RUTCOR
RESEARCH
REPORT

ESTIMATING THE AVERAGE EFFECT SIZE AND THE PROPORTION OF MARKERS WITHOUT EFFECT IN GENOMEWIDE ASSOCIATION STUDIES

József Bukszár[a] Edwin J. C. G. van den Oord[b]

RRR 9-2011, JUNE 27, 2011

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA
[b]Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA

Rutcor Research Report
RRR 9-2011, June 27, 2011

Estimating the average effect size and the proportion of markers without effect in genomewide association studies

József Bukszár Edwin J. C. G. van den Oord

Abstract. It has recently become possible to screen hundreds of thousands of genetic markers for their association with diseases. Knowledge of the proportion of markers without effect, p_0, and the effect sizes in these massive data sets has an intrinsic value and is required for a wide variety of applications. While numerous algorithms have been developed to estimate p_0, hardly any method is available to estimate effect sizes. We propose a maximum likelihood (ML) and a quasi-maximum likelihood (QML) approach for the simultaneous estimation of p_0 and the average effect size. The point estimate of any p_0 estimator can also be used in a 2-step procedure to estimate the average effect size through these (Q)ML methods. To avoid arbitrary choices of the fine-tuning parameter, needed for some p_0 estimators, we also developed a novel p_0 estimator where an (optimal) fine-tuning parameter is determined automatically through an iterative procedure. All estimators are illustrated for case-control studies for which we first derive an accurate approximation for the distribution of Pearson's statistic that depends on a single effect size parameter only. The two-step method appeared more accurate than the simultaneous estimation of both parameters. In this twostep procedure ML outperformed QML. ML combined with the Meinshausen-Rice estimator with fine-tuning parameter $\alpha = 0.5$ appeared to produce the best results in this genetic application. Our novel estimator was most precise among all studied p_0 estimators that did not require the pre-specification of a fine running parameter.

Acknowledgements: This work was supported by NIH grant R01 HG004240.

1 Introduction

Technology has recently become available that makes it possible to genotype up to 1 million genetic markers across the entire human genome. These large scale genetic studies offer great promise to expedite the discovery of genetic variants affecting susceptibility to common diseases [1], and [2]. Along with this potential come new statistical challenges, and accurate methods may need to be developed to better understand the properties of these massive data sets and extract meaningful information. Among the most basic features of these genetic data sets are the proportion of markers without effect p_0 and the effect sizes of markers with effects. Clearly, these two values are strongly connected; the average effect size cannot be interpreted without the number of markers with effect, or equivalently without p_0. In addition, information about both values is crucial for a wide variety of applications such as the control of false discoveries and designing follow-up studies that have adequate power to replicate the initial findings.

A variety of estimators for p_0 have been proposed [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. While numerous algorithms have been developed to estimate p_0, hardly any method is available to estimate effect sizes. To give an effect size estimator, or even just a proper definition of the effect size requires the distribution of the test statistic under the alternative. In fields such as genetics good approximations for the statistic distribution under the alternative are often available. In this article we propose maximum likelihood (ML) and quasi-maximum likelihood (QML) approaches that incorporate the test statistic distribution for the simultaneous estimation of p_0 and the average effect size. The ML approach relies on the fact that in a given set of markers the proportion of markers without effect (null markers) is a fixed number, p_0. The quasi-ML approach is based on the frequently used mixture model, in which the proportion of null-markers in a given set of markers is a binomially distributed random variable whose expected value is p_0, which is to be estimated. Also both ML approaches can be used to estimate the average effect size only based on a p_0 estimate. Existing and novel p_0 estimators are compared in Numerical Results section. An important feature of a p_0 estimator is whether it does or does not depend on any fine-tuning parameter(s). The choice of the fine-tuning parameters may strongly influence the precession of a p_0 estimator. Since the optimal choice usually depends on unknown parameters in a particular application, some researchers may feel uncomfortable to use p_0 estimators with fine-tuning parameters. We propose a novel p_0 estimator whose fine tuning parameter is not pre-specified, but computed adaptively by an iterative method. In principle, a similar adaptive version could be given for other (existing) p_0 estimators as well. However, this would require the calculation of their mean square error, which is usually a non-trivial task that needs further investigation.

We illustrate our estimators for case-control studies with single nucleotide polymorphisms (SNPs, [14]) that are one of the most important tools for mapping the genetic determinants of complex human diseases in large scale studies [15]. SNPs are bi-allelic so the statistical analysis often consists of calculating Pearson's statistic to test whether the frequency of the two alleles (A,a) or three SNP genotypes (AA, Aa, aa) differs between cases and controls.

A complication is that the asymptotic equivalent for the distribution of Pearson's statistic under the alternative hypothesis depends on many effect size parameters [16], which hampers the estimation. Therefore, we derived an accurate approximation for the distribution of Pearson's statistic that depends on a single effect size parameter Δ only.

2 Two maximum likelihood estimators

Suppose m hypothesis tests $H_1, ..., H_m$ are performed with statistics $T_1, ..., T_m$. *Exactly* m_0 of the m statistics follow the null distribution and the rest of them $m_1 = m - m_0$ follow alternative distribution whose density function is known but depends on an unknown parameter, called effect size, which may vary across the alternative hypotheses. We start with the aim of estimating the proportion of true null hypotheses $p_0 = m_0/m$ and the effect sizes of the true alternative hypotheses. In principle there are $m_1 = m - m_0$ effect sizes to estimate. However, even with $m_1 = 2$, the variances of the two effect size estimators were so large that we decided to use an alternative approach, which turned out to produce more precise effect size estimators. In this alternative approach we initially estimate the average effect size Δ. To estimate Δ we make the crude assumption that all individual effect sizes are identical. However, in Section 4 we show how this crude assumption can be "overwritten" when the actual individual effect sizes are calculated from Δ without making any assumption about their distribution.

Let $H_i = 0$ when null hypothesis i is true, and $H_i = 1$ otherwise. Note that vector $H = (H_1, ..., H_m)$ has m_0 0 and m_1 1 components. We assume that $H = (H_1, ..., H_m)$ is a random variable whose possible outcomes, the 0-1 vectors of length m with exactly m_1 1's, are taken with the same probability, $\binom{m}{m_1}^{-1}$. Note that $H_1, ..., H_m$ are not independent. Denote the distribution function of T_i by F_Δ (F_0) when $H_i = 1$ ($H_i = 0$), where we assume the same Δ, average effect size, for all alternatives. The joint distribution function of $T_1, ..., T_m$ on the test statistic values $t_1, ..., t_m$ will be

$$G(t_1, ..., t_m) = \frac{1}{\binom{m}{m_1}} \sum_H \Pr(T_1 < t_1, ..., T_m < t_m \mid H) =$$
$$\frac{1}{\binom{m}{m_1}} \sum_H \Pr(T_1 < t_1 \mid H) ... \Pr(T_m < t_m \mid H) = \frac{1}{\binom{m}{m_1}} \sum_H F_{H_1 \cdot \Delta}(t_1) ... F_{H_m \cdot \Delta}(t_m),$$

where the sums are on all possible outcomes of random variable H. By calculating the partial derivative $\frac{\partial^m G}{\partial t_1 ... \partial t_m}$ we obtain the likelihood function on the test statistic values $t_1, ..., t_m$, which has the form

$$L(m_1, \Delta) = \frac{1}{\binom{m}{m_1}} \sum_H f_{H_1 \cdot \Delta}(t_1) ... f_{H_m \cdot \Delta}(t_m) =$$
$$\frac{1}{\binom{m}{m_1}} \left(\prod_{i=1}^{m} f_0(t_i) \right) \sum_{\{i_1, ..., i_{m_1}\} \subseteq \{1, ..., m\}} \frac{f_\Delta(t_{i_1})}{f_0(t_{i_1})} \times ... \times \frac{f_\Delta(t_{i_{m_1}})}{f_0(t_{i_{m_1}})}, \quad (1)$$

where f_Δ (f_0) denotes the alternative (null) density function. The maximum likelihood estimator of p_0 and the average effect size are $\widehat{p}_0 = 1 - \widehat{m}_1/m$ and $\widehat{\Delta}$, respectively, where $\widehat{m}_1$ and $\widehat{\Delta}$ maximize function L. The effect sizes of alternative markers are unlikely to be the same. However, we do not know the distribution of the effect sizes in real life and even if we did this would add additional parameter(s) to the likelihood function which may hamper the parameter estimation. Therefore, we assumed uniform effect sizes but did study the robustness of this assumption by assuming non-uniform effect sizes in our simulations.

Due to enormous number of terms in the sum in (1) cannot be evaluated directly. We therefore developed the method below that is based on recursive series. Define $S(n)$ as

$$S(n) = \sum_{\{i_1,\ldots,i_n\}\subseteq\{1,\ldots,m\}} a_{i_1}\ldots a_{i_n} \tag{2}$$

for $n = 1, ..., m$, and $S(0) = 1$, where $a_i = \frac{f_\Delta(t_i)}{f_0(t_i)}$ for $i = 1, ..., m$. Then the maximum likelihood function can be re-written as

$$L(m_1, \Delta) = \frac{1}{\binom{m}{m_1}} \left(\prod_{i=1}^{m} f_0(t_i) \right) S(m_1). \tag{3}$$

One can verify the following sieve-formula

$$S(n) = \frac{1}{n} \sum_{i=1}^{n} (-1)^{i+1} R(i) S(n-i), \tag{4}$$

where $R(i) = \sum_{j=1}^{m} a_j^i$. By this formula we can calculate $S(m_1)$ in the real likelihood function in a recursive way starting from $S(0) = 1$ and $S(1) = \sum_{j=1}^{m} a_j$. The large spectrum of values of a_i's in combination with the recursive use of them will cause numerical problems when evaluating $S(n)$. Our computer implementation avoided these problems using a variety of techniques (see the R codes [17] on *http://www.people.vcu.edu/∼jbukszar/* for details). A major technique involved partitioning the set of a_is. That is, the distribution of a_i's is such that the vast majority of markers have values with small range. For this set we can use the recursive formula. For the remaining markers that have a very large range of a_i's, we created bins of 10 markers. Because there are only 10 markers in each bin, we don't need the recursive formula for which a large range is problematic. Instead, we calculated $S(n)$ directly using (2). The $S(n)$'s of all bins were then combined to calculate the $S(n)$ for all markers.

