Short communication

Optimal two-stage genome-wide association designs based on false discovery rate

Hansong Wang*, Daniel O. Stram

Department of Preventive Medicine, Division of Biostatistics and Genetic Epidemiology, Keck School of Medicine, University of Southern California, 1540 Alcazar Street, Suite 222, Los Angeles, CA 90033, USA

Received 24 April 2006; accepted 26 April 2006
Available online 19 May 2006

Abstract

Genome-wide association studies are likely to be conducted in large scale in the near future. In such studies, searching over hundreds of thousands of markers for the few ones that are associated with disease brings out the multiple-hypothesis testing problem in its severe form. We explore, in a two-stage design, how the use of false discovery rate (FDR) can alleviate the burden of a prohibitively strict significance level for single marker tests and still control the number of false positive findings, when there is more than one causal variant. FDR is the expected proportion of false positives among all significant findings. It can be approximated by \( (1 - p_0) \times (1 - \frac{1}{\text{power}}) \), where \( p_0 \) is the proportion of true causal markers, \( \alpha \) is the type I error rate and \( 1 - \beta \) the power of a two-stage study. When 500,000 SNPs are genotyped in the first stage with fixed SNP array and the most significant SNPs are genotyped in the second stage with standard but 20 times more expensive high-throughput techniques, up to 20% savings in the minimum genotyping cost is achieved for \( p_0 \) in the range of \( 1 \times 10^{-5} \) to \( 5 \times 10^{-4} \) and FDR in the range of 0.05 to 0.7, compared to when Bonferroni-corrected significance level is used. In terms of sample size, the saving is up to 60%. However, these savings come at a cost of more false positive findings.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Multiple comparison; False positive report probability; Whole-genome scan; Linkage disequilibrium (LD); Cost-efficient

1. Introduction

Genome-wide association is a promising approach for identifying common genetic variants underlying complex diseases (Hirschhorn and Daly, 2005). Enabled by recent developments in large-scale genotyping technology and enormous efforts in cataloging human genetic variation and LD patterns by the International HapMap Project, genome-wide association studies have been technologically possible only within the past few years (The International HapMap Consortium, 2005; Hirschhorn and Daly, 2005; Thomas et al., 2005) and there have already been some notable early successes (Klein et al., 2005). With this approach, a few hundred thousand well-chosen SNPs across the genome will be genotyped and tested, in an unbiased manner, for association with disease traits (Hirschhorn and Daly, 2005). This raises the need to control for false positive findings that arise merely by chance.

* Correspondence author. Tel.: +1 323 865 0408.
E-mail address: hansongw@usc.edu (H. Wang).
The traditional approach in multiple hypotheses testing to tackle this increased probability of making false discoveries has been to control the probability of making even one false discovery—the control of the family-wise error-rate (Benjamini et al., 2001). In the case of making \( m \) comparisons, for example, using Bonferroni correction requires that the significance level for each test is \( \alpha/m \) to ensure an overall error rate \( \leq \alpha \). Methods for controlling family-wise error rate impose an extremely small \( p \)-value for a single marker test to claim genome-wide significance and have been criticized for being conservative and not powerful enough to detect common variants with modest disease risk in complex diseases, especially when the markers are correlated (Benjamini et al., 2001; Lin, 2006; van den Oord and Sullivan, 2003). Some proposed to use permutation tests to assess empirical genome-wide significance level (Hirschhorn and Daly, 2005). Various ways of improving efficiency in permutation tests have also been developed (Dudbridge and Koeleman, 2004; Nyholt, 2004). However, most of these methods are not easily applied to design a study, since their statistical properties are dependent on empirical genome-wide LD structure.

Alternatively, under the assumption that there could be multiple true positive associations, one may consider controlling for false discovery rate (FDR) (Benjamini et al., 2001). FDR is the expected proportion of false positive features among all of those called significant (Storey and Tibshirani, 2003). When the number of tests \( m \) is large in practice, it is common to approximate FDR by the proportion of false tests among all that are expected to be significant (Storey and Tibshirani, 2003). FDR is determined by three factors: the proportion of true causal variants among all the markers \( p_0 \), the proportion of true causal variants to be discovered (PTD) and the significance level \( \alpha \) for each single test. If we consider PTD as the probability or power of detecting a single true causal gene \((1-\beta)\) and \( p_0 \) as the probability of being a true variant, then FDR is equivalent to the false positive report probability (FPRP) (Wacholder et al., 2004) and can be expressed as \((1-p_0)\alpha/(1-p_0)\alpha + p_0(1-\beta))\). It is advocated that FDR offers a good balance between false and true discoveries and has been applied in animal studies and micro-array analyses where expression levels of thousands of genes are measured to identify the ones that are differentially expressed (Benjamini et al., 2001; Storey and Tibshirani, 2003; van den Oord and Sullivan, 2003).

