes study, Sladek et al. [30] selected 57 SNPs from the first-stage analyses, tested these SNPs in an independent sample of 2,617 cases and 2,894 controls in the second stage, and finally confirmed association with 8 SNPs using the combined results from the two stages. In a glioma study, 34 SNPs were prioritized as showing significant associations in the first stage. The investigators then conducted a replication study of these 34 SNPs in three case-control series that included 5,498 individuals in the second stage, and confirmed the association with 14 SNPs [28]. In a lung cancer study, Amos et al. [2] selected the top 10 SNPs from the first-stage analyses, tested these SNPs in an additional sample of 711 cases and 632 controls from Texas and 2,013 cases and 3,062 controls from the United Kingdom in the second stage, and confirmed 2 SNPs associated with the risk of lung cancer.

In GWA studies, investigators typically choose the top SNPs in the first stage and then test the selected SNPs for replication in an independent sample in the second stage. Therefore, in the first stage of a GWA study, the investigators hope to include disease-associated SNPs in the set of

SNPs that are to be replicated. So, it is important to decide the number of top SNPs to select in the first stage. If more SNPs are selected in the first stage, more genotyping is required in the second stage; if fewer SNPs are selected in the first stage, some of the SNPs that are potentially causal might not be included for replication in the second stage. Mostly, in the first stage of current GWA studies, investigators select top SNPs for replication based on an arbitrary cutoff p-value (e.g. p-value $\leq 10^{-5}$ as in the glioma GWA study [28]) or an arbitrarily fixed number of SNPs (e.g. 10 top SNPs as in the lung cancer GWA study [2]). If the investigators are selecting a very large number of SNPs (e.g. top 1,000 SNPs) or a very liberal significance level (e.g. p-value $\leq 10^{-2}$), they might include more disease-associated SNPs for replication. However, such strategies would require a large number of SNPs to be genotyped. Moreover, it may not be necessary to select such a large number of top SNPs. The alternative is to select the top SNPs using the approaches based on controlling type I errors in the stage one, such as the Bonferroni correction or Benjamini-Hochberg False Discovery Rate (BH-FDR) approach. These approaches can control the type I error very well, but are usually very conservative, and might exclude the potentially interesting (disease-associated) SNPs from the analysis of second stage. Therefore, instead of using arbitrary thresholds or using the criterion based on controlling the type I errors, the purpose of this paper is to propose a selection criterion for the first stage of GWA studies based on the power of the multiple testing.

In this paper, we propose an approach for selecting top SNPs from the first stage by using the false-negative rate (FNR). FNR is a measure of the type II error rate for multiple testing, which is defined as the expected proportion of falsely not-rejected hypotheses among all alternative hypotheses [20, 21]. Delongchamp et al. [7] called the same quantity the Fraction of Genes Not Selected in their study. This is the definition of FNR that we will use throughout our paper. It should be noted that some other investigators have defined FNR differently [11, 24, 33]. Genovese and Wasserman [11] defined the False Nondiscovery Rate as the proportion of non-rejections that are incorrect. This was also referred to as the False Negative Rate by Sarkar [24] and the Miss Rate by Taylor et al. [33].

Using the proposed FNR approach, we selected SNPs on the basis of observed p-values and pre-specified multi-testing powers, which is also defined as $(1 - \text{false-negative rate})$, in the first stage of a GWA study. To test the performance of the FNR-based SNP selection approach, we performed simulation studies under different scenarios. We compared our FNR-based approach to the fixed p-value cutoff approach, fixed number of SNPs cutoff approach, Bonferroni correction and BH-FDR approach. We also applied the proposed approach to the analysis of SNP genotype data from 1,247 glioma patients and 2,232 controls. Our results from the simulation and real data analyses show that the proposed ap-

Table 1. Description of the outcomes of m multiple tests, where m is the total number of statistical tests, m_0 is the number of true null hypotheses, m_1 is the number of alternative hypotheses, and r is the total number of hypotheses rejected at a significance level α

True hypothesis	Non-significant	Significant	Total
Null	$m_0 - r_0$	r_0	m_0
Alternative	$m_1 - r_1$	r_1	m_1
Total	$m - r$	r	m

proach provides an FNR-based criterion to select top SNPs in the first stage, while attaining adequate power.

2. METHODS AND MATERIALS

2.1 Statistical methods

In our study, we considered a GWA study with m SNPs. The null hypothesis was no association between a SNP and the phenotype of interest, and the alternative hypothesis was that there is an association. Let m_0 denote the number of true null hypotheses and m_1 denote the number of alternative hypotheses, where $m_1 = m - m_0$. Therefore, in the GWA study, we had m_1 disease-associated SNPs and m_0 unassociated SNPs. We performed logistic regression analysis for each SNP and obtained a p-value for each SNP using Wald's test. Table 1 shows the outcomes of these m multiple tests at a specified significance level α.

From Table 1, we can see that, statistically, there are two groups of SNPs based on a given significance level: the SNPs whose estimated effects are declared significantly different than zero (r) and those declared not to be significantly different than zero $(m-r)$. The SNPs that were significant can be further classified into two groups: one group is from the true null hypotheses (r_0), in which the SNPs were selected as a consequence of type I error; the other group is from the true alternative hypotheses (r_1), in which the SNPs are true positives. Given m_0 true null hypotheses and a nominal significance level of α, statistically, $r_0 = \alpha \times m_0$ tests would show a false significant association (false positive, type I error) between the SNPs and the phenotype of interest, and $m_0 - r_0$ tests would show no significant association between the SNPs and the phenotype (true negative). Among all the m_1 true alternative hypotheses, $m_1 - r_1$ tests would show no association between the SNPs and the phenotype (false negative, type II error), and $r_1 = r - r_0 = r - \alpha \times m_0$ tests would indicate a significant association (true positive).

In this paper, we proposed an approach to select top SNPs in the first stage of a GWA study using a strategy based on FNR. FNR, as defined in this paper, is a measure of the type II error rate for multiple tests, or the expected proportion of falsely not-rejected hypotheses among all alternative hypotheses [20, 21]. In our study, FNR is defined as the proportion of SNPs that are associated with the disease of interest but not selected in the first stage of the GWA

study. By using a strategy based on FNR, we can obtain the number of SNPs that should be selected given a pre-specified overall study power based on performing m tests.

To begin with, assume that m_0 is known (we will describe how to estimate this value later). At a given significance level α, we can evaluate the power of m independent multiple tests by using the formula:

$$
\begin{aligned}
power = 1 - \beta &= 1 - \frac{m_1 - r_1}{m_1} \\
&= 1 - \frac{m - m_0 - (r - \alpha \times m_0)}{m - m_0} \\
&= \frac{r - \alpha \times m_0}{m - m_0},
\end{aligned}
$$

where m is the total number of SNPs, m_0 is the number of true unassociated SNPs, r is the number of SNPs rejected at a significance level α, m_1 is the number of true disease-associated SNPs, and r_1 is the number of true disease-associated SNPs found significant in the analysis. Therefore, β is the type II error rate and $1 - \beta$ is the study power.

Using the above formula, if the power of multiple testing is specified, the number of SNPs to be selected from among all true disease-associated SNPs to achieve this power can be derived. Let $p_1, \ldots, p_m$ be the p-values of the m SNPs, and denote $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)} \leq \cdots \leq p_{(m)}$ as the ranked p-values, so $p_{(k)}$ is the kth smallest p-value. If the top k SNPs with the lowest p-values are selected at the significance level α, then there are k SNPs with p-values less than the significance level α. For our purpose, we therefore substitute $p_{(k)}$ for α and k for r in the above formula, and then the power of multiple tests is given by:

$$
\begin{aligned}
power = 1 - \beta &= 1 - \frac{m_1 - r_1}{m_1} \\
&= 1 - \frac{m - m_0 - (k - p_{(k)} \times m_0)}{m - m_0} \\
&= \frac{r - p_{(k)} \times m_0}{m - m_0},
\end{aligned}
$$

where $p_{(k)}$ is the kth-order of p-value. From this formula, m is known and m_0 can be estimated (see below). The power is a function of k, and to calculate the number k, we started with $k = 1$ and stopped when for the first time, the power value in above formula was greater than or equal to the pre-specified power. Therefore, if the power is pre-specified, the ranking value k, which is the number of SNPs selected, can be obtained based on the observed p-values.

