DOI: 10.1534/genetics.105.049536

On the Sample Size Requirement in Genetic Association Tests When the Proportion of False Positives Is Controlled

Guohua Zou*,† and Yijun Zuo*,[1]

*Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824 and †Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

Manuscript received August 14, 2005
Accepted for publication September 21, 2005

ABSTRACT

With respect to the multiple-tests problem, recently an increasing amount of attention has been paid to control the false discovery rate (FDR), the positive false discovery rate (pFDR), and the proportion of false positives (PFP). The new approaches are generally believed to be more powerful than the classical Bonferroni one. This article focuses on the PFP approach. It demonstrates via examples in genetic association studies that the Bonferroni procedure can be more powerful than the PFP-control one and also shows the intrinsic connection between controlling the PFP and controlling the overall type I error rate. Since controlling the PFP does not necessarily lead to a desired power level, this article addresses the design issue and recommends the sample sizes that can attain the desired power levels when the PFP is controlled. The results in this article also provide rough guidance for the sample sizes to achieve the desired power levels when the FDR and especially the pFDR are controlled.

FOR multiple tests, the classical approach is to control the overall type I error rate [*i.e.*, the familywise error rate (FWER)]. Bonferroni correction is often used to this end. The method, however, often leads to a very stringent significance level for each test. As a remedial measure, the false discovery rate (FDR) was thus introduced recently and controlled in many investigations. Benjamini and Hochberg (1995) define the (unconditional) FDR as

$$\mathrm{FDR} = E\left(\frac{V}{R}\middle| R>0\right)\Pr(R>0),$$

where R is the number of positives, and V is the number of false positives. In practice, FDR will be controlled by α if the hypotheses corresponding to the k smallest P-values are rejected, where k is the largest j such that the jth smallest P-value $p_{(j)} \le j\alpha/m$, and m is the number of the hypotheses tested. Some numerical results show that controlling FDR can lead to higher powers (see, for example, Benjamini and Hochberg 1995; Sabatti *et al.* 2003). However, it should be pointed out that these power comparisons are not based on the same FWER level and hence do not necessarily imply that using FDR is more powerful than using Bonferroni correction at the same overall type I error level, because controlling the former has a different meaning from controlling the latter. Recently, the positive (or conditional) false discovery rate (pFDR) was discussed by Storey (2002) with

$$\mathrm{pFDR} = E\left(\frac{V}{R}\middle| R>0\right).$$

Storey (2002) also argues that it is more suitable than FDR in practice. Another related concept, the proportion of false positives (PFP), is suggested by Southey and Fernando (1998) and Fernando *et al.* (2004). They define the PFP as

$$\mathrm{PFP} = \frac{E(V)}{E(R)}.$$

These authors show through simulation studies that PFP is often close to FDR and pFDR. In general, PFP is closer to pFDR than to FDR. In fact, Storey (2003) shows that pFDR = PFP when the tests are independent and follow a mixture distribution. Clearly, when most null hypotheses are true, the discrepancy between FDR and either pFDR or PFP will increase. The connections and differences of these various measures have been discussed in Zaykin *et al.* (2000), Storey (2003), and Fernando *et al.* (2004). In practice, Benjamini and Hochberg's FDR is the easiest to estimate whereas pFDR is most difficult to control (although very important). Noting that PFP is closer to pFDR, we focus on PFP below.

In genetic association studies, it is usually very likely that a marker tested is not associated with the disease of interest; that is, most null hypotheses are true. This

[1]*Corresponding author:* Department of Statistics and Probability, Michigan State University, A440 Wells Hall, East Lansing, MI 48824. E-mail: zuo@msu.edu

means that a very stringent significance level may be sometimes appropriate in these settings. Otherwise, the false positive rate will be very high, a situation the applied researchers often want to avoid. On the basis of this observation, we present some examples that compare the power of the classical FWER controlling procedure with the PFP controlling one when the levels of FWER and PFP are set to be the same. It turns out that the Bonferroni approach can outperform the PFP approach in the aforementioned settings. Furthermore, we demonstrate that for a specified problem, controlling PFP is in fact approximately equivalent to Bonferroni correction by setting corresponding (different) levels of PFP and FWER. In this regard, Westfall *et al.* (1997) consider the conditions under which the Bonferroni correction behaves as the posterior probability (*i.e.*, essentially pFDR or PFP).

On the other hand, if we control only the PFP level, then a low power may result; that is, controlling only PFP is not sufficient to achieve a desired power level in multiple tests. An intuitive explanation is that PFP considers only the tests that are rejected. When the power is low, most (often even all) true alternatives will not be rejected. Therefore, simultaneous consideration of both PFP and power is very relevant and important. This article considers mainly the design issue for multiple case-control association tests when PFP is controlled. The sample sizes that can lead to the desired PFP and power levels are recommended.