Even with the current genotyping technology, it is still expensive to carry out a large-scale genome-wide association study and most studies use a multistage design (Thomas et al., 2005) to increase efficiency. In a typical two-stage design, all the \( m \) markers are genotyped on a portion of study participants in the first stage and only the most promising markers are genotyped on the remaining subjects in the second stage. A well-designed two-stage study has been shown to save genotyping cost significantly, mainly because many non-causal markers have been eliminated in the first stage (Saito and Kamatani, 2002; Satagopan and Elston, 2003; Satagopan et al., 2002).

van den Oord and Sullivan (2003) devised a procedure to find the minimum cost for a two-stage study while controlling for FDR. But they treated the second stage as a replicate set of the first so that the statistics from the two stages are independent. Intuitively, pooling together all subjects at the end of the second stage would provide better power to detect true causal genes, which was verified by a recent simulation study (Skol et al., 2006). Moreover, with today’s genotyping availability, the per-genotype costs in the first stage and the second stage are dramatically different. Recently cost ratios appear to be in the range of 15–20, comparing the per-genotype costs of using the fixed SNP array in the first stage versus other high throughput genotyping platforms in the second stage. While the costs of both stage I and stage II genotyping technologies are rapidly dropping, it seems likely that a large cost differential will remain for the foreseeable future. In the original study and the corresponding software, van den Oord and Sullivan (2003) did not allow the per-genotype cost to vary between stages.

We previously reported on finding the optimal two-stage designs with specific power levels and per genotype cost ratios of the two stages while controlling the family-wise type I error rate (Wang et al., 2006). Here we use a similar method to investigate how much one can save on genotyping cost and sample size with a relaxed overall type I error rate by applying the FDR method, using the most recent per genotype costs and assuming a total of 500,000 SNP markers are genotyped in the first stage. We compare the results to those when Bonferroni correction is used.

2. Methods

Assume a population-based case-control study is conducted so that all cases and controls are unrelated. We choose a multiplicative genetic risk model and use \( \psi \) to denote the increased relative risk (odds ratio) of carrying each additional copy of the disease allele. We wish to test \( H_0: \psi = 1 \) against \( H_1: \psi \neq 1 \). Let \( S \) denote the statistic based on one case and \( K \) controls (we call it one case-control unit), where \( K \) is the control:case ratio. For example, \( S \) can be the difference of allele frequencies between the case and control subjects. Without loss of generality, \( S \) can be defined with variance
one and mean zero under $H_0$ and $\mu$ under $H_A$, where we call $\mu$ the “design effect” of a SNP. Then the statistic based on $n$ cases and $nK$ controls, $nS$, is asymptotically distributed as $N(0, n)$ under $H_0$ and $N(n\mu, n)$ under $H_A$. We use the same method as the program QUANTO (Gauderman, 2002; Longmate, 2002) to calculate $\mu$ for a given baseline disease probability and $\psi$ (Appendix). For a one-stage design with power 1-$\beta$ and type I error rate $\alpha$ (2-sided), $n$ is calculated as $(Z_{1-\alpha/2} + Z_{1-\beta})^2/\mu^2$, where $Z_\beta$ denotes the $\beta$th percentile of a standard normal distribution.