As we discussed previously, to compute the number of SNPs selected, or k, we need to estimate m_0, denoted by $\hat{m}_0$, the number of true null hypotheses [3, 26, 31]. In our study, we employed three different methods to estimate the value of m_0. The first method we used was the adaptive linear step-up procedure [3, 31, 32]. For this approach, $\hat{m}_0$ was computed as $(m - r(\lambda))/(1 - \lambda)$, where λ was a tuning parameter

for the p-values between the null and alternative hypotheses and $r(\lambda)$ was the number of p-values less than or equal to the parameter λ. It has been shown that, the estimator $\hat{m}_0$ is unbiased if all p-values were from null hypotheses (i.e. from uniform $(0, 1)$ distribution) [5, 32]. However, when both null and alternative p-values are included, the estimate of $\hat{m}_0$ tends to be overestimated. When λ approaches 0, the bias of the estimate gets larger and the estimate is too conservative; while when λ approaches 1, the bias gets smaller but the variance of this estimate gets larger. Therefore, selecting an appropriate λ is significant in efficiently estimating $\hat{m}_0$. In general, a bootstrapping procedure is suggested to obtain the optimal λ. However, in Storey's paper [31], he also suggested that when the proportion of true alternative hypotheses is very small among all hypotheses, the best λ should be close to 0. This is indeed the case in the GWA studies, where the proportion of the disease-associated SNPs is likely to be small. Therefore, in this paper, we used different values of λ, such as 10^{-5}, 5×10^{-5}, 10^{-4}, and 10^{-3}. The second method is referred to as the two-stage linear step-up procedure [3]. For this method, we used a modified significance level $\alpha' = \alpha/(1+\alpha)$, where α is the nominal statistical significance level. We then evaluated the number of hypotheses rejected by $r = max\{k : p_{(k)} \leq \alpha' \times \frac{k}{m}, k = 1, \ldots, m\}$, and $\hat{m}_0$ was estimated as $m - r$. The last approach is the adaptive Benjamini-Hochberg procedure [3]. In this procedure, $m_0(k) = (m+1-k)/(1-p_{(k)})$ is defined as a function of ranked p-values and corresponding ranks. The procedure started with $k = 2$ and stopped at the smallest k value for which $m_0(k) > m_0(k-1)$. The estimated $\hat{m}_0$ is the minimum of $(m_0(k), m)$, rounding up to the next highest integer. In the simulation study, we found that the two-stage and the adaptive Banjamini-Hochberg approaches were relatively low powered and the adaptive linear step-up procedure was more robust than the other two methods; therefore, we only report results using this procedure in the paper.

2.2 Simulation study

In order to investigate the performance of the FNR-based approach for selection of top SNPs in the first stage of a GWA study, we applied forward-time simulation software (genomeSIM) to simulate large-scale genomic data in a population [8]. The specific parameters used for simulation are detailed in Table S1 in the Supplementary Material (http://www.intlpress.com/SII/p/2011/4-2/SII-4-2-bondy-supplement.zip). We simulated 100 replicates, each with 4,000 individuals and 500,000 SNPs for each individual.

We used a logistic regression model to simulate the case-control status. We studied five different models using different numbers of causal SNPs (5, 10, 15, 20, and 25 causal SNPs). For simplicity, we also assumed that the causal SNPs were independent (i.e., no linkage disequilibrium among causal SNPs). Within each model, we defined a range of odds ratios (ORs) for the causal SNPs. The ORs used for

Table 2. Parameters for the five simulation models. Different numbers of causal SNPs with a range of OR values were employed for different models. The corresponding heritability for each model is given as the expected squared residual of the observed and predicted disease status

| Model | # | Odds ratios | | | | | | | Herit-ability |
		1.2	1.3	1.4	1.5	1.6	1.7	1.8	
1	5	1	1	1	1	1	0	0	6.7%
2	10	2	2	2	1	1	1	1	14.2%
3	15	3	2	2	2	2	2	2	20.5%
4	20	3	3	3	3	3	3	2	25.2%
5	25	4	4	4	4	3	3	3	28.2%

#: number of causal SNPs

simulating case-control status are listed in Table 2. For example, for Model 4, we assumed that there were 20 causal SNPs associated with the disease: 3 SNPs with OR = 1.2, 3 SNPs with OR = 1.3, 3 SNPs with OR = 1.4, 3 SNPs with OR = 1.5, 3 SNPs with OR = 1.6, 3 SNPs with OR = 1.7, and 2 SNPs with OR = 1.8. We further denote $Y_j = \{0, 1\}$, $j = 1, \ldots, M$, as the outcome variables of case-control status of M individuals in the study, with 0 representing the individuals in the control group and 1 representing the individuals in the case group. So the logistic regression model is defined below:

$$\text{Logit}(p(Y_j = 1)) = \beta_0 + \sum_{i=1}^{n} \beta_i X_{ij},$$

where $X_{ij}(i = 1, \ldots, n, j = 1, \ldots, M)$ represent the categorical random variable for each individual with respect to the value of the three genotypes $\{0, 1, 2\}$ for n causal SNPs. $\beta_i(i = 1, \ldots, n)$ were the logistic regression coefficients, which are equal to $\text{Log}(\text{ORs})$. For each model, using specific intercept coefficient β_0, we randomly selected 1,000 cases and 1,000 controls from the 4,000 individuals for each replicate. In our study, we coded the genotypes as an additive model. The FNR approach is, however, not limited to the additive model and can readily be applied to the dominant or recessive models using appropriate genotype coding. The heritability associated with the simulated models was calculated using the expected squared residual between the observed and the predicted disease status [16–18]. The values of heritability are listed in Table 2. For example, for Model 2, the 10 causal SNPs explained 14.2% of the residual variance. In this study, we performed the statistical analyses using R (v 2.8) and Matlab (v R2007a).

3. RESULTS

Table 3 shows the numbers of SNPs that should be selected by the five models using the FNR approach given different pre-specified multi-testing powers. The results were

also based on the following parameters: $\lambda = 10^{-5}, 5 \times 10^{-5}, 10^{-4}$, and 10^{-3}. We reported the median numbers of SNPs selected, as well as the 1st and the 3rd quartile (Q_1 and Q_3) numbers of SNPs selected, based on 100 replicates. Each replicate included 1,000 cases and 1,000 controls. The results of the FNR approach are also reported according to the pre-specified multi-testing powers of 50%, 60%, 70%, 80%, and 90%.

Table 3 shows that to achieve higher power, more SNPs need to be selected as expected. For example, in Model 4, for $\lambda = 10^{-4}$, to achieve the pre-specified multi-testing power of 50%, the median number of SNPs selected was 48, and the Q_1 and Q_3 numbers were 38 and 54, respectively; to achieve the pre-specified power of 60%, the median number of SNPs selected was 62; to achieve the pre-specified power of 70%, the median number of SNPs selected was 75; to achieve the pre-specified power of 80%, the median number of SNPs selected was 92; and to achieve the pre-specified power of 90%, the median number of SNPs selected was 111. Similar trends were observed for other models.

In most cases, when the number of causal SNPs increased, the number of SNPs selected also increased, given the same parameter value (Table 3). For example, given $\lambda = 10^{-4}$, to achieve the pre-specified multi-testing power of 80% in Model 1, when there were 5 causal SNPs, we needed to select 28 top SNPs using the FNR approach. When the numbers of causal SNPs were 10, 15, and 20, we needed to select the top 46, 70, and 92 SNPs, respectively, to achieve the same power. However, using the same parameter in Model 5, where the number of causal SNPs was 25, to achieve 80% pre-specified power, we only needed to select 81 SNPs. This result could be due to variations in the ORs that were used in different simulation scenarios. Furthermore, as shown in Table 3, we also found that the number of SNPs selected increased as the parameter λ value increased, which is expected because the number of alternative hypotheses increases in proportion to the value of λ. For example, in Model 4, to achieve the pre-specified power of 80%, we needed to select the top 33 SNPs using our FNR approach when $\lambda = 10^{-5}$. When the parameter λ values were 5×10^{-5}, 10^{-4}, and 10^{-3}, we needed to select the top 63, 92, and 456 SNPs, respectively, to achieve the same power based on the median of 100 replicates. The large difference between the number of top SNPs selected using $\lambda = 10^{-4}$ and $\lambda = 10^{-3}$ could be due to the non-uniform property of the distribution of p-values.