A similar estimator of p_0 can be proposed that use the log-likelihood function assuming the mixture model, which also reduces the number of terms:

$$\ell_{\text{quasi}}(p_0, \Delta) = \sum_{i=1}^{m} \log \{p_0 f_0(t_i) + (1 - p_0) f_\Delta(t_i)\}. \tag{5}$$

In the mixture model $H_1, ..., H_m$ are independent Bernoulli random variables with $\Pr(H_i = 0) = p_0$ and $\Pr(H_i = 1) = 1 - p_0$. Thus the selected markers are assumed to be drawn from a

population of markers with mixing proportions p_0 and $1-p_0$. The maximum likelihood estimate of p_0 and the average effect size will be the pair of $\left(\widehat{p}_0, \widehat{\Delta}\right)$ that maximizes the function in (5) on the observed data $t_1, ..., t_m$. Whereas the likelihood in (1) estimates the number of markers in the selected set of markers, (5) estimates the proportion of null markers in the population of all possible markers by drawing them from a large number of markers. It could be argued that in a study one is typically interested in the properties of the markers that have been genotyped rather than the properties of the (fictitious) population of all possible markers. Hence we call the likelihood of the mixture model a quasi-likelihood as it is a somewhat indirect way of estimating the proportion of effects among the genotyped markers. The mathematical connection between the quasi- and (real) likelihood function can be given as

$$L_{\text{quasi}}\left(p_0, \Delta\right) = \sum_{k=0}^{m} \binom{m}{k} p_0^{m-k} \left(1-p_0\right)^k L\left(k, \Delta\right),$$

where L_{quasi} is the quasi-likelihood function.

Note that we can combine the (quasi-)maximum likelihood estimator with any p_0 estimator to estimate the average effect size. Particularly, the ML estimate of average effect size will be $\widehat{\Delta}$ that maximize $L\left(m - mp_0^*, \Delta\right)$ or $\ell_{\text{quasi}}\left(p_0^*, \Delta\right)$, where p_0^* is an arbitrary p_0 estimate smaller than 1. This 2-step procedure average effect size estimator will be examined in the Numerical Methods section.

2.1 A single value approximation for Pearson's statistic

The (quasi) maximum likelihood estimator requires an approximating density function under the null, f_0, and the alternative, f_Δ, where Δ is a parameter. Clearly, f_0 and f_Δ depend on a particular application. In this section, we give f_Δ for unmatched case-control study when Pearson's statistic is used. It is well-known that central chi-square density function is typically chosen for f_0 in this setting.

In the case of the classical unmatched case-control study, data are often summarized in a contingency table:

$$\begin{bmatrix} x_1 & \cdots & x_\upsilon \\ y_1 & \cdots & y_\upsilon \end{bmatrix}, \tag{6}$$

where $\upsilon \geq 2$ refers to the total number of categories, $i = 1, ..., \upsilon$ is category index, and x_i and y_i are the number of cases and controls in category i, respectively. To analyze this table, Pearson's statistic is usually computed to test for differences between cases and controls. Denote the total sample size as n, and let γ and $\delta = 1-\gamma$ be the proportions of cases and controls in the total sample, respectively, i.e. $\sum_i x_i = n\gamma$ and $\sum_i y_i = n\delta$. Recall that Pearson's statistic on contingency table (6) is given by

$$T = \sum_{i=1}^{\upsilon} \left[\frac{\left(x_i - \frac{(x_i+y_i)\gamma n}{n}\right)^2}{\frac{(x_i+y_i)\gamma n}{n}} + \frac{\left(y_i - \frac{(x_i+y_i)\delta n}{n}\right)^2}{\frac{(x_i+y_i)\delta n}{n}} \right] = \sum_{i=1}^{m} \frac{\left(\delta x_i - \gamma y_i\right)^2}{\gamma\delta\left(x_i+y_i\right)}. \tag{7}$$

Define the *probability table*

$$\left[\begin{array}{ccc} p_1 & \cdots & p_\upsilon \\ q_1 & \cdots & q_\upsilon \end{array}\right], \tag{8}$$

where p_i $(i = 1, ..., \upsilon)$ is the probability that a randomly chosen case falls into category i, and q_i $(i = 1, ..., \upsilon)$ is the probability that a randomly chosen control falls into category i. We define Δ as

$$\Delta = \sqrt{\gamma\delta \sum\nolimits_{i=1}^{\upsilon} \frac{\left(p_i - q_i\right)^2}{\gamma p_i + \delta q_i}}.$$

Our single value approximation of Pearson's statistic on $2 \times \upsilon$ contingency tables has the form

$$\chi_{\upsilon-2} + \left(1 - \Delta^2\right) \chi_1 \left(\frac{n\Delta^2}{1 - \Delta^2}\right), \tag{9}$$

where $\chi_{\upsilon-2}$ is a (central) chi-square random variable with $\upsilon - 2$ degrees of freedom for $\upsilon > 2$, $\chi_{\upsilon-2} \equiv 0$ for $\upsilon = 2$, and $\chi_1 \left(\frac{n\Delta^2}{1-\Delta^2}\right)$ is a chi-square random variable with 1 degree of freedom and non-centrality parameter $\frac{n\Delta^2}{1-\Delta^2}$ (see Appendix for detail). Since $\Delta = 0$ if and only if $p_i = q_i$ for every i, the term in (9) is a central chi-square random variable with $\upsilon - 1$ degrees of freedom under the null hypothesis. The single scalar Δ can be interpreted as the *effect size*. For 2×2 tables, for example, Δ can be given as a function of the odds ratio $o = (p_1/p_2) / (q_1/q_2)$ and control allele frequency q_1 by

$$\Delta = \frac{\sqrt{\gamma\delta}\sqrt{q_1\left(1 - q_1\right)}\left(o - 1\right)}{\sqrt{\left(\left(o - 1\right)\left(\gamma + \delta q_1\right) + 1\right)\left(\left(o - 1\right)\delta q_1 + 1\right)}}. \tag{10}$$

In classic works on power analysis [18], categorical data analysis [19], and text books [20], the distribution of Pearson's statistic is often approximated with a non-central chi-square distribution with $\upsilon - 1$ degrees of freedom and non-centrality parameter $n\Delta^2$

$$\chi_{\upsilon-1}\left(n\Delta^2\right), \tag{11}$$

which also depends on the same single value Δ only. This approximation can be obtained by altering the derivation of (9) shown in the Appendix, particularly, by setting $\varepsilon_{\upsilon-1}$ to 1. However, $\varepsilon_{\upsilon-1}$ is always less than 1 under the alternative hypothesis, which gives a hint that this common approximation is probably less accurate than (9). In our simulations the approximation in (9) appeared more precise than the one in (11). However, f_Δ can be chosen as the density in (11) and used in our maximum likelihood framework.

3 Conservative estimator

In this section, we give a p_0 estimator whose near-optimal fine-tuning parameter can be computed adaptively, which would be very difficult for existing p_0 estimators with fine-tuning parameters. This p_0 estimator does not rely on the test statistic distribution under the alternative but capitalizes on the fact that in large scale genetic studies p_0 is close to 1.

We calculate a cut-off value c in such a way that the probability that a null marker has test statistic value higher than c is k/m. If we denote the total number of markers whose test statistic value is higher than c as d, then this estimate of p_0 is

$$\widehat{p}_0 = 1 - \frac{d-k}{m}.$$

Note that the expected number of null markers with test statistic value higher than cut-off c is km_0/m rather than k. This estimator can therefore be expected to be conservatively biased, hence we will call it Conservative estimator. However, because $p_0 = m_0/m$ is close to 1 we would expect the bias to be small.

Let us denote the effect sizes of alternative hypotheses as $\Delta_1, ..., \Delta_{m_1}$. By taking the expectation we obtain

$$E\left(\widehat{p}_0\right) = p_0 + \frac{1}{m}\sum_{i=1}^{m_1}\left(1 + k/m - \pi^{(i)}\right),$$

where $\pi^{(i)} = 1 - F_{\Delta_i}\left(F_0^{-1}\left(1-k/m\right)\right)$, and F_0 and F_Δ are the proper cumulative distribution functions under the null and alternative hypothesis, respectively. In our numerical examples assuming a case-control design we will use the central chi-square cdf with $\upsilon - 1$ degrees for F_0 and the approximation in (9) for F_Δ. Note that the bias of $\widehat{p}_0$ is positive and $E\left(\widehat{p}_0\right) < 1$. This is because $k/m + F_{\Delta_i}\left(F_0^{-1}\left(1-k/m\right)\right) < 1$ or equivalently $F_0^{-1}\left(1-k/m\right) < F_{\Delta_i}^{-1}\left(1-k/m\right)$.