For a two-stage design, the full set of $m$ markers (SNPs) is genotyped on $n_1$ case-control units in the first stage. A stage I test of association is performed on all markers at a significance level $z_1$. The significant markers are then genotyped on an additional set of $n_2$ case-control units, and the stage II statistics based on all $(1 + K)(n_1 + n_2)$ subjects are evaluated at significance level $z_2$, to determine whether a SNP is associated with disease. The overall type I error rate $\alpha = \Pr((n_1S > c_1, (n_1 + n_2)S > c_2)_{H_0}$ is approximated by $2 \Pr(n_1S > c_1, (n_1 + n_2)S > c_2)_{H_0}$, because $\Pr(n_1S < c_1, (n_1 + n_2)S > c_2)_{H_0}$ and $\Pr(n_1S > c_1, (n_1 + n_2)S < c_2)_{H_0}$ are small compared to the overall $\alpha$, and power 1-$\beta$ is approximated by $\Pr(n_1S > c_1, (n_1 + n_2)S > c_2)_{H_A}$, where $c_1$ and $c_2$ are the critical values for the two stages that satisfy $z_1 = \Pr(n_1S > c_1)_{H_0}$, $1 - \beta_1 = \Pr(n_1S > c_1)_{H_A}$, $z_2 = \Pr((n_1 + n_2)S > c_2)_{H_0}$ and $1 - \beta_2 = \Pr((n_1 + n_2)S > c_2)_{H_A}$. Assume that there are $D$ causal variants and that each one is in complete LD with one of the markers. We consider the power to detect the smallest design effect $\mu$ as the power of the study.

Given the total number of causal variants $D$, study power 1-$\beta$ and FDR, $\alpha$ for a single SNP test is calculated as FDR(1- $\beta$) $p_0/(1-p_0)$ (1-FDR), where $p_0 = D/m$. The expected number of false positive findings is calculated as FDR(1-$\beta$)D/(1-FDR), where we treat the proportion of true variants that one wishes to discover as if it is the power of the study 1-$\beta$. For a study with $m = 500,000$, $p_0$ is typically small so that $\alpha$ increases approximately linearly with the number of true causal genes $D$ for given FDR and power. When Bonferroni correction is used, the type I error $\alpha_B = 0.05/m$.

The cost function of a two-stage design is $(1 + K)[t_1n_1m + t_2n_2(m - D)]x_1 + D(1 - \beta_1)]$, where $t_1$ and $t_2$ are the per genotype costs in the two stages. Our goal is to achieve specific power 1-$\beta$ and type I error rate $\alpha$ with the minimum expected cost. A program in R was designed to search for $x_1$ and 1-$\beta_1$ that give the minimum cost and make the empirical power and type I error less than 0.01% away from the nominal levels (for details, see (Wang et al., 2006)). We also search for two-stage studies that require the least sample size among all studies that cost less than 1.1 fold of the minimum expected cost—we call such designs “nearly optimal”. The “nearly optimal” design would be helpful when only a limited number of cases can be considered for participation, e.g., if the disease of interest is relatively rare or the recruiting cost is high. We also assume $m = 500,000$, $t_1 = 0.018$, $t_2/t_1 = 20$, $K = 1$ through this paper.

3. Results and discussion

The design parameters $x_1$, 1-$\beta_1$, $x_2$ and 1-$\beta_2$ that reach the minimum expected cost do not depend on $\mu$ (Wang et al., 2006), but do depend on the per genotype cost ratio of stage II vs. stage I, the overall power 1-$\beta$ and significance level $\alpha$, which is a function of FDR and $D$. For the purpose of our calculation, we fix the causal allele frequency $p = 0.2$ and odds ratio $\psi = 1.4$, corresponding to a design effect $\mu$ of 0.14. This can be viewed as the effect that is least possible to be detected among all the causal ones. Because our goal is to see how much the design parameters $x_1$, 1-$\beta_1$, $x_2$ and 1-$\beta_2$ and the minimum expected cost change as a result of a more liberal type I error rate than given by Bonferroni correction, we only consider FDR, 1-$\beta$ and $p_0$ that are realistic and also give $\alpha > 0.05/m$. This generally occurs when $D \geq 2$, 1-$\beta \geq 0.5$ and FDR $\geq 0.05$.