We also evaluated the number of top SNPs selected using the traditional approaches for comparison: Bonferroni correction, BH-FDR approach, and fixed p-value cutoff (see Table 4). As expected, the approaches based on controlling type I errors are very conservative. For example, in Model 1, there were 5 causal SNPs in the model, but only 3 top SNPs were selected using Bonferroni correction, and 4 top SNPs were selected using BH-FDR approach, at a genome-wide 5% level of significance. More interestingly, we observed that the numbers of top SNPs selected using the type I

Table 3. The median $(Q_1 - Q_3)^*$ numbers of SNPs selected in the first stage of a two-stage GWA study using the FNR-based approach for the five simulation models, given pre-specified multi-testing powers (50%, 60%, 70%, 80% and 90%) and different parameter λ values (10^{-5}, 5×10^{-5}, 10^{-4}, and 10^{-3}), based on 100 replicates

Model	Parameter	Pre-specified multi-testing powers				
(#)	values	50%	60%	70%	80%	90%
	$\lambda = 10^{-5}$	$3(2-5)$	$4(2-6)$	$5(3-7)$	$6(4-9)$	$8(4-11)$
1	$\lambda = 5 \times 10^{-5}$	$5(1-10)$	$7(1-14)$	$9(2-19)$	$12(2-25)$	$15(2-30)$
(5)	$\lambda = 10^{-4}$	$10(2-20)$	$16(3-30)$	$21(3-39)$	$28(4-46)$	$36(4-54)$
	$\lambda = 10^{-3}$	$130(1-254)$	$180(1-314)$	$223(1-379)$	$284(1-418)$	$335(1-464)$
	$\lambda = 10^{-5}$	$7(4-8)$	$8(5-11)$	$10(7-13)$	$12(8-16)$	$14(9-18)$
2	$\lambda = 5 \times 10^{-5}$	$11(8-17)$	$16(11-22)$	$22(15-28)$	$27(18-36)$	$35(21-43)$
(10)	$\lambda = 10^{-4}$	$19(11-26)$	$25(15-35)$	$34(19-43)$	$46(26-55)$	$56(38-66)$
	$\lambda = 10^{-3}$	$172(85-246)$	$246(138-317)$	$314(187-378)$	$395(228-465)$	$476(273-525)$
	$\lambda = 10^{-5}$	$12(10-15)$	$15(13-19)$	$18(15-22)$	$21(18-26)$	$25(21-30)$
3	$\lambda = 5 \times 10^{-5}$	$24(18-29)$	$31(24-37)$	$39(31-45)$	$46(38-55)$	$56(46-64)$
(15)	$\lambda = 10^{-4}$	$35(27-41)$	$45(36-53)$	$58(47-68)$	$70(61-83)$	$86(74-100)$
	$\lambda = 10^{-3}$	$194(147-229)$	$272(204-314)$	$349(280-412)$	$430(372-491)$	$535(461-581)$
	$\lambda = 10^{-5}$	$19(16-23)$	$24(20-28)$	$28(24-33)$	$33(28-39)$	$38(33-44)$
4	$\lambda = 5 \times 10^{-5}$	$34(30-39)$	$43(36-50)$	$52(45-60)$	$63(54-71)$	$73(64-84)$
(20)	$\lambda = 10^{-4}$	$48(38-54)$	$62(50-68)$	$75(63-86)$	$92(78-103)$	$111(95-124)$
	$\lambda = 10^{-3}$	$193(159-228)$	$272(215-312)$	$355(304-419)$	$456(392-523)$	$573(506-628)$
	$\lambda = 10^{-5}$	$15(11-17)$	$18(14-22)$	$21(17-25)$	$25(20-29)$	$29(23-34)$
5	$\lambda = 5 \times 10^{-5}$	$28(22-32)$	$36(29-40)$	$44(37-49)$	$54(44-60)$	$63(54-71)$
(25)	$\lambda = 10^{-4}$	$40(32-47)$	$52(44-61)$	$65(54-76)$	$81(65-91)$	$94(83-108)$
	$\lambda = 10^{-3}$	$212(170-257)$	$288(228-339)$	$379(311-428)$	$461(399-528)$	$566(508-627)$

\#: number of causal SNPs
*Q_1: 1st quartile Q_3: 3rd quartile
λ: parameter used for estimating number of null hypotheses

Table 4. The median $(Q_1 - Q_3)^*$ numbers of SNPs selected in the first stage of a two-stage GWA study using the traditional approach, including Bonferroni correction and BH-FDR at 5% genome-wide significance level, and a fixed p-value cutoff 10^{-5}, for the five simulation models

Model (#)	Approaches	# of Selection
1	Bonferroni correction	$3(3-4)$
(5)	BH-FDR	$4(3-5)$
	Fixed $p = 10^{-5}$	$12(9-14)$
2	Bonferroni correction	$6(5-7)$
(10)	BH-FDR	$9(7-11)$
	Fixed $p = 10^{-5}$	$18(14-22)$
3	Bonferroni correction	$11(9-12)$
(15)	BH-FDR	$18(14-23)$
	Fixed $p = 10^{-5}$	$30(26-36)$
4	Bonferroni correction	$14(12-16)$
(20)	BH-FDR	$29(22-37)$
	Fixed $p = 10^{-5}$	$44(38-51)$
5	Bonferroni correction	$11(9-12)$
(25)	BH-FDR	$21(16-25)$
	Fixed $p = 10^{-5}$	$34(28-40)$

\#: number of causal SNPs
*Q_1: 1st quartile Q_3: 3rd quartile
p: p-value cutoff

error-based approaches are similar to those obtained using our FNR-based approach with a stringent parameter value $\lambda = 10^{-5}$: the results from Bonferroni correction are similar to those obtained from our approach with pre-specified multiple testing power of 50%, and the results from BH-FDR are similar to those from our approach with pre-specified multiple testing power of $60\% \sim 70\%$. For the traditional p-value cutoff of p-value $< 10^{-5}$, it is expected that in most cases, the number of top SNPs selected increases as the number of causal SNPs increases. For example, in Model 1, where there were 5 causal SNPs, 12 SNPs ($Q_1 = 9$ and $Q_3 = 14$) had p-values less than the specified threshold, based on the median of 100 replicates. The numbers of top SNPs selected were 18, 30, and 44 when the numbers of causal SNPs were 10, 15, and 20, respectively. In Model 5, when there were 25 causal SNPs, 34 top SNPs were selected, which could be due to variations in ORs in the different models. Many GWA studies select arbitrary numbers of top SNPs in stage one, so we also employed a fixed number of SNPs cutoff for each model, including selections of 10, 20, and 30 top SNPs, which are concordant with the commonly used cutoffs in current GWA studies. Obviously, for all the traditional approaches discussed here, for each model, there is only one number of top SNPs selected in this situation, since no false negative rate will be attached to these approaches.

Table 5. Comparison of the median numbers of causal SNPs selected for different pre-specified odds ratios for the five simulation models using the FNR-based approach (given pre-specified power 80%) with traditional approaches