A natural idea is to choose a value for fine tuning parameter k that minimizes the mean square error

$$\begin{aligned} MSE\left(k\right) &= E(\widehat{p}_0 - p_0)^2 \\ &= \frac{1}{m^2}\left[\left\{\sum_{i=1}^{m_1}\left(1 + k/m - \pi^{(i)}\right)\right\}^2 + p_0 k\left(1 - \frac{k}{m}\right) + \sum_{i=1}^{m_1}\pi^{(i)}\left(1-\pi^{(i)}\right)\right. \\ &\left. +2\sum_{1\le i<j\le m} Cov\left(I_{\{T_i>c\}}, I_{\{T_j>c\}}\right)\right], \end{aligned} \tag{12}$$

where I_A is the indicator random variable of event A, i.e. I_A equals 1 if A occurs and 0 otherwise, and T_i is test statistic for marker i. It is reasonable to assume that T_i and T_j are independent if one of them corresponds to a null marker and the other one to a non-null marker. Furthermore,

$$Cov\left(I_{\{T_i>c\}} \,|\, H_0,\ I_{\{T_j>c\}} \,|\, H_0\right) = \Pr\left(T_i > c, T_j > c \,|\, H_0\right) - \left(\frac{k}{m}\right)^2 \tag{13}$$

and

$$Cov\left(I_{\{T_i>c\}} \,|\, H_1,\ I_{\{T_j>c\}} \,|\, H_1\right) = \Pr\left(T_i > c, T_j > c \,|\, H_1\right) - \pi^{(i)}\pi^{(j)}. \tag{14}$$

For example, for 2×2 tables the probability in (13) can be computed by exploiting the result in (9) that T_i and T_j are each a square of a normal random variable with correlation C.

A practical problem is that the value of k that minimizes the MSE defined in (12) depends on the unknown parameters p_0 and Δ_i through $\pi^{(i)}$. In addition, although the covariances among the markers can be observed, some simplifying assumptions may be required for practical reasons. A first approach for choosing k is to make assumptions about these unknowns and calculate the corresponding optimum. For example, assume a study with $100,000$ markers and a total sample size (cases+controls) of $1,500$. Because each individual contributes 2 alleles, this implies $n = 3,000$ for testing whether the frequencies of the two alleles (A,a) differs between cases and controls. In Figure 1 we assumed the range $p_0 =$ 0.9995...99995 implying 5...50 markers with effects. In large scale genetic studies where markers are typically selected without any prior knowledge regarding their relation to the outcome of interest, $p_0 = 0.99995$ is probably the more realistic value [21] but studying this broad range will help obtain an impression about the variability in the optimal k. For the average effect size Δ we assumed values of 0.07, 0.08, and.0.09. For example, $\Delta = 0.08$ would be obtained with uniformly distributed effect sizes Δ_i that range from 0.04 to 0.12 as calculated by (10) when assuming minor allele frequencies ranging from 0.1 to 0.5 and allelic odds ratios ranging from 1.3 to 1.7.

Figure 1 shows that for $p_0 = 0.99995$, the optimal k is close to 1 regardless of Δ. For smaller values of p_0, the optimal k is more dependent on Δ. Figure 1 assumes independent markers. We also studied the impact of covariances among the markers on the optimal k. The dependency among markers in the human genome can be characterized by a block structure where within a block correlations are high and between blocks correlations are low. In addition, geneticists will generally avoid genotyping markers that are highly correlated because those markers will be redundant [22]. Based on these observations, we assumed blocks of five markers with statistical association among the markers of $R^2 = 0.5$. To calculate the MSE with (12) we need C^2 rather than R^2. However, in Appendix we show that under the null hypothesis $C^2 = R^2$ and a more general proof including the alternative hypothesis can be found on our web site *http://www.vipbg.vcu.edu/~edwin/*. Results (not shown) indicated that the optimal k hardly depended on R^2 when $R^2 \leq 0.5$ suggesting that in large scale genetic studies correlations can be ignored when choosing k. To enable researchers to tailor the choice of k to their specific study we also put the R-code for these calculations on our web site.

A second approach is to choose k in an adaptive way. First, we calculate $\widehat{p}_0$ with some k, e.g. $k = 1$. Second, for the value of $\widehat{p}_0$ we use either maximum likelihood or quasi maximum likelihood method to estimate the effect size. Third, for $\widehat{p}_0$ and the effect size estimate we calculate the optimal k. We repeat steps 1 to 3 with this new k until there is no noticeable change in k.

4 Numerical Results

In the first part of this section we compare the proposed p_0 estimators with existing ones. In the second part we compare the maximum likelihood (ML) and quasi-maximum likelihood (QML) average effect size estimators based on the p_0 estimators examined in the first

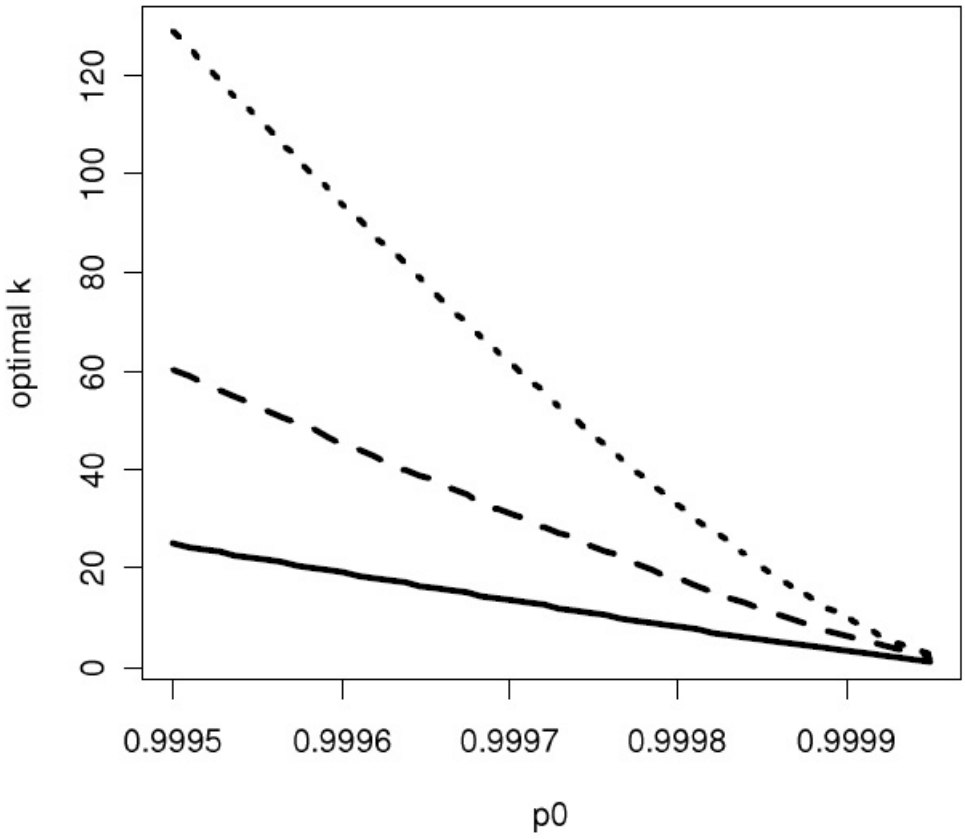

Figure 1: The optimal k that minimizes the MSE of the Conservative p_0 estimator for different values of p_0 and effect sizes when $m = 100,000$ and $n = 3,000$. Effect size was 0.07 (dotted line), 0.08 (dashed line), and 0.09 (solid line).

part. An important feature of a p_0 estimator is whether it does or does not depend on any fine-tuning parameter(s). As we will see, the choice of the fine-tuning parameters may strongly influence the precession of a p_0 estimator. As a result, some researchers may feel uncomfortable to use p_0 estimators with fine-tuning parameters. Therefore, we show their numerical results separately, results of p_0 estimators without (with) fine-tuning parameters are shown in Table 1 and 3 (Table 2 and 4).

Simulations were performed to evaluate the proposed maximum likelihood (ML), quasi-maximum likelihood (QML), and Conservative estimator (CON) and its adaptive variant described in the last paragraph of the previous section (ADA CON), and they are compared with existing p_0 estimators. We found two studies [13] and [5] comparing multiple and non-overlapping sets of estimators for p_0. Our estimators were compared with the Lowest slope [9] and Location based estimator [5] that were shown to have the most favorable properties in these two studies. In addition, estimators developed by Storey [7] and Meinshausen and Rice [23] were included. Storey's estimator is commonly used and Meinshausen-Rice method

is specifically designed for scenarios like this one where p_0 is close to 1.

Similar to the scenario studied in Figure 1, in the first (baseline) condition we assumed $100,000$ independent markers, $p_0 = 0.99995$, effect sizes Δ_i in the range $0.04, \ldots, 0.12$, and a total sample size (cases+controls) of $1,500$. In the second condition we assumed blocks of five dependent markers with $R^2 = 0.5$ within each block. In conditions 3 the samples size was four times of the sample size in the baseline condition and in condition 4 we doubled the number of markers of the baseline condition. The maximum likelihood methods are computer intensive and we therefore limited ourselves to $1,000$ replications in each condition.

Table 1: The mean, standard deviation and root of the mean square error (RMSE) of p_0 estimates on thousand replications. In the first row in every condition the number of times the estimator crashed and the number of times it was 1 are also indicated, respectively. In each condition $p_0 = 0.99995$, and effect sizes are $0.04, ..., 0.12$ (equidistant).