Figs. 1 and 2 plot, respectively, the relative reduction in minimum genotyping cost and sample size when $D$ is in (5, 250) and FDR in (0.05, 0.70), for optimal two-stage studies that achieve 90% and 50% power; in each case, total cost or sample size is compared, within the same power level, to those when $D = 1$ and $x_B$ is applied (indicated by the letter “B” on the plot). As expected, genotyping cost and sample size both decrease as $\alpha$ increases with either FDR or $D$. Maximum savings occur at the extreme values of $D = 250$ and FDR = 0.70 when power is 0.50; at this point, it is also worth noting that up to 292 false positive findings are expected to be picked up along with 125 true positive ones. The pattern and magnitude of decreasing in cost and sample size according to FDR and $D$ is similar for power levels 0.50 and 0.90, therefore it is most likely preserved for other power levels between 0.50 and 0.90. The relative reductions in sample size and cost, however, do not agree with each other. For example, at 1-$\beta$ = 0.90, FDR = 0.7 and $D = 250$, 17% savings in minimum genotyping cost ($1554 vs. $1863 K) result from applying the FDR method, while the corresponding saving in sample size is about 57% (1184 vs. 2775). This is because the proportion of sample size allocated to the first stage, $k_1 = n_1/(n_1 + n_2)$, is different for the two situations—$k_1$ is 0.33 when $\alpha_B$ is applied but is
0.65 for FDR = 0.7 and D = 250. To understand this somewhat unexpected result, we turn to the per-subject cost ratio of stage II vs. stage I, which we use CR to denote. From the Methods section, \( CR = (t_2/t_1)(1-p_0)z_1 + (1-\beta_1)p_0 \). Comparing the optimal design parameters as the overall significance level \( \alpha \) goes up due to application of FDR (Table 1), we note that the optimal \( z_1 \) increases for a more relaxed \( \alpha \), while the optimal \( 1-\beta_1 \) stays about the same within the same power level; thus as \( D \) and \( p_0 = D/m \) increase, \( CR \) tends to increase. Because it is then relatively more expensive to genotype a subject in stage II, the procedure assigns more subjects to the first stage, i.e. increase \( k_1 \), in order to get the minimum cost. However, because the absolute per-subject cost in stage I is still much higher than that in stage II, putting more subjects in stage I causes the minimum cost decrease not as fast as the reduction in sample size, as FDR and \( D \) increase. This finding makes the FDR method especially attractive when sample size is a more stringent constraint than monetary issues.

Tables 1 and 2 present the design parameters for optimal and nearly optimal two-stage studies when 90% and 50% of the true variants are expected to be discovered, i.e. \( 1-\beta = 0.90 \) and 0.50, respectively. In each case, the fraction of subjects optimally allocated to stage I \( k_1 \), the significance level \( z_1 \) and power \( 1-\beta_1 \) for moving a marker from stage I to stage II, the cost fraction of the optimal two-stage study relative to an equally powered one-stage study, and the power \( (1 - \beta_N) \) of a one-stage study that uses the sample size of the optimal two-stage study, are shown. When Bonferroni correction is applied at \( \alpha_B = 10^{-7} \), the optimal design parameters \( z_1, 1-\beta_1, x_2 \) and total sample size \( n_1 + n_2 \) do not depend on \( D \) over the range (1, 250); but the minimum cost does increase slightly as \( D \) goes up, which is also evident in Figs. 1 and 2 (indicated by the letter “B”). When the overall \( \alpha \) increases to 1800 fold of \( \alpha_B \) both \( z_1 \) and \( 1-\beta_1 \) increase. Specifically, \( z_1 \) increases up to 3-fold compared to that when \( \alpha_B \) is applied whereas \( 1-\beta_1 \) changes less radically. In
Fig. 2. Relative reduction in sample size of optimal two-stage design for different assumptions on study power, $D$ and FDR, compared to those when Bonferroni correction is used and $D = 1$ within the same power level. Per genotype cost ratio for stage II vs. stage I is 20. “B” stands for Bonferroni correction. Causal allele frequency is 0.2 and odds ratio is 1.4 with log-additive penetrance.

contrast, the significance level of the second stage, in which $(n_1 + n_2)$ subjects are combined for hypothesis testing, is highly related to the overall $\alpha$ rate—$\alpha_2$ is always slightly higher than $\alpha$, reflecting the fact that a marker has to pass the criteria of both stages to be claimed as significant of the combined study.

Although using a more liberal $\alpha$ significance level results in cost and sample size savings, the benefit of using a 2-stage design decreases as the overall $\alpha$ increases with FDR and $D$. Column 5 in Tables 1 and 2 displays the cost fraction of an optimal two-stage study relative to that of a one-stage study, both of which achieve the same power when rejected at the same adjusted $\alpha$ level. For example, when $1-\beta = 0.90$ and FDR = 0.5 (Table 1), for $D = 5$, $\alpha = 9.0 \times 10^{-6}$, the cost fraction is 59%; for $D = 100$, $\alpha = 1.8 \times 10^{-4}$, the cost fraction becomes 72%.