Model (#)	Approaches	Parameter values	Odds ratios						
			1.8	1.7	1.6	1.5	1.4	1.3	1.2
1 (5)	Bonferroni correction		$0^a/0^b$	0/0	1/1	1/1	1/1	0/1	0/1
	BH-FDR		0/0	0/0	1/1	1/1	1/1	0/1	0/1
	Fixed p-value cutoff	$p = 10^{-5}$	0/0	0/0	1/1	1/1	1/1	0/1	0/1
		10	0/0	0/0	1/1	1/1	1/1	0/1	0/1
	Fixed number cutoff	20	0/0	0/0	1/1	1/1	1/1	0/1	0/1
		30	0/0	0/0	1/1	1/1	1/1	1/1	0/1
		$\lambda = 10^{-5}$	0/0	0/0	1/1	1/1	1/1	0/1	0/1
	FNR-based	$\lambda = 5 \times 10^{-5}$	0/0	0/0	1/1	1/1	1/1	0/1	0/1
		$\lambda = 10^{-4}$	0/0	0/0	1/1	1/1	1/1	0/1	0/1
		$\lambda = 10^{-3}$	0/0	0/0	1/1	1/1	1/1	1/1	0/1
2 (10)	Bonferroni correction		1/1	1/1	1/1	1/1	0/2	0/2	0/2
	BH-FDR		1/1	1/1	1/1	1/1	1/2	1/2	0/2
	Fixed p-value cutoff	$p = 10^{-5}$	1/1	1/1	1/1	1/1	1/2	1/2	0/2
		10	1/1	1/1	1/1	1/1	1/2	1/2	0/2
	Fixed number cutoff	20	1/1	1/1	1/1	1/1	1/2	1/2	0/2
		30	1/1	1/1	1/1	1/1	2/2	1/2	0/2
		$\lambda = 10^{-5}$	1/1	1/1	1/1	1/1	1/2	1/2	0/2
	FNR-based	$\lambda = 5 \times 10^{-5}$	1/1	1/1	1/1	1/1	1/2	1/2	0/2
		$\lambda = 10^{-4}$	1/1	1/1	1/1	1/1	2/2	1/2	0/2
		$\lambda = 10^{-3}$	1/1	1/1	1/1	1/1	2/2	2/2	0/2
3 (15)	Bonferroni correction		1/2	2/2	2/2	1/2	1/2	0/2	0/3
	BH-FDR		2/2	2/2	2/2	2/2	2/2	0/2	0/3
	Fixed p-value cutoff	$p = 10^{-5}$	2/2	2/2	2/2	2/2	2/2	0/2	0/3
		10	1/2	2/2	2/2	1/2	1/2	0/2	0/3
	Fixed number cutoff	20	2/2	2/2	2/2	2/2	2/2	0/2	0/3
		30	2/2	2/2	2/2	2/2	2/2	0/2	0/3
		$\lambda = 10^{-5}$	2/2	2/2	2/2	2/2	2/2	0/2	0/3
	FNR-based	$\lambda = 5 \times 10^{-5}$	2/2	2/2	2/2	2/2	2/2	0/2	0/3
		$\lambda = 10^{-4}$	2/2	2/2	2/2	2/2	2/2	1/2	0/3
		$\lambda = 10^{-3}$	2/2	2/2	2/2	2/2	2/2	1/2	0/3
4 (20)	Bonferroni correction		2/2	3/3	2/3	3/3	1/3	0/2	0/3
	BH-FDR		2/2	3/3	2/3	3/3	1/3	0/2	0/3
	Fixed p-value cutoff	$p = 10^{-5}$	2/2	3/3	2/3	3/3	2/3	0/2	0/3
		10	2/2	2/3	2/3	2/3	1/3	0/2	0/3
	Fixed number cutoff	20	2/2	3/3	2/3	3/3	1/3	0/2	0/3
		30	2/2	3/3	2/3	3/3	1/3	0/2	0/3
		$\lambda = 10^{-5}$	2/2	3/3	2/3	3/3	1/3	0/2	0/3
	FNR-based	$\lambda = 5 \times 10^{-5}$	2/2	3/3	2/3	3/3	2/3	0/2	0/3
		$\lambda = 10^{-4}$	2/2	3/3	2/3	3/3	2/3	0/2	0/3
		$\lambda = 10^{-3}$	2/2	3/3	3/3	3/3	2/3	2/2	0/3
5 (25)	Bonferroni correction		3/3	2/3	1/3	2/4	0/4	0/4	0/4
	BH-FDR		3/3	2/3	2/3	3/4	0/4	0/4	0/4
	Fixed p-value cutoff	$p = 10^{-5}$	3/3	3/3	2/3	3/4	4/4	1/4	0/4
		10	3/3	2/3	1/3	1/4	0/4	0/4	0/4
	Fixed number cutoff	20	3/3	2/3	2/3	3/4	1/4	0/4	0/4
		30	3/3	3/3	2/3	3/4	3/4	1/4	0/4
		$\lambda = 10^{-5}$	3/3	3/3	2/3	3/4	2/4	1/4	0/4
	FNR-based	$\lambda = 5 \times 10^{-5}$	3/3	3/3	3/3	3/4	4/4	1/4	0/4
		$\lambda = 10^{-4}$	3/3	3/3	3/3	3/4	4/4	1/4	0/4
		$\lambda = 10^{-3}$	3/3	3/3	3/3	4/4	4/4	2/4	0/4

#: number of causal SNPs
[a] Denotes how many causal SNPs were selected
[b] Denotes how many causal SNPs in total in the models
p: p-value cutoff
λ: parameter used for estimating number of null hypotheses

Table 6. *The median type I error probabilities and median observed powers of the FNR-based approach in stage one analysis for the five simulation models, given pre-specified multi-testing powers (50%, 60%, 70%, 80% and 90%) and different parameter λ values (10^{-5}, 5×10^{-5}, 10^{-4}, and 10^{-3}), based on 100 replicates*

| | | Pre-specified multi-testing powers | | | | | | | | | |
| | | 50% | | 60% | | 70% | | 80% | | 90% | |
Model (#)	Parameter values	Type I errors	Observed powers	Type I errors	Observed powers	Type I errors	Observed powers	Type I errors	Observed powers	Type I errors	Observed powers
1 (5)	$\lambda = 10^{-5}$	0	0.60	2.00E-6	0.60	4.00E-6	0.60	6.00E-6	0.60	1.00E-5	0.60
	$\lambda = 5 \times 10^{-5}$	6.00E-6	0.60	8.00E-6	0.60	1.20E-5	0.60	1.60E-5	0.60	2.20E-5	0.60
	$\lambda = 10^{-4}$	1.50E-5	0.60	2.40E-5	0.60	3.50E-5	0.60	4.90E-5	0.60	6.20E-5	0.60
	$\lambda = 10^{-3}$	2.51E-4	0.80	3.51E-4	0.80	4.37E-4	0.80	5.61E-4	0.80	6.60E-4	0.80
2 (10)	$\lambda = 10^{-5}$	4.00E-6	0.50	6.00E-6	0.50	8.00E-6	0.50	1.20E-5	0.60	1.60E-5	0.60
	$\lambda = 5 \times 10^{-5}$	1.20E-5	0.50	1.90E-5	0.60	3.20E-5	0.60	4.20E-5	0.60	5.60E-5	0.60
	$\lambda = 10^{-4}$	2.50E-5	0.60	3.60E-5	0.60	5.50E-5	0.70	7.80E-5	0.70	9.80E-5	0.70
	$\lambda = 10^{-3}$	3.31E-4	0.70	4.76E-4	0.80	6.11E-4	0.80	7.75E-4	0.80	9.38E-4	0.80
3 (15)	$\lambda = 10^{-5}$	8.00E-6	0.53	1.20E-5	0.60	1.80E-5	0.60	2.40E-5	0.60	3.10E-5	0.60
	$\lambda = 5 \times 10^{-5}$	2.80E-5	0.60	4.20E-5	0.67	5.70E-5	0.67	7.10E-5	0.67	9.10E-5	0.67
	$\lambda = 10^{-4}$	5.10E-5	0.67	7.10E-5	0.67	9.50E-5	0.67	1.20E-4	0.67	1.49E-4	0.67
	$\lambda = 10^{-3}$	3.65E-4	0.73	5.21E-4	0.73	6.77E-4	0.73	8.40E-4	0.73	1.05E-3	0.73
4 (20)	$\lambda = 10^{-5}$	1.70E-5	0.55	2.40E-5	0.55	3.30E-5	0.55	4.20E-5	0.60	5.20E-5	0.60
	$\lambda = 5 \times 10^{-5}$	4.50E-5	0.60	6.20E-5	0.60	7.90E-5	0.65	1.00E-4	0.65	1.21E-4	0.65
	$\lambda = 10^{-4}$	7.20E-5	0.60	9.70E-5	0.65	1.25E-4	0.65	1.56E-4	0.65	1.95E-4	0.70
	$\lambda = 10^{-3}$	3.55E-4	0.70	5.12E-4	0.75	6.82E-4	0.75	8.82E-4	0.75	1.11E-3	0.75
5 (25)	$\lambda = 10^{-5}$	8.00E-6	0.40	1.30E-5	0.44	1.80E-5	0.48	2.40E-5	0.48	3.20E-5	0.52
	$\lambda = 5 \times 10^{-5}$	3.10E-5	0.52	4.40E-5	0.56	5.80E-5	0.56	7.50E-5	0.60	9.30E-5	0.60
	$\lambda = 10^{-4}$	5.40E-5	0.56	7.40E-5	0.56	9.90E-5	0.60	1.30E-4	0.60	1.59E-4	0.64
	$\lambda = 10^{-3}$	3.91E-4	0.68	5.42E-4	0.68	7.22E-4	0.72	8.86E-4	0.72	1.10E-3	0.72

#: number of causal SNPs
λ: parameter used for estimating number of null hypotheses

We compared the number of causal SNPs with different ORs selected using the FNR-based method, given the different λ values and a pre-specified power of 80%, to the number of causal SNPs selected using the traditional approaches discussed above (Table 5). In the simulated data, the numbers and locations of the causal SNPs were known for the different simulation models. Therefore, given a number of how many top SNPs were selected, the exact number of causal SNPs selected and the corresponding ORs were known. As expected, when the ORs were high (i.e., OR = 1.7 or 1.8), all approaches showed very similar results and the corresponding causal SNPs were selected with a higher probability. All the causal SNPs associated with OR = 1.8 were selected for all 5 models by using different approaches, except for the approaches using the Bonferroni correction and the fixed top 10 SNPs in Model 3, which only selected 1 of the 2 causal SNPs associated with OR = 1.8. We observed a similar trend for the causal SNPs associated with OR = 1.7. When the ORs decrease, it is not surprising that all approaches would select the corresponding causal SNPs with a lower probability. When the OR was small (i.e., OR = 1.2), we found that given the sample size, none of the approaches could identify the corresponding causal SNPs. However, we also observed that when the OR was moderate, such as OR = 1.3 and 1.4, which are similar to the ORs reported in the current

GWA studies, our FNR-based approach could identify more causal SNPs than the traditional approaches. For example, in Model 2, there are 2 SNPs with OR = 1.3. Using the traditional approaches, we only can identify half of the causal SNPs (1/2), whereas using our FNR-based approach with $\lambda = 10^{-3}$, we can identify all the causal associated SNPs (2/2) based on the median of the 100 replicates. In Model 3 where there were 2 causal SNPs associated with OR = 1.3, none of the causal SNPs (0/2) was selected using the traditional approaches, but 1 of 2 causal SNPs was selected using the FNR-based approach with parameters $\lambda = 10^{-4}$ and 10^{-3}.