	ML	QML	ADA CON	LOW S	LBE
	$m = 100,000$, $n = 1,500$, independent				
# crashed, # 1's	1, 0	1, 0	0, 32	0, 351	0, 496
Mean	.999934	.999927	.999975	.999986	.909963
Standard dev.	8.48×10^{-4}	8.50×10^{-4}	1.23×10^{-5}	1.57×10^{-5}	1.42×10^{-2}
RMSE	8.48×10^{-4}	8.51×10^{-4}	2.81×10^{-5}	3.92×10^{-5}	1.74×10^{-2}
	$m = 100,000$, $n = 1,500$, $R^2 = 0.5$, 5×5				
# crashed, # 1's	9, 7	9, 10	0, 86	0, 411	0, 504
Mean	.999806	.999753	.999976	.999985	.989565
Standard dev.	2.46×10^{-3}	2.70×10^{-3}	1.36×10^{-5}	1.89×10^{-5}	1.52×10^{-2}
RMSE	2.47×10^{-3}	2.70×10^{-3}	2.98×10^{-5}	3.99×10^{-5}	1.84×10^{-2}
	$m = 100,000$, $n = 6,000$, independent				
# crashed, # 1's	0, 0	0, 0	0, 0	0, 40	0, 496
Mean	.999963	.999963	.999959	.999968	.909963
Standard dev.	5.89×10^{-6}	5.89×10^{-6}	4.97×10^{-6}	1.73×10^{-5}	1.42×10^{-2}
RMSE	1.44×10^{-5}	1.44×10^{-5}	1.04×10^{-5}	2.51×10^{-5}	1.74×10^{-2}
	$m = 200,000$, $n = 1,500$, independent				
# crashed, # 1's	2, 0	2, 0	0, 0	0, 41	0, 483
Mean	.999970	.999966	.999977	.999982	.986509
Standard dev.	2.04×10^{-5}	2.54×10^{-5}	7.26×10^{-6}	1.06×10^{-5}	1.91×10^{-2}
RMSE	2.86×10^{-5}	3.03×10^{-5}	2.78×10^{-5}	3.33×10^{-5}	2.34×10^{-2}

ML: Maximum likelihood, QML: Quasi maximum likelihood, ADA CON: Adaptive conservative, LOW S: Lowest slope, LBE: Location based estimator

In Table 1 and 2 p_0 estimators with and without fine-tuning parameters, respectively, are compared. Since it is computationally intensive to evaluate ML function when the number of markers with effect is large, we used QML instead of ML function when the estimated number of markers with effect exceeded 150. This happened 6 and 23 times in the first two conditions, respectively. We also indicated the number of replications out of 1000 when the p_0 estimators crashed. Only ML and QML p_0 estimators crashed. The ML is slightly more precise than QML, both of them are outperformed by ADA CON, which gives the most

Table 2: The mean, standard deviation and root of the mean square error (RMSE) of p_0 estimates on thousand replications. In the first row in every condition the number of times the estimator crashed and the number of times it was 1 are also indicated, respectively. In each condition $p_0 = 0.99995$, and effect sizes are $0.04, ..., 0.12$ (equidistant).

	CON	STO1	STO2	MEI-RIC1	MEI-RIC2
	$m = 100,000$, $n = 1,500$, independent				
# crashed, # 1's	0, 29	0, 195	0, 10	0, 8	0, 1
Mean	.999975	.994352	.999962	.999980	.999971
Standard dev.	1.28×10^{-5}	7.15×10^{-3}	2.19×10^{-5}	7.74×10^{-6}	1.16×10^{-5}
RMSE	2.78×10^{-5}	9.08×10^{-3}	2.50×10^{-5}	3.10×10^{-5}	2.39×10^{-5}
	$m = 100,000$, $n = 1,500$, $R^2 = 0.5$, 5×5				
# crashed, # 1's	0, 66	0, 286	0, 22	0, 42	0, 12
Mean	.999976	.994016	.999961	.999980	.999976
Standard dev.	1.41×10^{-5}	7.70×10^{-3}	2.73×10^{-5}	9.41×10^{-6}	1.46×10^{-5}
RMSE	2.94×10^{-5}	9.72×10^{-3}	2.95×10^{-5}	3.19×10^{-5}	2.54×10^{-5}
	$m = 100,000$, $n = 6,000$, independent				
# crashed, # 1's	0, 0	0, 201	0, 40	0, 0	0, 0
Mean	.999954	.994353	.999944	.999957	0.999950
Standard dev.	1.17×10^{-5}	7.15×10^{-3}	2.28×10^{-5}	5.65×10^{-6}	1.01×10^{-5}
RMSE	1.25×10^{-5}	9.08×10^{-3}	2.36×10^{-5}	9.13×10^{-6}	1.01×10^{-5}
	$m = 200,000$, $n = 1,500$, independent				
# crashed, # 1's	0, 0	0, 240	0, 0	0, 1	0, 0
Mean	.999977	.996117	.999964	.999982	.999975
Standard dev.	7.47×10^{-6}	5.21×10^{-3}	1.66×10^{-5}	5.44×10^{-6}	7.43×10^{-6}
RMSE	2.76×10^{-5}	6.47×10^{-3}	2.15×10^{-5}	3.22×10^{-5}	2.57×10^{-5}

CON: Conservative with $k = 1$, STO1 and 2: Storey with grid $\Gamma = (0, 0.05, ..., 0.95)$ and $\Gamma = \big(0, 10^{-8}, ..., 10^{-7}, \quad 2 \times 10^{-7}, ..., 10^{-6}, \quad 2 \times 10^{-6}, ..., 10^{-5}, \quad 2 \times 10^{-5}, ..., 10^{-4}\big)$, resp., MEI-RIC1 and 2: Meinshausen-Rice with $\alpha = 0.1$ and $\alpha = 0.5$, resp.

precise estimates of p_0 among the estimators not relying on fine-tuning parameters (Table 1). The drawback of ADA CON vs ML is that it estimates p_0 to be 1 32 and 86 times out of 1000 in condition 1 and 2 respectively, in which cases the effect size estimator could not be used. Basically we could use the ML estimators more times than ADA CON for average effect size estimates. Lowest slope (LOW S) was just a bit less good than ADA CON in terms of root of the mean square error (RMSE), however, it estimated p_0 to be 1 substantially more times than ADA CON. The Location based estimator (LBE) estimated p_0 to be 1 in roughly half of the simulations and returned a value lower than 0.99991 in the other half, which is reflected in its high variance.

In Table 2 we show the results of Storey's method with the grid he suggests in his article [7] (STO1), and also with the grid $\Gamma = (0, 10^{-8}, ..., 10^{-7}, \ 2 \times 10^{-7}, ..., 10^{-6}, \ 2 \times 10^{-6}, ..., 10^{-5}, \ 2 \times 10^{-5}, ..., 10^{-4})$ (STO2), with which we found Storey's method perform well in these particular numerical examples. We also show Meinshausen-Rice [23] method with linear

bounding function and fine-tuning parameters $\alpha = 0.1$ (MEI-RIC1) and $\alpha = 0.5$ (MEI-RIC2). Here $\alpha = 0.1$ is their choice, albeit in a different context, and $\alpha = 0.5$ is our suggestion for using Meinshausen-Rice to estimate p_0. Table 2 shows that among the estimators relying on fine-tuning parameters, the Conservative method (CON) with $k = 1$, STO2, MEI-RIC1, and MEI-RIC2 performed roughly equally well in terms of RMSE. However, it was MEI-RIC2 that estimated p_0 to be 1 the fewest times and had the smallest bias. The Conservative method performed well even though the optimal k, as calculated theoretically using (12), was different from $k = 1$ (e.g. $k \cong 5$ in condition 4). The RMSE obtained with the optimal k was often very similar to the RMSE obtained with $k = 1$ suggesting that the Conservative method is fairly robust against mis-specifications of k.

Estimates of the average effect size are shown in Table 3 and 4. The two ML methods estimate Δ simultaneously with p_0. For all other estimators we used a two step procedure described in the last paragraph of section 2. Average effect size estimates based on p_0 estimators without (with) fine-tuning parameters are shown in Table 3 (Table 4). Simulations where p_0 estimates were 1 were removed from the analysis, because the average effect size cannot be estimated. Estimates based on the perfect p_0 estimator, i.e. the one that always returns the real p_0 (0.99995), are also shown in Table 3. The rationale of adding the perfect p_0 estimator to the table is twofold. First, it shows the limit of the precision that can be achieved by ML or QML average effect size estimators. Second, it sheds light on the different behavior of ML vs. QML estimators. The perfect p_0 based ML estimator outperforms the perfect p_0 based QML substantially in terms of bias, standard deviation and RMSE in all conditions. The difference between the performance of the ML and QML average effect size estimators is more spectacular for higher sample sizes. Similarly, also ML average effect size estimator based on a good p_0 estimator, e.g. STO2 and MEI-RIC2, outperforms QML average effect size estimator based on the same p_0 estimator. Here again the difference between the performance of the two ML average effect size estimators is more spectacular for higher sample sizes, e.g. ML based on any reasonably good p_0 estimator except LOW S outperforms QML when $n = 6000$. It is intuitive, because good p_0 estimators are more precise for higher samples sizes, i.e. they "are closer to the perfect p_0 estimator". The reason why LOW S based ML does not perform well is that LOW S is highly upward biased.