The nearly optimal design is a two-stage study that achieves the desired overall power and type I error rate but requires the minimum $(n_1 + n_2)$ among all studies that cost less than 110% of the minimum. Significance level and power of the first stage, $\alpha_1$ and $1-\beta_1$, of the nearly optimal design is both higher than those in the optimal design. As much as 17–24% reduction in sample size can be accrued when 10% more cost is spent for a carefully designed study, which would be helpful in situations when sample size too is a restricting concern.

As discussed in the previous paper (Wang et al., 2006), since the design parameters do not depend on $\mu$, Tables 1 and 2 provide general guidance for designing a two-stage study. First, one needs to calculate the $\alpha$ level for a single test from FDR, $D$ and $1-\beta$; then use a standard sample size calculation software to get total sample size $(n_1 + n_2)$ by putting in $\alpha$ level and $(1-\beta_N)$; $k_1$ is used next to figure out sample size of stage I. At the end of stage I, $\alpha_1$ is to be used for determining whether a SNP is genotyped in stage II; at the end of stage II, one compares $p$-value based on all
Table 1
Optimal and nearly optimal two-stage design parameters for $m = 500,000$, overall power $1-\beta = 0.90$, $t_2/t_1 = 20$ and various $D$ and FDR levels.

<table>
<thead>
<tr>
<th>$D$</th>
<th>FDR</th>
<th>$\alpha$ ($\times 10^{-7}$)</th>
<th>FP$^b$</th>
<th>CF$^c$</th>
<th>$(1-\beta_N)^d$</th>
<th>$k_1^e$</th>
<th>$x_1$ ($\times 10^{-3}$)$^f$</th>
<th>$1-\beta_1$</th>
<th>$x_2$ ($\times 10^{-7}$)$^f$</th>
<th>$1-\beta_2$</th>
<th>$x_2$ ($\times 10^{-7}$)$^f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B$^a$</td>
<td>1</td>
<td>0.05</td>
<td>0.470</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.447</td>
</tr>
<tr>
<td>5</td>
<td>B$^a$</td>
<td>4.7</td>
<td>0.05</td>
<td>0.471</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.447</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.526</td>
<td>0.977</td>
<td>0.373</td>
<td>4.0</td>
<td>0.908</td>
<td>17.2</td>
<td>0.917</td>
<td>0.503</td>
<td>5.6</td>
<td>0.954</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>1.13</td>
<td>0.549</td>
<td>0.975</td>
<td>0.390</td>
<td>4.4</td>
<td>0.908</td>
<td>39</td>
<td>0.916</td>
<td>0.527</td>
<td>6.1</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>4.5</td>
<td>0.593</td>
<td>0.970</td>
<td>0.427</td>
<td>5.1</td>
<td>0.909</td>
<td>154</td>
<td>0.913</td>
<td>0.578</td>
<td>6.9</td>
</tr>
<tr>
<td>10</td>
<td>B$^a$</td>
<td>1</td>
<td>0.05</td>
<td>0.471</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.447</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.525</td>
<td>0.978</td>
<td>0.368</td>
<td>4.0</td>
<td>0.907</td>
<td>17</td>
<td>0.917</td>
<td>0.505</td>
<td>5.3</td>
<td>0.954</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>1.046</td>
<td>0.975</td>
<td>0.389</td>
<td>4.2</td>
<td>0.908</td>
<td>35</td>
<td>0.916</td>
<td>0.523</td>
<td>6.1</td>
<td>0.955</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>2.25</td>
<td>0.571</td>
<td>0.973</td>
<td>0.407</td>
<td>4.7</td>
<td>0.908</td>
<td>78</td>
<td>0.914</td>
<td>0.551</td>
<td>6.5</td>
</tr>
<tr>
<td>10</td>
<td>B$^a$</td>
<td>1</td>
<td>0.05</td>
<td>0.471</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.449</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.544</td>
<td>0.977</td>
<td>0.382</td>
<td>4.3</td>
<td>0.907</td>
<td>33</td>
<td>0.917</td>
<td>0.505</td>
<td>5.3</td>
<td>0.954</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>2.04</td>
<td>0.567</td>
<td>0.974</td>
<td>0.403</td>
<td>4.6</td>
<td>0.908</td>
<td>70</td>
<td>0.914</td>
<td>0.523</td>
<td>6.1</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>4.5</td>
<td>0.594</td>
<td>0.971</td>
<td>0.425</td>
<td>5.1</td>
<td>0.909</td>
<td>156</td>
<td>0.913</td>
<td>0.578</td>
<td>6.9</td>
</tr>
<tr>
<td>10</td>