These findings provide strong support that compared to the traditional fixed p-value cutoffs, fixed number of top SNPs cutoffs and the type I error-based approaches, the FNR-based approach has more power to identify moderate significant causal SNPs when using a relatively liberal parameter λ. On the basis of the results from our simulation studies, we would like to recommend a parameter value of $\lambda = 10^{-3}$ for selecting top SNPs for stage one of GWA studies. In the simulation studies, because the disease-causal SNPs were pre-defined, we estimated the type I error probabilities and the observed powers for all approaches in the stage one analysis. Table 6 reports the median type I error probabilities and median observed powers using the pro-

Table 7. The median type I error probabilities and median observed powers of the traditional approaches in stage one analysis, including Bonferroni correction and BH-FDR at 5% genome-wide significance level, a fixed p-value cutoff 10^{-5}, and fixed number cutoffs, for the five simulation models

Model (#)	Approaches	Parameter values	Type I errors	Observed powers
	Bonferroni correction		0	0.60
1	BH-FDR		0	0.60
(5)	Fixed p-value cutoff	$p = 10^{-5}$	1.60E-5	0.60
	Fixed number cutoff	10	1.40E-5	0.60
		20	3.20E-5	0.80
		30	5.20E-5	0.80
	Bonferroni correction		2.00E-6	0.50
2	BH-FDR		6.00E-6	0.60
(10)	Fixed p-value cutoff	$p = 10^{-5}$	2.40E-5	0.60
	Fixed number cutoff	10	8.00E-6	0.60
		20	2.80E-5	0.60
		30	4.60E-5	0.70
	Bonferroni correction		6.00E-6	0.53
3	BH-FDR		1.80E-5	0.60
(15)	Fixed p-value cutoff	$p = 10^{-5}$	4.20E-5	0.67
	Fixed number cutoff	10	4.00E-6	0.53
		20	2.20E-5	0.60
		30	4.00E-5	0.67
	Bonferroni correction		8.00E-6	0.50
4	BH-FDR		3.60E-5	0.60
(20)	Fixed p-value cutoff	$p = 10^{-5}$	6.30E-5	0.60
	Fixed number cutoff	10	3.00E-6	0.43
		20	1.80E-5	0.55
		30	3.60E-5	0.60
	Bonferroni correction		4.00E-6	0.36
5	BH-FDR		1.80E-5	0.48
(25)	Fixed p-value cutoff	$p = 10^{-5}$	4.20E-5	0.54
	Fixed number cutoff	10	2.00E-6	0.36
		20	1.60E-5	0.48
		30	3.40E-5	0.52

#: number of causal SNPs

Table 8. The numbers of SNPs selected given different pre-specified powers in the glioma GWA dataset, at different parameter λ values

Parameter values	Pre-specified multi-testing powers				
	50%	60%	70%	80%	90%
$\lambda = 10^{-5}$	16	19	23	27	31
$\lambda = 5 \times 10^{-5}$	35	47	56	65	80
$\lambda = 10^{-4}$	49	61	77	89	113
$\lambda = 10^{-3}$	363	478	557	663	751

λ: parameter used for estimating number of null hypotheses

the FNR-based approach increased with the increase of pre-specified multi-testing power and the value of the parameter λ. Table 7 reports the median type I error probabilities and median observed powers for the standard approaches.

4. APPLICATION TO REAL GLIOMA DATA

In addition to the simulated data, we applied the proposed FNR-based approach to data from a GWA study of glioma we have recently conducted [28]. Glioma is a rare and diffusely infiltrating brain disease. To investigate the FNR-based approach proposed in this paper, we used SNP genotype data from this whole-genome association analysis. The GWA study was based on genotyping $1,247$ glioma patients and $2,232$ controls in the first stage for $499,139$ autosomal SNPs. Using these data, we applied the FNR-based approach to estimate the number of top SNPs to be selected given different pre-specified multi-testing powers in the first stage. These results are shown in Table 8. With the use of the FNR approach with $\lambda = 10^{-4}$, 49 SNPs were selected to achieve 50% power, and 61, 77, 89, and 113 SNPs were selected to achieve powers of 60%, 70%, 80%, and 90%, respectively. In the original GWA study of glioma, 34 SNPs were selected in the first stage using an arbitrary threshold p-value $< 10^{-5}$. Subsequent replication of these 34 SNPs was performed in three case-control series totaling $5,498$ individuals confirmed that 14 SNPs were significantly associated with glioma and identified 5 distinct genetic loci. Using the FNR approach to achieve 80% power, we would have selected 89 SNPs in the first stage, which could have identified the same 5 genetic loci. It needs to be noted that all the genetic regions we proposed to select in the first stage have been validated in the second stage of replication in this GWA study of glioma. The identification of more susceptive SNPs to be repeated in the second stage might allow us to identify more genetic loci. Actually, if the investigators could have selected more top SNPs as we suggested using the FNR-based approach with a liberal parameter value of λ (e.g. $\lambda = 10^{-4}$ or 10^{-3}), it would be possible to identify two additional loci associated with glioma which were recently discovered (unpublished data). Most importantly, compared to the traditional arbitrary cutoff approach, using the FNR-based approach, one can select top SNPs in stage one with pre-specified confidence.

posed FNR-based approach with different values of λ parameter. As expected, for the FNR-based approach, as the pre-specified multi-testing power and the value of the parameter λ increased, the type I error probabilities increased because more SNPs were selected. However, it is important to note that the GWA significance (5×10^{-8}) employed at the end of the experiment (stage two) will control the overall type I error probabilities. Also, the observed powers of

5. DISCUSSION

The purpose of this paper is to provide a criterion that the investigators can follow to select top SNPs in the first stage of a GWA study based on controlling the type II errors. The findings from both our analysis of simulated data and our GWA study of glioma indicate that the proposed approach provides an FNR-based criterion to select more potential disease-associated SNPs with moderate significance according to pre-specified powers. Using the FNR-based approach, the number of SNPs to be selected in the first stage on the basis of observed p-values and a pre-specified multi-testing power can be ascribed, thus controlling the false negative rate.

To illustrate the performance of the FNR approach, we conducted simulation studies of five different scenarios, with respect to different numbers of actual causal SNPs for a range of OR values. As expected, the simulation results showed that more SNPs need to be selected when the pre-specified multi-testing power increases, as well as when the number of actual causal SNPs increases. We compared the FNR-based approach using pre-specified powers to the traditional approaches for selecting SNPs, including fixed p-value cutoffs, fixed number of SNPs cutoffs, Bonferroni correction and BH-FDR approach. The approaches based on controlling type I errors (e.g. Bonferroni correction and BH-FDR) are conservative and select fewer top SNPs for replication in stage two. When the disease-associated SNPs are highly significant, both the traditional approaches and the FNR-based approach can identify them. But when the disease-associated SNPs are only moderately significant, the traditional approaches may lose power to identify them, whereas the FNR-based approach will have more power to identify this kind of disease-associated SNP. In the simulations, we assumed that all the causal SNPs are in linkage equilibrium, therefore, single SNP analysis is valid and provides an unbiased estimate of marker effect size. We performed a proof of principle simulation study to investigate the impact of linkage disequilibrium (LD) among causal SNPs. We simulated two scenarios. In one scenario the two disease-causal SNPs were in LD ($r^2 = 0.4$) and in the other scenario they were not in LD ($r^2 = 0$). We found that the p-values were less significant when causal SNPs were in LD compared to when causal SNPs were not in LD. Therefore, none of the approaches, including the standard approaches and the proposed FNR-based approach, selected the two causal SNPs in LD in the first stage (data not shown). The issue related with multiple causal SNPs in varying linkage disequilibrium should be further investigated. One of the limitations of our simulation study is that our simulation models with multiple disease causing loci had larger heritability (as shown in Table 2), therefore, signal to noise ratio or the complexity of the trait may have some impact on the results.