Note that when sample size is 3000 or higher, QML average effect size estimator based on any reasonably good p_0 estimator (incl. Perfect) are equally precise, both the mean and the mean square error are practically the same for all of them. Interestingly enough, ML behaves in the opposite way, i.e. the ML average effect size estimate is getting even more sensitive to the value of the given p_0 with increasing sample size. Therefore the accuracy of the p_0 estimator is crucial for ML average effect size estimates while it is almost irrelevant for QML average effect size estimates especially for higher sample sizes. This phenomenon also explains why it is advisable to use p_0 estimator with little bias for ML average effect size estimator. ML based on MEI-RIC2 is more precise than ML based on MEI-RIC1 when $n = 6000$ because MEI-RIC2 p_0 estimator has less bias than MEI-RIC1 even though it has higher RMSE. In the other scenarios STO2 gives the best ML average effect size estimates due to its low bias and relatively low standard deviation. STO2 is based on

Table 3: The mean, standard deviation and root of the mean square error (RMSE) of ML and QML average effect size estimates based on different p_0 estimators on thousand replications. In the first row in every condition the number of times the p_0 estimator crashed and the number of times it was 1 are also indicated, respectively. In each condition $p_0 = 0.99995$, and effect sizes are $0.04, ..., 0.12$ (equidistant).

		Perfect	ML	QML	ADA CON	LOW S	LBE
		$m = 100,000$, $n = 1,500$, independent					
ML	# crashed, # 1's	0, 0	1, 0	-	0, 32	0, 351	0, 496
	Mean	.0888	.1100		.1106	.1131	.0028
	Standard dev.	8.67×10^{-3}	1.84×10^{-2}		1.46×10^{-2}	1.67×10^{-2}	1.14×10^{-2}
	RMSE	1.24×10^{-2}	3.52×10^{-2}		3.39×10^{-2}	3.70×10^{-2}	7.81×10^{-2}
QML	# crashed, # 1's	0, 0		1, 0	0, 32	0, 351	0, 496
	Mean	.0999		.1057	.1082	.1076	.0016
	Standard dev.	1.86×10^{-2}	-	2.06×10^{-2}	1.65×10^{-2}	1.61×10^{-2}	9.65×10^{-3}
	RMSE	2.73×10^{-2}		3.29×10^{-2}	3.27×10^{-2}	3.19×10^{-2}	7.90×10^{-2}
		$m = 100,000$, $n = 1,500$, $R^2 = 0.5$, 5×5					
ML	# crashed, # 1's	0, 0	9, 7	-	0, 86	0, 411	0, 504
	Mean	.0865	.1070		.1088	.1118	.0032
	Standard dev.	1.51×10^{-2}	1.68×10^{-2}		1.28×10^{-2}	1.91×10^{-2}	1.08×10^{-2}
	RMSE	1.65×10^{-2}	3.18×10^{-2}		3.15×10^{-2}	3.71×10^{-2}	7.76×10^{-2}
QML	# crashed, # 1's	0, 0		9, 10	0, 86	0, 411	0, 504
	Mean	.0937		.0999	.1056	.1057	.0010
	Standard dev.	2.12×10^{-2}	-	2.16×10^{-2}	1.37×10^{-2}	1.57×10^{-2}	3.92×10^{-3}
	RMSE	2.52×10^{-2}		2.94×10^{-2}	2.90×10^{-2}	3.02×10^{-2}	7.91×10^{-2}
		$m = 100,000$, $n = 6,000$, independent					
ML	# crashed, # 1's	0, 0	0, 0	-	0, 0	0, 40	0, 496
	Mean	.0807	.0942		.0898	.0978	.0021
	Standard dev.	3.91×10^{-3}	7.58×10^{-3}		6.31×10^{-3}	1.52×10^{-2}	9.30×10^{-3}
	RMSE	3.98×10^{-3}	1.61×10^{-2}		1.17×10^{-2}	2.34×10^{-2}	7.84×10^{-2}
QML	# crashed, # 1's	0, 0		0, 0	0, 0	0, 40	0, 496
	Mean	.0941		.0942	.0941	.0943	.0013
	Standard dev.	7.54×10^{-3}	-	7.52×10^{-3}	7.55×10^{-3}	7.51×10^{-3}	8.49×10^{-3}
	RMSE	1.60×10^{-2}		1.61×10^{-2}	1.60×10^{-2}	1.61×10^{-2}	7.92×10^{-2}
		$m = 200,000$, $n = 1,500$, independent					
ML	# crashed, # 1's	0, 0	2, 0	-	0, 0	0, 41	0, 483
	Mean	.0875	.1055		.1067	.1103	.0006
	Standard dev.	6.27×10^{-3}	1.45×10^{-2}		8.47×10^{-3}	1.21×10^{-2}	5.76×10^{-4}
	RMSE	9.80×10^{-3}	2.94×10^{-2}		2.80×10^{-2}	3.26×10^{-2}	7.94×10^{-2}
QML	# crashed, # 1's	0, 0		2, 0	0, 0	0, 41	0, 483
	Mean	.0935		.1023	.1041	.1047	.0006
	Standard dev.	1.01×10^{-2}	-	1.41×10^{-2}	1.03×10^{-2}	9.76×10^{-3}	5.76×10^{-4}
	RMSE	1.69×10^{-2}		2.64×10^{-2}	2.62×10^{-2}	2.65×10^{-2}	7.94×10^{-2}

ML: Maximum likelihood, QML: Quasi maximum likelihood, ADA CON: Adaptive conservative, LOW S: Lowest slope, LBE: Location based estimator

the grid $\Gamma = (0, 10^{-8}, ..., 10^{-7}, \ 2 \times 10^{-7}, ..., 10^{-6}, \ 2 \times 10^{-6}, ..., 10^{-5}, \ 2 \times 10^{-5}, ..., 10^{-4})$, which may work worse for other numerical examples, of course. The selection of fine-tuning parameters is beyond the scope of this article. However, for Meinshausen-Rice estimator $\alpha = 0.5$ seems a reasonable choice, because $1 - \alpha$ is a tight lower bound for the probability that the estimator is a lower bound for p_0 [23]. Therefore, with the choice $\alpha = 0.5$ the

Table 4: The mean, standard deviation and root of the mean square error (RMSE) of ML and QML average effect size estimates based on different p_0 estimators on thousand replications. In the first row in every condition the number of times the p_0 estimator crashed and the number of times it was 1 are also indicated, respectively. In each condition $p_0 = 0.99995$, and effect sizes are $0.04, ..., 0.12$ (equidistant).

		CON	STO1	STO2	MEI-RIC1	MEI-RIC2
		$m = 100,000$, $n = 1,500$, independent				
ML	# crashed, # 1's	0, 29	0, 195	0, 10	0, 8	0, 1
	Mean	.1086	.0071	.1006	.1146	.1056
	Standard dev.	1.47×10^{-2}	1.67×10^{-2}	1.48×10^{-2}	1.27×10^{-2}	1.26×10^{-2}
	RMSE	3.22×10^{-2}	7.48×10^{-2}	2.53×10^{-2}	3.68×10^{-2}	2.85×10^{-2}
QML	# crashed, # 1's	0, 29	0, 195	0, 10	0, 8	0, 1
	Mean	.1075	.0011	.1027	.1104	.1067
	Standard dev.	1.64×10^{-2}	5.10×10^{-3}	2.07×10^{-2}	1.53×10^{-2}	1.65×10^{-2}
	RMSE	3.20×10^{-2}	7.91×10^{-2}	3.07×10^{-2}	3.40×10^{-2}	3.14×10^{-2}
		$m = 100,000$, $n = 1,500$, $R^2 = 0.5$, 5×5				
ML	# crashed, # 1's	0, 66	0, 286	0, 22	0, 42	0, 12
	Mean	.1066	.0061	.0987	.1120	.1029
	Standard dev.	1.51×10^{-2}	1.55×10^{-2}	1.64×10^{-2}	1.06×10^{-2}	1.31×10^{-2}
	RMSE	3.06×10^{-2}	7.55×10^{-2}	2.49×10^{-2}	3.37×10^{-2}	2.63×10^{-2}
QML	# crashed, # 1's	0, 66	0, 286	0, 22	0, 8	0, 12
	Mean	.1039	.0014	.0969	.1072	.1022
	Standard dev.	1.55×10^{-2}	6.72×10^{-3}	2.35×10^{-2}	1.25×10^{-2}	1.59×10^{-2}
	RMSE	2.85×10^{-2}	7.89×10^{-2}	2.89×10^{-2}	2.99×10^{-2}	2.73×10^{-2}
		$m = 100,000$, $n = 6,000$, independent				
ML	# crashed, # 1's	0, 0	0, 201	0, 40	0, 0	0, 0
	Mean	.0860	.0056	.0781	.0880	.0825
	Standard dev.	1.01×10^{-2}	1.59×10^{-2}	1.09×10^{-2}	6.48×10^{-3}	8.10×10^{-3}
	RMSE	1.17×10^{-2}	7.61×10^{-2}	1.11×10^{-2}	1.03×10^{-2}	8.48×10^{-3}
QML	# crashed, # 1's	0, 0	0, 201	0, 40	0, 0	0, 0
	Mean	.0941	.0007	.0908	.0941	.0941
	Standard dev.	7.53×10^{-3}	3.40×10^{-4}	1.83×10^{-2}	7.53×10^{-3}	7.53×10^{-3}
	RMSE	1.60×10^{-2}	7.94×10^{-2}	2.13×10^{-2}	1.60×10^{-2}	1.60×10^{-2}
		$m = 200,000$, $n = 1,500$, independent				
ML	# crashed, # 1's	0, 0	0, 240	0, 0	0, 1	0, 0
	Mean	.1062	.0048	.0973	.1118	.1043
	Standard dev.	8.41×10^{-3}	1.40×10^{-2}	9.38×10^{-3}	7.78×10^{-3}	7.52×10^{-3}
	RMSE	2.75×10^{-2}	7.65×10^{-2}	1.96×10^{-2}	3.27×10^{-2}	2.54×10^{-2}
QML	# crashed, # 1's	0, 0	0, 240	0, 0	0, 1	0, 0
	Mean	.1039	.0007	.0989	.1062	.1031
	Standard dev.	1.02×10^{-2}	5.08×10^{-4}	1.09×10^{-2}	9.59×10^{-2}	9.97×10^{-3}
	RMSE	2.60×10^{-2}	7.93×10^{-2}	2.18×10^{-2}	2.79×10^{-2}	2.52×10^{-2}