<td>B$^a$</td>
<td>1</td>
<td>0.05</td>
<td>0.472</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.449</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.573</td>
<td>0.974</td>
<td>0.407</td>
<td>4.6</td>
<td>0.908</td>
<td>82.8</td>
<td>0.914</td>
<td>0.550</td>
<td>6.9</td>
<td>0.956</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>5.098</td>
<td>0.971</td>
<td>0.426</td>
<td>5.3</td>
<td>0.909</td>
<td>172</td>
<td>0.912</td>
<td>0.577</td>
<td>7.5</td>
<td>0.957</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>11.3</td>
<td>0.628</td>
<td>0.966</td>
<td>0.457</td>
<td>5.7</td>
<td>0.910</td>
<td>383</td>
<td>0.910</td>
<td>0.547</td>
<td>6.9</td>
</tr>
<tr>
<td>10</td>
<td>B$^a$</td>
<td>1</td>
<td>0.05</td>
<td>0.473</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.449</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.598</td>
<td>0.971</td>
<td>0.429</td>
<td>5.0</td>
<td>0.909</td>
<td>164</td>
<td>0.912</td>
<td>0.577</td>
<td>7.4</td>
<td>0.957</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>10.0</td>
<td>0.625</td>
<td>0.969</td>
<td>0.448</td>
<td>5.6</td>
<td>0.909</td>
<td>347</td>
<td>0.911</td>
<td>0.612</td>
<td>7.6</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>22.5</td>
<td>0.658</td>
<td>0.964</td>
<td>0.482</td>
<td>6.2</td>
<td>0.910</td>
<td>763</td>
<td>0.910</td>
<td>0.647</td>
<td>8.8</td>
</tr>
<tr>
<td>10</td>
<td>B$^a$</td>
<td>1</td>
<td>0.05</td>
<td>0.472</td>
<td>0.982</td>
<td>0.329</td>
<td>3.3</td>
<td>0.906</td>
<td>1.7</td>
<td>0.921</td>
<td>0.447</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.598</td>
<td>0.971</td>
<td>0.429</td>
<td>5.0</td>
<td>0.909</td>
<td>164</td>
<td>0.912</td>
<td>0.577</td>
<td>7.4</td>
<td>0.957</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>10.0</td>
<td>0.625</td>
<td>0.969</td>
<td>0.448</td>
<td>5.6</td>
<td>0.909</td>
<td>347</td>
<td>0.911</td>
<td>0.612</td>
<td>7.6</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>22.5</td>
<td>0.658</td>
<td>0.964</td>
<td>0.482</td>
<td>6.2</td>
<td>0.910</td>
<td>763</td>
<td>0.910</td>
<td>0.647</td>
<td>8.8</td>
</tr>
</tbody>
</table>

$^a$Using Bonferroni correction to control the type I error rate $\alpha$.

$b$Number of accepted false positive findings, calculated as FDR(1-\(\beta\))/D/(1-FDR).

$c$Cost fraction of an optimally designed two-stage study relative to a one-stage study that achieves power $1-\beta$ when tested at significance level $\alpha$.

$d$Power of a one-stage study that uses the sample size of an optimal two-stage study and tests at significance level $\alpha$.

$e k_1 = n_1/(n_1+n_2)$.

$^f k_1 = n_1/(n_1+n_2)$.

Although there is some integration error involved, our method provides a practical way to find the minimum genotyping cost in a two-stage study while controlling false positive findings with FDR. This would be useful when one believes that there are multiple true disease genes and that the penalty of making some false positive findings is not so severe. A recent paper discussed similar calculations (Kraft, 2006) and our method is different from his in two aspects. First, we consider the power of discovering the SNP with the minimum design effect as the power of the study, i.e. the worst possible scenario, while he explicitly requested effects of all causal variants be input as parameters, resulting in different power levels for each variant in calculating FDR. Although it sounds more precise, it is not always possible to determine all the causal variants and their effects before hand. Second, he treated the second stage as a replicate set of the first stage, which, as discussed earlier, is not as powerful as pooling the two stages together.
Optimal and nearly optimal two-stage design parameters for \( m = 500,000 \), overall power \( 1-\beta = 0.50 \), \( t_2/t_1 = 20 \) and various \( D \) and FDR levels. Causal allele frequency is 0.2 and odds ratio is 1.4 with log-additive penetrance.