When using the FNR-based approach proposed in this paper, the selection of appropriate parameter λ in estimating $\hat{m}_0$ is very important. It has been shown that $\hat{m}_0$ can be overestimated when the parameter λ is very small [32]. An alternative approach for estimating $\hat{m}_0$ has been proposed [32] where $\hat{m}_0$ was estimated by smoothing the function $\hat{m}_0(\lambda)$ over a range of values of λ, based on natural cubic spline with 3 degrees of freedom. However, this approach might not be suitable for GWA studies because it estimates $\hat{m}_0$ at the limiting value of $\lambda = 1$. Thus, the estimated proportion of disease-associated SNPs among all the SNPs will be extremely high. However, this is not the case from the findings of the current GWA studies, where only a handful of SNPs were discovered. Moreover, based on p-values obtained from our simulated GWA data, we did not find a pattern of p-values suggested by Storey and Tibshirani, therefore, this approach is not directly applicable for GWA studies. Because the proportion of the unassociated SNPs (not associated with the disease of interest) among all the SNPs could be close to 1 and the p-values corresponding to the disease-associated SNPs are always assumed to be more significant than those corresponding to the unassociated SNPs, we used small values of λ as suggested in [31], such as 10^{-5}, 5×10^{-5}, 10^{-4} and 10^{-3} for our FNR-based approach. From our simulation results, we found that the average values of estimated $\hat{m}_0$ were not dramatically different for different values of λ. For example, when the number of causal SNPs is 15 (Model 3), the medians of estimated $\hat{m}_0$ based on 100 replicates were 499970, 499960, 499950 and 499945, respectively, for λ of 10^{-5}, 5×10^{-5}, 10^{-4} and 10^{-3}. Thus, the estimated $\hat{m}_0$ was not overestimated in our simulation studies. This phenomenon could be due to multiple SNPs in linkage disequilibrium with causal SNPs. Therefore, p-values associated with these SNPs will also be significant.

To select the parameter value of λ for GWA studies, we would like to recommend using a liberal $\lambda = 10^{-3}$. Although more top SNPs will be selected in the first stage using $\lambda = 10^{-3}$ (usually hundreds of top SNPs) than with the traditional approaches, it is still feasible for replications in GWA studies because of the rapid development of genotyping techniques and therefore the decrease of genotyping cost. Furthermore, because a pre-specific power-based criterion is attached to our FNR-based approach, this approach provides a more optimal selection criterion than traditional approaches. To identify disease-associated SNPs with smaller ORs (such as 1.2), much larger sample sizes in stage one of the two-stage design may be required. Our methodology of choosing the number of SNPs is, however, valid irrespective of the magnitude of the odds ratio. It should also be noted that the type I error can be controlled in the second-stage analysis of GWA studies, because the SNPs selected in stage one with the use of our FNR approach are not final, and they have to meet the GWA significance (5.0×10^{-8}) at the end of the stage two analysis [1]. Furthermore, to achieve higher power, much larger samples would be needed in order to detect the causal SNPs with small true ORs (i.e. OR = 1.2). Therefore, a GWA study with 1,000 cases and 1,000 controls

is likely to have low power to select 90% of the small effect causal SNPs.

We also applied the FNR approach to a recently published GWA dataset. Selecting a bit larger number of SNPs in stage one, our approach led to the same conclusion as the original GWA study using an arbitrary p-value cut-off. Furthermore, using arbitrary p-value or fixed number cut-offs, one just selects top SNPs, without confidence about the disease-associated variants being selected; on the other hand, using the proposed FNR-based approach, we have confidence that most of the moderately significant SNPs in the GWA study of glioma were selected in stage one for replication.

In conclusion, we present an FNR-based approach for selecting top SNPs given a pre-specified power based on the ranked p-values. This approach will select a relatively larger number of top SNPs in stage one that could include more moderately significant SNPs. The type II error for stage one can be controlled, and type I error can be controlled at the end of the stage two analysis in the GWA studies.

ACKNOWLEDGEMENTS

REFERENCES

[1] ALTSHULER, D., DALY, M. J., and LANDER, E. S. (2008). Genetic mapping in human disease. *Science* **322** 881–888.

[2] AMOS, C. I., WU, X., BRODERICK, P., GORLOV, I. P., GU, J., EISEN, T., DONG, Q., ZHANG, Q., GU, X., VIJAYAKRISHNAN, J., SULLIVAN, K., MATAKIDOU, A., WANG, Y., MILLS, G., DOHENY, K., TSAI, Y. Y., CHEN, W. V., SHETE, S., SPITZ, M. R., and HOULSTON, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* **40** 616–622.

[3] BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery fate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25** 60–83.

[4] BISHOP, D. T., DEMENAIS, F., ILES, M. M., HARLAND, M., TAYLOR, J. C., CORDA, E., RANDERSON-MOOR, J., AITKEN, J. F., AVRIL, M. F., AZIZI, E., BAKKER, B., BIANCHI-SCARRA, G., BRESSAC-DE, P. B., CALISTA, D., CANNON-ALBRIGHT, L. A., CHIN, A. W., DEBNIAK, T., GALORE-HASKEL, G., GHIORZO, P., GUT, I., HANSSON, J., HOCEVAR, M., HOIOM, V., HOPPER, J. L., INGVAR, C., KANETSKY, P. A., KEFFORD, R. F., LANDI, M. T., LANG, J., LUBINSKI, J., MACKIE, R., MALVEHY, J., MANN, G. J., MARTIN, N. G., MONTGOMERY, G. W., VAN NIEUWPOORT, F. A., NOVAKOVIC, S., OLSSON, H., PUIG, S., WEISS, M., VAN WORKUM, W., ZELENIKA, D., BROWN, K. M., GOLDSTEIN, A. M., GILLANDERS, E. M., BOLAND, A., GALAN, P., ELDER, D. E., GRUIS, N. A., HAYWARD, N. K., LATHROP, G. M., BARRETT, J. H., and BISHOP, J. A. (2009). Genome-wide association study identifies three loci associated with melanoma risk. *Nature Genetics* **41** 920–925.

[5] BLACK, M. A. (2004). A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **66** 297–304. MR2062377

[6] BURTON, P. R., CLAYTON, D. G., CARDON, L. R., CRADDOCK, N., DELOUKAS, P., DUNCANSON, A., KWIATKOWSKI, D. P., MCCARTHY, M. I., OUWEHAND, W. H., SAMANI, N. J., TODD, J. A., DONNELLY, P., BARRETT, J. C., DAVISON, D., EASTON, D., EVANS, D. M., LEUNG, H. T., MARCHINI, J. L., MORRIS, A. P., SPENCER, C. C., TOBIN, M. D., ATTWOOD, A. P., BOORMAN, J. P., CANT, B., EVERSON, U., HUSSEY, J. M., JOLLEY, J. D., KNIGHT, A. S., KOCH, K., MEECH, E., NUTLAND, S., PROWSE, C. V., STEVENS, H. E., TAYLOR, N. C., WALTERS, G. R., WALKER, N. M., WATKINS, N. A., WINZER, T., JONES, R. W., MCARDLE, W. L., RING, S. M., STRACHAN, D. P., PEMBREY, M., BREEN, G., ST, C. D., CAESAR, S., GORDON-SMITH, K., JONES, L., FRASER, C., GREEN, E. K., GROZEVA, D., HAMSHERE, M. L., HOLMANS, P. A., JONES, I. R., KIROV, G., MOSKIVINA, V., NIKOLOV, I., O'DONOVAN, M. C., OWEN, M. J., COLLIER, D. A., ELKIN, A., FARMER, A., WILLIAMSON, R., MCGUFFIN, P., YOUNG, A. H., FERRIER, I. N., BALL, S. G., BALMFORTH, A. J., BARRETT, J. H., BISHOP, T. D., ILES, M. M., MAQBOOL, A., YULDASHEVA, N., HALL, A. S., BRAUND, P. S., DIXON, R. J., MANGINO, M., STEVENS, S., THOMPSON, J. R., BREDIN, F., TREMELLING, M., PARKES, M., DRUMMOND, H., LEES, C. W., NIMMO, E. R., SATSANGI, J., FISHER, S. A., FORBES, A., LEWIS, C. M., ONNIE, C. M., PRESCOTT, N. J., SANDERSON, J., MATTHEW, C. G., BARBOUR, J., MOHIUDDIN, M. K., TODHUNTER, C. E., MANSFIELD, J. C., AHMAD, T., CUMMINGS, F. R., JEWELL, D. P., WEBSTER, J., BROWN, M. J., LATHROP, M. G., CONNELL, J., DOMINICZAK, A., MARCANO, C. A., BURKE, B., DOBSON, R., GUNGADOO, J., LEE, K. L., MUNROE, P. B., NEWHOUSE, S. J., ONIPINLA, A., WALLACE, C., XUE, M., CAULFIELD, M., FARRALL, M., BARTON, A., BRUCE, I. N., DONOVAN, H., EYRE, S., GILBERT, P. D., HILDER, S. L., HINKS, A. M., JOHN, S. L., POTTER, C., SILMAN, A. J., SYMMONS, D. P., THOMSON, W., WORTHINGTON, J., DUNGER, D. B., WIDMER, B., FRAYLING, T. M., FREATHY, R. M., LANGO, H., PERRY, J. R., SHIELDS, B. M., WEEDON, M. N., HATTERSLEY, A. T., HITMAN, G. A., WALKER, M., ELLIOTT, K. S., GROVES, C. J., LINDGREN, C. M., RAYNER, N. W., TIMPSON, N. J., ZEGGINI, E., NEWPORT, M., SIRUGO, G., LYONS, E., VANNBERG, F., HILL, A. V., BRADBURY, L. A., FARRAR, C., POINTON, J. J., WORDSWORTH, P., BROWN, M. A., FRANKLYN, J. A., HEWARD, J. M., SIMMONDS, M. J., GOUGH, S. C., SEAL, S., STRATTON, M. R., RAHMAN, N., BAN, M., GORIS, A., SAWCER, S. J., COMPSTON, A., CONWAY, D., JALLOW, M., NEWPORT, M., SIRUGO, G., ROCKETT, K. A., BUMPSTEAD, S. J., CHANEY, A., DOWNES, K., GHORI, M. J., GWILLIAM, R., HUNT, S. E., INOUYE, M., KENIRY, A., KING, E., MCGINNIS, R., POTTER, S., RAVINDRARAJAH, R., WHITTAKER, P., WIDDEN, C., WITHERS, D., CARDIN, N. J., DAVISON, D., FERREIRA, T., PEREIRA-GALE, J., HALLGRIMSDÓTTIR, I. B., HOWIE, B. N., SU, Z., TEO, Y. Y., VUKCEVIC, D., BENTLEY, D., BROWN, M. A., COMPSTON, A., FARRALL, M., HALL, A. S., HATTERSLEY, A. T., HILL, A. V., PARKES, M., PEMBREY, M., STRATTON, M. R., MITCHELL, S. L., NEWBY, P. R., BRAND, O. J., CARR-SMITH, J., PEARCE, S. H., MCGINNIS, R., KENIRY, A., DELOUKAS, P., REVEILLE, J. D., ZHOU, X., SIMS, A. M., DOWLING, A., TAYLOR, J., DOAN, T., DAVIS, J. C., SAVAGE, L., WARD, M. M., LEARCH, T. L., WEISMAN, M. H., and BROWN, M. (2007). Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics* **39** 1329–1337.