CON: Conservative with $k = 1$, STO1 and 2: Storey with grid $\Gamma = (0, 0.05, ..., 0.95)$ and $\Gamma = \left(0, 10^{-8}, ..., 10^{-7}, \quad 2 \times 10^{-7}, ..., 10^{-6}, \quad 2 \times 10^{-6}, ..., 10^{-5}, \quad 2 \times 10^{-5}, ..., 10^{-4}\right)$, resp., MEI-RIC1 and 2: Meinshausen-Rice with $\alpha = 0.1$ and $\alpha = 0.5$, resp.

probability that Meinshausen-Rice estimator is either lower or higher than the real p_0 is roughly equal, 0.5, so intuitively the estimator should be close to the median resulting in a low bias, and as we have seen above ML average effect size estimator based on a p_0 estimator with a low bias is precise. ML average effect size estimators based on Meinshausen-Rice p_0 estimator give the best results for higher sample size. A grid may or may not exist for Storey estimator that outperforms Meinshausen-Rice estimator.

5 Discussion

In this article we proposed maximum likelihood (ML) and quasi-maximum likelihood (QML) approaches that incorporate the test statistic distribution for the simultaneous estimation of p_0 and the average effect size. The point estimate of any p_0 estimator can also be used in a 2-step procedure to estimate the average effect size through these (Q)ML methods. All estimators were illustrated for case-control studies for which we derived an accurate approximation for the distribution of Pearson's statistic that depends on a single effect size parameter only. The 2-step method appeared more accurate than the simultaneous estimation of both parameters. In this 2-step procedure ML outperformed QML. Moreover, ML combined with Meinshausen-Rice p_0 estimator with fine-tuning parameter $\alpha = 0.5$ [23] appeared to produce the best results in this genetic application. To avoid arbitrary choices of the fine-tuning parameter, needed for some p_0 estimators, we also developed a novel, called Adaptive Conservative, p_0 estimator, where an (optimal) fine-tuning parameter is determined automatically through an iterative procedure. The Adaptive Conservative estimator was the most precise among all studied p_0 estimators that did not require the pre-specification of a fine running parameter.

First we discuss the p_0 estimators, then the average effect size estimators because the latter ones rely on the former ones. We proposed two estimators of p_0. The Conservative estimator, which requires a fine-tuning parameter, and its adaptive version, which does not require any fine-tuning parameter. However, the adaptive version relies on average effect size estimate, therefore, utilizes that the test statistic distribution with some possibly unknown parameters is known under the alternative. Both estimators take advantage of the information that in large scale genetic studies p_0 must be very close to 1.

Through simulations, the proposed p_0 estimators were compared with existing ones that are commonly used or showed the most favorable properties in the two studies comparing multiple and non-overlapping sets of estimators. The Conservative, Meinshausen-Rice and Storey's methods with suitably chosen fine-tuning parameters appeared to be the best and roughly equally good in terms of RMSE. However, finding the optimal or good enough fine-tuning parameter for p_0 estimators may be a difficult problem by itself. Among the p_0 estimators not relying on fine-tuning parameters, the adaptive variant of the Conservative estimator appeared to be the most precise. In principle, an adaptive version could be given for the Meinshausen-Rice estimator. However, this would require the calculation of the RMSE of the Meinshausen-Rice estimator, which uses a supremum, hence this is a non-trivial task that needs further investigation.

We proposed a maximum likelihood (ML) and quasi-ML approach for the simultaneous estimation of p_0, and the average effect size for applications where the alternative distribution is known up to some unknown parameters. With our maximum likelihood (ML) approach we used the likelihood function of the sample of a multiple hypothesis test to estimate p_0, and the average effect size *in the sample*. In the alternative quasi-maximum likelihood (QML) approach we used the likelihood function of the mixture model instead of the "real" likelihood function. The fundamental principle difference between the ML and QML approach is that ML (correctly) assumes that the number of true null hypotheses is constant in a given setup whereas QML assumes it is a binomially distributed random variable. Both the ML and QML methods also can be used in combination with a p_0 estimator to estimate the average effect size.

The ML vs. QML average effect size estimators based on different p_0 estimators, and the versions that simultaneously estimate p_0 and the average effect size were also compared through simulations. Both ML and QML average effect size estimators when based on good p_0 estimators appeared more accurate than their version that simultaneously estimate p_0 and the average effect size. Moreover, ML were more precise than QML when both were based on (the same) good p_0 estimators. It is especially true for higher sample size, where QML is not even sensitive to the precession of p_0 estimator. ML behaves in the opposite way, i.e. the accuracy of the p_0 estimator is crucial for ML average effect size estimates. This phenomenon also explains why it is advisable to use p_0 estimator with little bias for ML average effect size estimator. An example is that ML based on Meinshausen-Rice p_0 estimator gave better effect size estimate for sample size $n = 6000$ when Meinshausen-Rice's fine-tuning parameter was $\alpha = 0.5$ than when it was $\alpha = 0.1$, in spite of the fact that the latter one resulted in somewhat smaller variance and RMSE. An important feature of a p_0 estimator when it is used to estimate the average effect size in combination of ML or QML method is the number of times it (erroneously) estimates p_0 to be 1. Since the number of markers with effect is estimated to be zero, the average effect size cannot be estimated. The Lowest Slope and Location Based Estimator did not perform well from this aspect in the underlying applications, although they proved very good in others [9] and [5]. As a conclusion, we suggest to use ML method based on a good p_0 estimator to estimate average effect size. Meinshausen-Rice p_0 estimator with $\alpha = 0.5$ seems to be a good choice because of its low bias, or the adaptive version of the conservative p_0 estimator is recommended for researchers who feel uncomfortable to use estimators that require any fine-tuning parameters. Alternatively, simulations can be run to find good fine-tuning parameters. In addition, p_0 estimators more accurate than the existing ones may be invented in the future. These p_0 estimators when combined with our ML method may result in a better average effect size estimator.

Knowledge of the proportion of markers without effect, p_0 and the effect sizes have an intrinsic value and is required for a wide variety of applications. Some applications are discussed in the Introduction. However, other extensions and applications of the proposed theoretical framework are conceivable that we are currently exploring in detail. An example involves the estimation of the individual effect sizes using the estimated average effect size

conditional on the number of markers with effect. In this application only the ML can be used. The QML cannot be used due to its different behavior. Furthermore, in case-control studies, population stratification can cause spurious associations between marker alleles and disease status when both disease prevalence and allele frequencies differ among subgroups. Using the principle of genomic control [24] and [25], our methods can be further adapted to obtain estimates of p_0 and Δ that take stratification into account.

Appendix

Derivation of the formula in (9):

In our article ([16]) we showed that an asymptotic approximation to Pearson's statistic on a $2 \times \upsilon$ contingency table is given by

$$Z_1^2 + ... + Z_\upsilon^2, \tag{15}$$

where $Z_1, ..., Z_\upsilon$ are independent normal random variables with

$$E(Z_i) = \mathbf{a}_i \mu \text{ and } Var(Z_i) = \lambda_i, \tag{16}$$

where

$$\mu^T = \left(\frac{(p_1 - q_1)\sqrt{\gamma\delta n}}{\sqrt{\gamma p_1 + \delta q_1}}, ..., \frac{(p_m - q_m)\sqrt{\gamma\delta n}}{\sqrt{\gamma p_m + \delta q_m}} \right),$$

and $\mathbf{a}_i$, $i = 1, ..., \upsilon$, are the unit length eigenvectors of matrix J with corresponding eigenvalues λ_i, and the entries of J are given by

$$J_{ij} = \begin{cases} -\frac{1}{\sqrt{p_i\gamma+q_i\delta}\sqrt{p_j\gamma+q_j\delta}} \left[\left(1 + \frac{(p_i-q_i)\delta}{2(p_i\gamma+q_i\delta)}\right) \left(1 + \frac{(p_j-q_j)\delta}{2(p_j\gamma+q_j\delta)}\right) \gamma q_i q_j + \right. \\ \left. \left(1 - \frac{(p_i-q_i)\gamma}{2(p_i\gamma+q_i\delta)}\right) \left(1 - \frac{(p_j-q_j)\gamma}{2(p_j\gamma+q_j\delta)}\right) \delta p_i p_j \right] & \text{if } i \neq j \\ \frac{1}{p_i\gamma+q_i\delta} \left[\left(1 + \frac{(p_i-q_i)\delta}{2(p_i\gamma+q_i\delta)}\right)^2 \gamma q_i (1-q_i) + \left(1 - \frac{(p_i-q_i)\gamma}{2(p_i\gamma+q_i\delta)}\right)^2 \delta p_i (1-p_i) \right] & \text{if } i = j. \end{cases} \tag{17}$$

It can be verified that under the usual null hypothesis, i.e. when $p_i = q_i$ for all $i = 1, ..., \upsilon$, the sum in (15) follows the chi-square distribution with $\upsilon - 1$ degree of freedom. In fact, under the null hypothesis $\mu = \mathbf{0}$ and J reduces to J^0 with $J^0_{ij} = -\sqrt{p_i p_j}$ if $i \neq j$ and $J^0_{ij} = 1 - p_i$ if $i = j$. Thus, J^0 has two eigenvalues, 1 with multiplicity $\upsilon - 1$, and 0 with multiplicity 1, which implies that $\upsilon - 1$ of the Z's in (15) has standard normal distribution and the remaining one is 0 with probability 1.