<table>
<thead>
<tr>
<th>( D )</th>
<th>( \alpha (\times 10^{-7}) )</th>
<th>FP(^b)</th>
<th>CF(^c)</th>
<th>((1-\beta))(^d)</th>
<th>( k_1)(^e)</th>
<th>( \chi_1^2 ) ((\times 10^{-3}))(^f)</th>
<th>( 1-\beta_1 ) ((\times 10^{-7}))</th>
<th>( \chi_2^2 ) ((\times 10^{-7}))</th>
<th>( 1-\beta_1 ) ((\times 10^{-7}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \begin{array}{c} 1 \end{array} )</td>
<td>0.05</td>
<td>0.39</td>
<td>0.846</td>
<td>0.231</td>
<td>2.7</td>
<td>0.521</td>
<td>3.3</td>
<td>0.615</td>
</tr>
<tr>
<td>5</td>
<td>( \begin{array}{c} 1 \end{array} )</td>
<td>0.05</td>
<td>0.39</td>
<td>0.846</td>
<td>0.231</td>
<td>2.7</td>
<td>0.521</td>
<td>3.3</td>
<td>0.615</td>
</tr>
<tr>
<td>10</td>
<td>( \begin{array}{c} 1 \end{array} )</td>
<td>0.05</td>
<td>0.39</td>
<td>0.846</td>
<td>0.231</td>
<td>2.7</td>
<td>0.521</td>
<td>3.3</td>
<td>0.615</td>
</tr>
<tr>
<td>20</td>
<td>( \begin{array}{c} 1 \end{array} )</td>
<td>0.05</td>
<td>0.39</td>
<td>0.846</td>
<td>0.231</td>
<td>2.7</td>
<td>0.521</td>
<td>3.3</td>
<td>0.615</td>
</tr>
<tr>
<td>50</td>
<td>( \begin{array}{c} 1 \end{array} )</td>
<td>0.05</td>
<td>0.39</td>
<td>0.846</td>
<td>0.231</td>
<td>2.7</td>
<td>0.521</td>
<td>3.3</td>
<td>0.615</td>
</tr>
<tr>
<td>100</td>
<td>( \begin{array}{c} 1 \end{array} )</td>
<td>0.05</td>
<td>0.39</td>
<td>0.846</td>
<td>0.231</td>
<td>2.7</td>
<td>0.521</td>
<td>3.3</td>
<td>0.615</td>
</tr>
</tbody>
</table>

\(^a\) Using Bonferroni correction to control the overall type I error rate \( \alpha \).
\(^b\) Number of accepted false positive findings, calculated as FDR\((1-\beta)D/(1-FDR)\).
\(^c\) Cost fraction of an optimally designed two-stage study relative to a one-stage study that achieves power \( 1-\beta \) when tested at significance level \( \alpha \).
\(^d\) Power of a one-stage study that uses the sample size of an optimal two-stage study and tests at significance level \( \alpha \).
\(^e\) \( k_1 = n_1/(n_1+n_2) \).
\(^f\) \( \chi_1^2 \): significance level of stage S.

One reason that Bonferroni correction is criticized often in a genome-wide study is that it does not give the correct \( \alpha \) level when the markers are correlated. When the markers are highly correlated, Lin (2006) discussed a Monte Carlo procedure to assess empirical power in two-stage studies, with given sample size \( n_1 \) and \( n_2 \) and critical value of stage I (analogous to \( c_1 \) in our method). With his program, to design a two-stage study with the minimum cost, one would simulate the genotype data to be tested, choose sample sizes in both stages and the critical value of the first stage test statistic and use his program to find the study power for these choices of parameters, then search through possible sample sizes and critical values of stage I for a study design that returns the minimum cost and also achieves the nominal power and type I error rate. This is helpful but could also be cumbersome in the design stage.