[7] 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. MR2089454

[8] DUDEK, S. M., MOTSINGER, A. A., VELEZ, D. R., WILLIAMS, S. M., and RITCHIE, M. D. (2006). Data simulation software for whole-genome association and other studies in human genetics. *Pacific Symposium on Biocomputing* **11** 499–510.

[9] EASTON, D. F. and EELES, R. A. (2008). Genome-wide association studies in cancer. *Human Molecular Genetics* **17** R109–R115.

[10] EELES, R. A., KOTE-JARAI, Z., GILES, G. G., OLAMA, A. A., GUY, M., JUGURNAUTH, S. K., MULHOLLAND, S., LEONGAMORNLERT, D. A., EDWARDS, S. M., MORRISON, J., FIELD, H. I., SOUTHEY, M. C., SEVERI, G., DONOVAN, J. L., HAMDY, F. C., DEARNALEY, D. P., MUIR, K. R., SMITH, C., BAGNATO, M., RDERN-JONES, A. T., HALL, A. L., O'BRIEN, L. T., GEHR-SWAIN, B. N., WILKINSON, R. A., COX, A., LEWIS, S., BROWN, P. M., JHAVAR, S. G., TYMRAKIEWICZ, M., LOPHATANANON, A., BRYANT, S. L., HORWICH, A., HUDDART, R. A., KHOO, V. S., PARKER, C. C., WOODHOUSE, C. J., THOMPSON, A., CHRISTMAS, T., OGDEN, C., FISHER, C., JAMIESON, C., COOPER, C. S., ENGLISH, D. R., HOPPER, J. L., NEAL, D. E., and EASTON, D. F. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics* **40** 316–321.

[11] GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64** 499–517. MR1924303

[12] HELGADOTTIR, A., THORLEIFSSON, G., MAGNUSSON, K. P., GRETARSDOTTIR, S., STEINTHORSDOTTIR, V., MANOLESCU, A., JONES, G. T., RINKEL, G. J., BLANKENSTEIJN, J. D., RONKAINEN, A., JAASKELAINEN, J. E., KYO, Y., LENK, G. M., SAKALIHASAN, N., KOSTULAS, K., GOTTSATER, A., FLEX, A., STEFANSSON, H., HANSEN, T., ANDERSEN, G., WEINSHEIMER, S., BORCH-JOHNSEN, K., JORGENSEN, T., SHAH, S. H., QUYYUMI, A. A., GRANGER, C. B., REILLY, M. P., AUSTIN, H., LEVEY, A. I., VACCARINO, V., PALSDOTTIR, E., WALTERS, G. B., JONSDOTTIR, T., SNORRADOTTIR, S., MAGNUSDOTTIR, D., GUDMUNDSSON, G., FERRELL, R. E., SVEINBJORNSDOTTIR, S., HERNESNIEMI, J., NIEMELA, M., LIMET, R., ANDERSEN, K., SIGURDSSON, G., BENEDIKTSSON, R., VERHOEVEN, E. L., TEIJINK, J. A., GROBBEE, D. E., RADER, D. J., COLLIER, D. A., PEDERSEN, O., POLA, R., HILLERT, J., LINDBLAD, B., VALDIMARSSON, E. M., MAGNADOTTIR, H. B., WIJMENGA, C., TROMP, G., BAAS, A. F., RUIGROK, Y. M., VAN RIJ, A. M., KUIVANIEMI, H., POWELL, J. T., MATTHIASSON, S. E., GULCHER, J. R., THORGEIRSSON, G., KONG, A., THORSTEINSDOTTIR, U., and STEFANSSON, K. (2008). The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genetics* **40** 217–224.

[13] HIRSCHHORN, J. N. and DALY, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6** 95–108.

[14] HOULSTON, R. S., WEBB, E., BRODERICK, P., PITTMAN, A. M., DI BERNARDO, M. C., LUBBE, S., CHANDLER, I., VIJAYAKRISHNAN, J., SULLIVAN, K., PENEGAR, S., CARVAJAL-CARMONA, L., HOWARTH, K., JAEGER, E., SPAIN, S. L., WALTHER, A., BARCLAY, E., MARTIN, L., GORMAN, M., DOMINGO, E., TEIXEIRA, A. S., KERR, D., CAZIER, J. B., NIITTYMAKI, I., TUUPANEN, S., KARHU, A., AALTONEN, L. A., TOMLINSON, I. P., FARRINGTON, S. M., TENESA, A., PRENDERGAST, J. G., BARNETSON, R. A., CETNARSKYJ, R., PORTEOUS, M. E., PHAROAH, P. D., KOESSLER, T., HAMPE, J., BUCH, S., SCHAFMAYER, C., TEPEL, J., SCHREIBER, S., VOLZKE, H., CHANG-CLAUDE, J., HOFFMEISTER, M., BRENNER, H., ZANKE, B. W., MONTPETIT, A., HUDSON, T. J., GALLINGER, S., CAMPBELL, H., and DUNLOP, M. G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics* **40** 1426–1435.

[15] HUNG, R. J., MCKAY, J. D., GABORIEAU, V., BOFFETTA, P., HASHIBE, M., ZARIDZE, D., MUKERIA, A., SZESZENIA-DABROWSKA, N., LISSOWSKA, J., RUDNAI, P., FABIANOVA, E., MATES, D., BENCKO, V., FORETOVA, L., JANOUT, V., CHEN, C., GOODMAN, G., FIELD, J. K., LILOGLOU, T., XINARIANOS, G., CASSIDY, A., MCLAUGHLIN, J., LIU, G., NAROD, S., KROKAN, H. E., SKORPEN, F., ELVESTAD, M. B., HVEEM, K., VATTEN, L., LINSEISEN, J., CLAVEL-CHAPELON, F., VINEIS, P., BUENO-DE-MESQUITA, H. B., LUND, E., MARTINEZ, C., BINGHAM, S., RASMUSON, T., HAINAUT, P., RIBOLI, E., AHRENS, W., BENHAMOU, S., LAGIOU, P., TRICHOPOULOS, D., HOLCATOVA, I., MERLETTI, F., KJAERHEIM, K., AGUDO, A., MACFARLANE, G., TALAMINI, R., SIMONATO, L., LOWRY, R., CONWAY, D. I., ZNAOR, A., HEALY, C., ZELENIKA, D., BOLAND, A., DELEPINE, M., FOGLIO, M., LECHNER, D., MATSUDA, F., BLANCHE, H., GUT, I., HEATH, S., LATHROP, M., and BRENNAN, P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452** 633–637.