The approximation in (15) is the asymptotic equivalent ([16]), which is at least as accurate as the central chi-square with $\upsilon - 1$ degrees of freedom for approximating the distribution of Pearson's statistic under the null hypothesis. However, there are too many parameters involved in J. Therefore, we now give another approximation that involves just one parameter

and will subsequently be used to estimate p_0. Note that all small fractions in parentheses in (17) are close to 0. By deleting these fractions, we obtain matrix G whose entries are

$$G_{ij} = \begin{cases} -\frac{1}{\sqrt{p_i\gamma+q_i\delta}\sqrt{p_j\gamma+q_j\delta}}\left[\gamma q_i q_j + \delta p_i p_j\right] & \text{if } i \neq j \\ \frac{1}{p_i\gamma+q_i\delta}\left[\gamma q_i\left(1-q_i\right) + \delta p_i\left(1-p_i\right)\right] & \text{if } i = j. \end{cases} \tag{18}$$

We show that the approximation based on matrix G rather than J has the form given in (9) if $\gamma = \delta$. If $\gamma = \delta$, then the entries of G equal

$$G_{ij} = \begin{cases} -\frac{1}{\sqrt{p_i+q_i}\sqrt{p_j+q_j}}\left[q_i q_j + p_i p_j\right] & \text{if } i \neq j \\ \frac{1}{p_i+q_i}\left[q_i\left(1-q_i\right) + p_i\left(1-p_i\right)\right] & \text{if } i = j. \end{cases} \tag{19}$$

First we show that the eigenvalues of G are 1 with multiplicity $m-2$, 0 with multiplicity 1 and $2\sum_{i=1}^{m}\frac{p_i q_i}{p_i+q_i}$ with multiplicity 1. Matrix G can be written in the form

$$G = I - D\left(\mathbf{p}\mathbf{p}^T + \mathbf{q}\mathbf{q}^T\right)D,$$

where $D = diag\left(\frac{1}{\sqrt{p_i+q_i}}\right)$, $\mathbf{p} = (p_1, ..., p_\upsilon)^T$ and $\mathbf{q} = (q_1, ..., q_\upsilon)$. First we verify that vector $\mathbf{x} = D^{-1}\mathbf{1} = (\sqrt{p_1+q_1}, ..., \sqrt{p_m+q_m})$ is an eigenvector of G with eigenvalue 0 by

$$G\mathbf{x} = \mathbf{x} - D\left(\mathbf{p}\mathbf{p}^T + \mathbf{q}\mathbf{q}^T\right)\mathbf{1} = \mathbf{x} - D\left(\mathbf{p}\left(\mathbf{p}^T\mathbf{1}\right) + \mathbf{q}\left(\mathbf{q}^T\mathbf{1}\right)\right) = \mathbf{x} - D\left(\mathbf{p}+\mathbf{q}\right) = \mathbf{x} - \mathbf{x} = \mathbf{0},$$

where $\mathbf{1}$ is a υ-dimensional vector with components 1. Furthermore, if $\mathbf{x} = D^{-1}\mathbf{z}$, where $\mathbf{z}$ is orthogonal to both $\mathbf{p}$ and $\mathbf{q}$, i.e. $\mathbf{p}^T\mathbf{z} = \mathbf{q}^T\mathbf{z} = 0$, then $\mathbf{x}$ is an eigenvector of G with eigenvalue 1, because

$$G\mathbf{x} = \mathbf{x} - D\left(\mathbf{p}\mathbf{p}^T + \mathbf{q}\mathbf{q}^T\right)\mathbf{z} = \mathbf{x} - D\left(\mathbf{p}\mathbf{p}^T\mathbf{z} + \mathbf{q}\mathbf{q}^T\mathbf{z}\right) = \mathbf{x}.$$

Consequently, the eigenvectors of G with eigenvalue 1 span a $(\upsilon-2)$-dimensional eigenspace if $\mathbf{p} \neq \mathbf{q}$, and a $(\upsilon-1)$-dimensional eigenspace if $\mathbf{p} = \mathbf{q}$. Thus, if $\mathbf{p} = \mathbf{q}$, then there is no other eigenvector of G, and if $\mathbf{p} \neq \mathbf{q}$, then there is one more left, particularly $D\left(\mathbf{p}-\mathbf{q}\right)$. This is verified by

$$\left\{I - D\left(\mathbf{p}\mathbf{p}^T + \mathbf{q}\mathbf{q}^T\right)D\right\}D\left(\mathbf{p}-\mathbf{q}\right) = D\left\{\left(\mathbf{p}-\mathbf{q}\right) - \left(\mathbf{p}\mathbf{p}^T + \mathbf{q}\mathbf{q}^T\right)D^2\left(\mathbf{p}-\mathbf{q}\right)\right\} \overset{*}{=}$$

$$D\left\{\left(\mathbf{p}-\mathbf{q}\right) + \left(\mathbf{p}-\mathbf{q}\right)\mathbf{q}^T D^2\left(\mathbf{p}-\mathbf{q}\right)\right\} = D\left(\mathbf{p}-\mathbf{q}\right)\left\{1 + \mathbf{q}^T D^2\left(\mathbf{p}-\mathbf{q}\right)\right\}, \tag{20}$$

where at $*$ we used that $\mathbf{p}^T D^2\left(\mathbf{p}-\mathbf{q}\right) = -\mathbf{q}^T D^2\left(\mathbf{p}-\mathbf{q}\right)$, which holds because of

$$\mathbf{p}^T D^2\left(\mathbf{p}-\mathbf{q}\right) + \mathbf{q}^T D^2\left(\mathbf{p}-\mathbf{q}\right) = \left(\mathbf{p}+\mathbf{q}\right)^T D^2\left(\mathbf{p}-\mathbf{q}\right) = \mathbf{1}^T\left(\mathbf{p}-\mathbf{q}\right) = 0.$$

The equality in (20) also shows that the eigenvalue corresponding to eigenvector $D\left(\mathbf{p}-\mathbf{q}\right)$ is

$$1+\mathbf{q}^T D^2\left(\mathbf{p}-\mathbf{q}\right) = 1+\sum\nolimits_j \frac{q_j\left(p_j-q_j\right)}{p_j+q_j} = \sum\nolimits_j \frac{q_j\left(p_j+q_j\right)}{p_j+q_j} + \sum\nolimits_j \frac{q_j\left(p_j-q_j\right)}{p_j+q_j} = 2\sum\nolimits_j \frac{p_j q_j}{p_j+q_j}.$$

Finally, note that

$$D\left(\mathbf{p}-\mathbf{q}\right)=\left(\frac{p_1-q_1}{\sqrt{p_1+q_1}},...,\frac{p_\upsilon-q_\upsilon}{\sqrt{p_\upsilon+q_\upsilon}}\right)^T=\sqrt{\frac{2}{n}}\mu.$$

Since G is symmetric, all eigenvectors corresponding to eigenvalue 1 are orthogonal to eigenvector $D\left(\mathbf{p}-\mathbf{q}\right)$, and hence to μ.

The approximation to Pearson's statistic based on matrix G rather than J has the form

$$U_1^2+...+U_\upsilon^2,$$

where $U_1,...,U_\upsilon$ are independent normal random variables with $E\left(U_i\right)=\mathbf{b}_i\mu$ and $Var\left(U_i\right)=\varepsilon_i$, and $\mathbf{b}_i$, $i=1,...,\upsilon$, are the unit length eigenvectors of G with corresponding eigenvalues ε_i. As we have shown above the eigenvalues of G are $\varepsilon_1=...=\varepsilon_{\upsilon-2}=1$, $\varepsilon_{\upsilon-1}=2\sum_{j=1}^{\upsilon}\frac{p_jq_j}{p_j+q_j}$, $\varepsilon_\upsilon=0$ and $\mathbf{b}_1,...,\mathbf{b}_{\upsilon-2}$ are orthogonal to μ, hence $U_\upsilon=0$ and $U_1^2+...+U_{\upsilon-2}^2$ is a central chi-square random variable with $\upsilon-2$ degrees of freedom. Furthermore, since

$$\mathbf{b}_{\upsilon-1}^T=\frac{1}{\sqrt{\sum_{j=1}^{\upsilon}\frac{\left(p_j-q_j\right)^2}{p_j+q_j}}}\left(\frac{p_1-q_1}{\sqrt{p_1+q_1}},...,\frac{p_\upsilon-q_\upsilon}{\sqrt{p_\upsilon+q_\upsilon}}\right),$$

we have

$$E\left(U_{\upsilon-1}\right)=\mathbf{b}_{\upsilon-1}^T\mu=\sqrt{\frac{n}{2}\sum_{j=1}^{\upsilon}\frac{\left(p_j-q_j\right)^2}{p_j+q_j}}=\sqrt{n}\Delta\ ,$$

where

$$\Delta=\sqrt{\frac{1}{2}\sum\nolimits_{j=1}^{\upsilon}\frac{\left(p_j-q_j\right)^2}{p_j+q_j}}.$$

Since

$$1-\Delta^2=\frac{1}{2}\sum\nolimits_{j=1}^{\upsilon}\frac{\left(p_j+q_j\right)^2}{p_j+q_j}-\frac{1}{2}\sum\nolimits_{j=1}^{\upsilon}\frac{\left(p_j-q_j\right)^2}{p_j+q_j}=2\sum\nolimits_{j=1}^{\upsilon}\frac{p_jq_j}{p_j+q_j},$$

we have

$$Var\left(U_{\upsilon-1}\right)=1-\Delta^2.$$

We obtained that the approximation to Pearson's statistic based on G is the one given in (9) if $\gamma=\delta$.