Skol et al. (2006) considered situations when the design effects \( \mu \) of the causal gene are different between the two stages, which may happen in studies where the \( n_2 \) subjects are sampled from a different population than the \( n_1 \) subjects. Our program can be modified to accommodate these situations of heterogeneity. However, when the design effects are believed to vary between stages, the optimal \( \chi_1^2 \) and \( 1-\beta_1 \) will depend on the risk allele frequency and odds ratio.
Acknowledgements

This work has been supported by Grants CA63464, Genetic Susceptibility to Cancer in Multiethnic Cohorts; GM58897, Computational Methods in Genetic Epidemiology; HG002790, Center for Excellence in Genomic Sciences, University of Southern California; 5P30 ES07048, Southern California Environmental Health Sciences Center; and P01 CA 17054-27A2, Unresolved Public Health Issues Related to HT and Cancer.

Appendix Calculating \( \mu \)

Let \( Y, G \) and \( \delta \) denote the binary disease status, the observed genotype and the observed causal allele dosage for a subject. Then \( \delta = 0, 1 \) and 2 corresponding to \( G = aa, Aa \) and \( AA \), respectively, where \( A \) is the causal allele. Choose a logistic disease model and a multiplicative genetic risk, \( \Pr (Y = 1|\delta) = (e^{\gamma_0+\gamma_1\delta}/(1 + e^{\gamma_0+\gamma_1\delta}) \). The baseline disease probability \( \Pr (Y = 1|\delta = 0) \) is expressed as \( e^{\gamma_0}/(1 + e^{\gamma_0}) \) and the odds ratio \( \psi \) associated with carrying each additional copy of \( A \) is \( e^{\gamma_1} \). The likelihood from \( n \) cases and \( nK \) controls is

\[
L(\gamma_0, \gamma_1) = \prod_{i=1}^{n} \frac{e^{\gamma_0+\gamma_1\delta_i}}{1 + e^{\gamma_0+\gamma_1\delta_i}} \prod_{j=1}^{nK} \frac{1}{1 + e^{\gamma_0+\gamma_1\delta_j}}.
\]

For an unmatched case-control study, the conditional distribution of \( G \) in cases and controls, \( \Pr (G|Y) \), can be calculated using Bayes formula with given baseline risk (thus \( \gamma_0 \)) and \( \psi \) (thus \( \gamma_1 \)). Given the “expected” distribution of \( G \) in cases and controls, one can treat them as real data and form the likelihood ratio test (LRT) on \( H_0: \gamma_1 = 0 \) vs. \( H_A: \gamma_1 \neq 0 \). Denote the natural logarithm of maximum likelihood from one case-control unit (one case and \( K \) controls) under \( H_0 \) and \( H_A \) as \( \hat{\gamma}_0 \) and \( \hat{\gamma}_A \), respectively. Let \( A = 2(\hat{\gamma}_A - \hat{\gamma}_0) \), then \( nA \) is the non-centrality parameter of the \( \chi^2 \) statistic for \( n \) cases and \( nK \) controls under \( H_A \) (Gauderman, 2002; Longmate, 2002). The number of cases required for a study with power 1-\( \beta \) when rejected at significance level \( \alpha \) (2-sided) is \( n = (Z_{1-\beta/2} + Z_{1-\beta})^2/\Delta \), where \( Z_\beta \) denotes the \( p \)th percentile of a standard normal distribution.

Let \( S \) denote the test statistic based on one case-control unit with unit variance, expected value 0 and \( \mu \) under \( H_0 \) and \( H_A \), respectively. Note that the exact form of \( S \) is not important. Then the statistic based on \( n \) cases and \( nK \) controls, \( nS \), is approximately distributed as \( N(0, n) \) and \( N(n\mu, n) \) under \( H_0 \) and \( H_A \). The statistic \( (nS)^2/n \) is approximately \( \chi^2 \), with non-centrality parameter \( n\mu^2 \) under \( H_A \). We equate \( \mu^2 \) to \( \Delta \) because a \( \chi^2 \) statistic from the same data should have the same non-centrality parameter. Simulation studies showed that sample size given by this method achieved accurate empirical power for a wide range of baseline risk, \( \psi \) and type I error when common methods are used in analysis. These include logistic regression and Pearson’s \( \chi^2 \) test on allelic association for \( 2 \times 2 \) tables (results not shown). From \( \mu \) we obtain the distribution of \( S, n_1S \) and \( (n_1 + n_2)S \) and the power function of a two-stage study.

References