[16] KATHIRESAN, S., MELANDER, O., GUIDUCCI, C., SURTI, A., BURTT, N. P., RIEDER, M. J., COOPER, G. M., ROOS, C., VOIGHT, B. F., HAVULINNA, A. S., WAHLSTRAND, B., HEDNER, T., CORELLA, D., TAI, E. S., ORDOVAS, J. M., BERGLUND, G., VARTIAINEN, E., JOUSILAHTI, P., HEDBLAD, B., TASKINEN, M. R., NEWTON-CHEH, C., SALOMAA, V., PELTONEN, L., GROOP, L., ALTSHULER, D. M., and ORHO-MELANDER, M. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics* **40** 189–197.

[17] LEE, S. H., VAN DER WERF, J. H., HAYES, B. J., GODDARD, M. E., and VISSCHER, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genetics* **4** e1000231.

[18] MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A., and VISSCHER, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

[19] MCPHERSON, R., PERTSEMLIDIS, A., KAVASLAR, N., STEWART, A., ROBERTS, R., COX, D. R., HINDS, D. A., PENNACCHIO, L. A., TYBJAERG-HANSEN, A., FOLSOM, A. R., BOERWINKLE, E., HOBBS, H. H., and COHEN, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science* **316** 1488–1491.

[20] NORRIS, A. W. and KAHN, C. R. (2006). Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. *Proceedings of the National Academy of Sciences of the United States of America* **103** 649–653.

[21] PAWITAN, Y., MICHIELS, S., KOSCIELNY, S., GUSNANTO, A., and PLONER, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21** 3017–3024.

[22] PEARSON, T. A. and MANOLIO, T. A. (2008). How to interpret a genome-wide association study. *JAMA* **299** 1335–1344.

[23] SAMANI, N. J., ERDMANN, J., HALL, A. S., HENGSTENBERG, C., MANGINO, M., MAYER, B., DIXON, R. J., MEITINGER, T., BRAUND, P., WICHMANN, H. E., BARRETT, J. H., KONIG, I. R., STEVENS, S. E., SZYMCZAK, S., TREGOUET, D. A., ILES, M. M., PAHLKE, F., POLLARD, H., LIEB, W., CAMBIEN, F., FISCHER, M., OUWEHAND, W., BLANKENBERG, S., BALMFORTH, A. J., BAESSLER, A., BALL, S. G., STROM, T. M., BRAENNE, I., GIEGER, C., DELOUKAS, P., TOBIN, M. D., ZIEGLER, A., THOMPSON, J. R., and SCHUNKERT, H. (2007). Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine* **357** 443–453.

[24] SARKAR, S. K. (2004). FDR-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference* **125** 119–137. MR2086892

[25] SAXENA, R., VOIGHT, B. F., LYSSENKO, V., BURTT, N. P., DE BAKKER, P. I., CHEN, H., ROIX, J. J., KATHIRESAN, S., HIRSCHHORN, J. N., DALY, M. J., HUGHES, T. E., GROOP, L., ALTSHULER, D., ALMGREN, P., FLOREZ, J. C., MEYER, J., ARDLIE, K., BENGTSSON, B. K., ISOMAA, B., LETTRE, G., LINDBLAD, U., LYON, H. N., MELANDER, O., NEWTON-CHEH, C., NILSSON, P., ORHO-MELANDER, M., RASTAM, L., SPELIOTES, E. K., TASKINEN, M. R., TUOMI, T., GUIDUCCI, C., BERGLUND, A., CARLSON, J., GIANNINY, L., HACKETT, R., HALL, L., HOLMKVIST, J., LAU-

rila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G. W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316** 1331–1336.

[26] Schweder, T. and Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69** 493–502.

[27] Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C. J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X. Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S., and Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316** 1341–1345.

[28] Shete, S., Hosking, F. J., Robertson, L. B., Dobbins, S. E., Sanson, M., Malmer, B., Simon, M., Marie, Y., Boisselier, B., Delattre, J. Y., Hoang-Xuan, K., El, H. S., Idbaih, A., Zelenika, D., Andersson, U., Henriksson, R., Bergenheim, A. T., Feychting, M., Lonn, S., Ahlbom, A., Schramm, J., Linnebank, M., Hemminki, K., Kumar, R., Hepworth, S. J., Price, A., Armstrong, G., Liu, Y., Gu, X., Yu, R., Lau, C., Schoemaker, M., Muir, K., Swerdlow, A., Lathrop, M., Bondy, M., and Houlston, R. S. (2009). Genome-wide association study identifies five susceptibility loci for glioma. *Nature Genetics* **41** 899–904.

[29] Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* **38** 209–213.

[30] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445** 881–885.

[31] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64** 479–498. MR1924302

[32] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100** 9440–9445. MR1994856

[33] Taylor, J., Tibshirani, R., and Efron, B. (2005). The 'miss rate' for the analysis of gene expression data. *Biostatistics* **6** 111–117.

[34] Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Jr., Hoover, R., Hayes, R. B., Hunter, D. J., and Chanock, S. J. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics* **40** 310–315.

[35] Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifs-

son, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T.,Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsater, A., Flex, A., Aben, K. K., de, V. F., Mulders, P. F., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., van Rij, A. M., Dieplinger, B., Haltmayer, M., Mayordomo, J. I., Kiemeney, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfingsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452** 638–642.

[36] Tomlinson, I. P., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A. M., Spain, S., Lubbe, S., Walther, A., Sullivan, K., Jaeger, E., Fielding, S., Rowan, A., Vijayakrishnan, J., Domingo, E., Chandler, I., Kemp, Z., Qureshi, M., Farrington, S. M., Tenesa, A., Prendergast, J. G., Barnetson, R. A., Penegar, S., Barclay, E., Wood, W., Martin, L., Gorman, M., Thomas, H., Peto, J., Bishop, D. T., Gray, R., Maher, E. R., Lucassen, A., Kerr, D., Evans, D. G., Schafmayer, C., Buch, S., Volzke, H., Hampe, J., Schreiber, S., John, U., Koessler, T., Pharoah, P., van, W. T., Morreau, H., Wijnen, J. T., Hopper, J. L., Southey, M. C., Giles, G. G., Severi, G., Castellvi-Bel, S., Ruiz-Ponte, C., Carracedo, A., Castells, A., Forsti, A., Hemminki, K., Vodicka, P., Naccarati, A., Lipton, L., Ho, J. W., Cheng, K. K., Sham, P. C., Luk, J., Agundez, J. A., Ladero, J. M., de la, H. M., Caldes, T., Niittymaki, I., Tuupanen, S., Karhu, A., Aaltonen, L., Cazier, J. B., Campbell, H., Dunlop, M. G., and Houlston, R. S. (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genetics* **40** 623–630.

[37] Wrensch, M., Jenkins, R. B., Chang, J. S., Yeh, R. F., Xiao, Y., Decker, P. A., Ballman, K. V., Berger, M., Buckner, J. C., Chang, S., Giannini, C., Halder, C., Kollmeyer, T. M., Kosel, M. L., LaChance, D. H., McCoy, L., O'Neill, B. P., Patoka, J., Pico, A. R., Prados, M., Quesenberry, C., Rice, T., Rynearson, A. L., Smirnov, I., Tihan, T., Wiemels, J., Yang, P., and Wiencke, J. K. (2009). Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nature Genetics* **41** 905–908.

[38] Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S., McCarthy, M. I., and Hattersley, A. T. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316** 1336–1341.

Zhuying Huang
Department of Epidemiology
The University of Texas M.D. Anderson Cancer Center
Houston, TX 77030
USA
E-mail address: zhuying_huang@yahoo.com

Jian Wang
Department of Epidemiology
The University of Texas M.D. Anderson Cancer Center
Houston, TX 77030
USA
E-mail address: jianwang@mdanderson.org

Chih-Chieh Wu
Department of Epidemiology
The University of Texas M.D. Anderson Cancer Center
Houston, TX 77030
USA
E-mail address: ccwu@mdanderson.org

Richard S. Houlston
Section of Cancer Genetics
Institute of Cancer Research
Sutton
United Kingdom
E-mail address: richard.houlston@icr.ac.uk

Melissa L. Bondy
Department of Epidemiology
The University of Texas M.D. Anderson Cancer Center
Houston, TX 77030
USA
E-mail address: mbondy@mdanderson.org

Sanjay Shete
Department of Epidemiology
The University of Texas M.D. Anderson Cancer Center
Houston, TX 77030
USA
E-mail address: sshete@mdanderson.org