Relation between the statistical association and correlation:

Now we prove that under the null hypothesis, the statistical association R^2 between two markers is C^2 if and only if the absolute value of the correlation between the two standard normal random variables whose squares approximate the corresponding Pearson's statistics on 2×2 contingency tables is C, particularly

Theorem 1 Let the contingency table at locus i be given as

$$\left[\begin{array}{cc} x_1^{(i)} & x_2^{(i)} \\ y_1^{(i)} & y_2^{(i)} \end{array} \right], \tag{21}$$

where $x_1^{(i)} + x_2^{(i)} = n\gamma$ and $y_1^{(i)} + y_2^{(i)} = n\delta$ for every i, and the "root" of Pearson's statistic at locus i is given as

$$P_n^{(i)} = \sqrt{n} \frac{\left(\delta x_1^{(i)} - \gamma y_1^{(i)}\right)}{\sqrt{\gamma\delta \left(x_1^{(i)} + y_1^{(i)}\right) \left(n - \left(x_1^{(i)} + y_1^{(i)}\right)\right)}}.$$

If the null hypothesis is true at both locus 1 with alleles A, a and locus 2 with alleles B, b, and the statistical association between locus 1 and 2 is C^2 ($C \geq 0$) then

$$\left(P_n^{(1)}, P_n^{(2)}\right) \xrightarrow{D} N\left(0, \left[\begin{array}{cc} 1 & C \\ C & 1 \end{array} \right]\right).$$

Proof. The statistical association between locus 1 and 2 is $r^2 / \left(p_1^{(1)} p_2^{(1)} p_1^{(2)} p_2^{(2)}\right)$ if and only if the haplotype probability table corresponding to this two biallelic loci has the form

$$\begin{array}{ccc} & A & a \\ B & p_1^{(1)} p_1^{(2)} + r & p_1^{(1)} p_2^{(2)} - r \\ b & p_2^{(1)} p_1^{(2)} - r & p_2^{(1)} p_2^{(2)} + r \end{array}. \tag{22}$$

The proof is based on the following theorem (delta method)

Theorem 2 Let (T_n), $n = 1, 2, ...,$ be a sequence of statistics such that

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow{D} N\left(0, \Sigma\right),$$

where $N\left(0, \Sigma\right)$ is multivariate normal distribution with 0 expectation and $m \times m$ covariance matrix Σ.

See [26] for the proof in the special case when g is a real-valued function. The proof of Theorem 2 is analogous.

Let $g : \mathbf{R}^k \to \mathbf{R}^m$ be a differentiable function. Then

$$\sqrt{n}\left(g\left(T_n\right) - g\left(\theta\right)\right) \xrightarrow{D} N\left(0, g'\left(\theta\right) \Sigma g'^T\left(\theta\right)\right).$$

Let T_n be defined as

$$T_n = \left(\frac{x_1^{(1)}}{\gamma n}, \frac{y_1^{(1)}}{\delta n}, \frac{x_1^{(2)}}{\gamma n}, \frac{y_1^{(2)}}{\delta n}\right).$$

The expected value of T_n is

$$\theta = \left(p_1^{(1)}, p_1^{(1)}, p_1^{(2)}, p_1^{(2)}\right),$$

because $Ex_1^{(i)} = n\gamma p_1^{(i)}$, $Ey_1^{(i)} = n\delta p_1^{(i)}$. From the central limit theorem we have that

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow{D} N\left(0, \Sigma\right),$$

where

$$\Sigma = \begin{bmatrix} \frac{1}{\gamma}p_1^{(1)}p_2^{(1)} & 0 & \frac{r}{\gamma} & 0 \\ 0 & \frac{1}{\delta}p_1^{(1)}p_2^{(1)} & 0 & \frac{r}{\delta} \\ \frac{r}{\gamma} & 0 & \frac{1}{\gamma}p_1^{(2)}p_2^{(2)} & 0 \\ 0 & \frac{r}{\delta} & 0 & \frac{1}{\delta}p_1^{(2)}p_2^{(2)} \end{bmatrix}.$$

Now define

$$g^*\left(u, v\right) = \frac{\sqrt{\gamma\delta}\left(u - v\right)}{\sqrt{\left(\gamma u + \delta v\right)\left(1 - \left(\gamma u + \delta v\right)\right)}}.$$

and

$$g\left(u_1, v_1, u_2, v_2\right) = \left(g^*\left(u_1, v_1\right), g^*\left(u_2, v_2\right)\right).$$

One can verify that

$$\sqrt{n}g\left(T_n\right) = \left(P_n^{(1)}, P_n^{(2)}\right) \text{ and } \sqrt{n}g\left(\theta\right) = \left(0, 0\right).$$

Since $g_u^{*\prime}\left(p_1, p_1\right) = \sqrt{\gamma\delta}/\sqrt{p_1p_2}$ and $g_v^{*\prime}\left(p_1, p_1\right) = -\sqrt{\gamma\delta}/\sqrt{p_1p_2}$, we have that

$$g'\left(\theta\right) = g'\left(p_1^{(1)}, p_1^{(1)}, p_1^{(2)}, p_1^{(2)}\right) =$$

$$\begin{bmatrix} \sqrt{\gamma\delta}/\sqrt{p_1^{(1)}p_2^{(1)}} & -\sqrt{\gamma\delta}/\sqrt{p_1^{(1)}p_2^{(1)}} & 0 & 0 \\ 0 & 0 & \sqrt{\gamma\delta}/\sqrt{p_1^{(2)}p_2^{(2)}} & -\sqrt{\gamma\delta}/\sqrt{p_1^{(2)}p_2^{(2)}} \end{bmatrix}.$$

One can verify that

$$g^{*\prime}\left(\theta\right)\Sigma g^{*\prime T}\left(\theta\right) = \begin{bmatrix} 1 & r/\sqrt{p_1^{(1)}p_2^{(1)}p_1^{(2)}p_2^{(2)}} \\ r/\sqrt{p_1^{(1)}p_2^{(1)}p_1^{(2)}p_2^{(2)}} & 1 \end{bmatrix},$$

and the statement of the theorem follows from Theorem 2.

References

[1] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet.* 2005; **6**:95-108.

[2] Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 2005; **6**:109-118.

[3] Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics.* 2003; **19**:1236-1242.

[4] Pounds S, Cheng C. Improving false discovery rate estimation. *Bioinformatics.* 2004; **20**:1737-1745.

[5] Dalmasso C, Broet P, Moreau T. A simple procedure for estimating the false discovery rate. *Bioinformatics.* 2005;21:660-668.

[6] Allison DB, Gadbury G, Heo M, Fernandez J, Lee C-K, Prolla TA, Weindruch R. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis.* 2002; **39**:1-20.

[7] Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B.* 2002; **64**:479-498.

[8] Turkheimer FE, Smith CB, Schmidt K. Estimation of the number of "true" null hypotheses in multivariate analysis of neuroimaging data. *Neuroimage.* 2001; **13**:920-930.

[9] Benjamini Y, Hochberg Y. On adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics.* 2000; **25**:60-83.

[10] Mosig MO, Lipkin E, Khutoreskaya G, Tchourzyna E, Soller M, Friedmann A. A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics.* 2001; **157**:1683-1698.

[11] Efron B, Tibshirani R, Storey JD, Tusher VG. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association.* 2001; **96**:1151-1160.

[12] Schweder T, Spjøtvoll E. Plots of p-values to evaluate many tests simultaneously. *Biometrika.* 1982;69:493-502.

[13] Hsueh H, Chen J, Kodell R. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J Biopharm Stat.* 2003; **13**:675-689.

[14] Chakravarti A. It's raining SNPs, hallelujah? *Nat Genet* 1998; **19**:216-217.

[15] Cardon LR, Bell JI. Association study designs for complex diseases. *Nat. Rev. Genet.* 2001; **2**:91-99.

[16] Bukszár J, Van den Oord EJCG. Accurate and efficient power calculations for $2 \times m$ tables in unmatched case-control designs. *Statistics in Medicine.* 2006; **25**:2632-2646.

[17] R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*; 2008.

[18] Cohen J. *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates Inc. Publishers: New Jersey, 1988; p 549.

[19] Agresti A. *Categorical Data Analysis.* New York, 1990; p 241.

[20] Weir BS. *Genetic data analysis II.* Sunderland, 1996.

[21] Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst.* 2004; **96**:434-442.

[22] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* 2004; **74**:106-120.

[23] Meinshausen N, Rice J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics.* 2006; **34**:373-393.

[24] Devlin B, Bacanu SA, Roeder K. Genomic Control to the extreme. *Nat Genet.* 2004; **36**:1129-1130.

[25] Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet.* 2004; **36**:512-517.

[26] Rao CR. *Linear statistical inference and its applications.* John Wiley & Sons: New York, London, Sydney, Toronto, 1993; pp 385-386.