METHODS

Genetic model and power calculation: We first give the formula for calculating the power under various genetic models (*cf.*, Zou and Zhao 2004). We consider two alleles, A and a, at a candidate marker, whose frequencies are p and $q = 1 - p$ in the population, respectively. For simplicity, we consider a case-control study with n cases and n controls. Let X_i denote the number of alleles A carried by the ith individual in the case group, and Y_i is defined similarly for the ith individual in the control group. Assuming Hardy-Weinberg equilibrium, each X_i and Y_i has a value of 2, 1, 0 with respective probabilities p^2, $2pq$, and q^2 under the null hypothesis of no association between the candidate marker and disease. When the candidate marker is associated with disease, we assume that the penetrances are f_2 for genotype AA, f_1 for genotype Aa, and f_0 for genotype aa. Without loss of generality, we let $f_2 \geq f_1 \geq f_0$. Note that the two alleles may be true functional alleles or may be in linkage disequilibrium with true functional alleles. Under this genetic model, the probabilities of having k copies of A among the cases, $m_k = P(X_i = k)$, and among the controls, $m'_k = P(Y_i = k)$, are

$$m_0 = \frac{q^2 f_0}{p^2 f_2 + 2pq f_1 + q^2 f_0},$$
$$m_1 = \frac{2pq f_1}{p^2 f_2 + 2pq f_1 + q^2 f_0},$$
$$m_2 = \frac{p^2 f_2}{p^2 f_2 + 2pq f_1 + q^2 f_0},$$
$$m'_0 = \frac{q^2(1-f_0)}{p^2(1-f_2) + 2pq(1-f_1) + q^2(1-f_0)},$$
$$m'_1 = \frac{2pq(1-f_1)}{p^2(1-f_2) + 2pq(1-f_1) + q^2(1-f_0)},$$
$$m'_2 = \frac{p^2(1-f_2)}{p^2(1-f_2) + 2pq(1-f_1) + q^2(1-f_0)}.$$

Let n_A and n_U denote the observed numbers of allele A in the case group and the control group and p_A and p_U denote the population frequencies of allele A in these two groups, respectively. Then the statistic to test the association between the candidate marker and disease is

$$t = \frac{(n_A - n_U)/(2n)}{\sqrt{\hat{p}(1-\hat{p})/n}},$$

where $\hat{p} = (n_A + n_U)/(4n)$.

Consider a one-sided test and use a significance level of α. The power of the test statistic t is

$$1 - \beta = \Phi\left(\frac{-z_\alpha\sqrt{\tilde{p}(1-\tilde{p})} + \sqrt{n}\mu}{\sigma}\right), \qquad (1)$$

where Φ is the cumulative standard normal distribution function, z_α is the upper 100αth percentile of the standard normal distribution, $\tilde{p} = \mu/2 + m'_2 + m'_1/2$ is the expected frequency of allele A under the genetic model, μ is the expected difference of estimated allele frequencies between cases and controls, which is given by

$$\mu = m_2 + \tfrac{1}{2}m_1 - m'_2 - \tfrac{1}{2}m'_1,$$

and σ^2/n is the corresponding variance with σ^2 being given by

$$\sigma^2 = \tfrac{1}{4}[4m_2 + m_1 - (2m_2 + m_1)^2 + 4m'_2 + m'_1 - (2m'_2 + m'_1)^2].$$

Thus, assuming M markers are tested, when Bonferroni correction is used, the power that a disease-associated marker is detected is given by

$$\Phi\left(\frac{-z_{\alpha/M}\sqrt{\tilde{p}(1-\tilde{p})} + \sqrt{n}\mu}{\sigma}\right).$$

Approximate equivalence of controlling FWER and PFP: In the genetic association test between gene and disease, for the case of testing M markers, Southey and Fernando (1998) define the PFP as

$$\mathrm{PFP} = \frac{\sum_{i=1}^{M}\alpha_i \Pr(H_0^{(i)})}{\sum_{i=1}^{M}[\alpha_i \Pr(H_0^{(i)}) + (1-\beta_i)\Pr(H_1^{(i)})]}, \qquad (2)$$

TABLE 1

The power comparison based on Bonferroni correction and controlling PFP for $M = 1000$ tests

		PFP		
	Bonferroni correction	$K = 1$	$K = 2$	$K = 5$
$p_A - p_U = 5\%$				
$p = 0.05$	0.626 (0.148)	0.588 (0.053)	0.652 (0.080)	0.732 (0.131)
$p = 0.2$	0.125 (0.018)	0.030 (0.000)	0.051 (0.000)	0.096 (0.001)
$p = 0.7$	0.085 (0.011)	0.009 (0.000)	0.018 (0.000)	0.044 (0.000)
$p_A - p_U = 10\%$				
$p = 0.05$	0.999 (0.760)	0.999 (0.745)	0.999 (0.791)	1.000 (0.847)
$p = 0.2$	0.882 (0.283)	0.878 (0.175)	0.909 (0.233)	0.941 (0.327)
$p = 0.7$	0.892 (0.268)	0.890 (0.149)	0.919 (0.209)	0.950 (0.308)

The overall significance level and PFP level are $\alpha = \gamma = 0.05$. The sample size is $n = 500$, and the values in parentheses correspond to the case of $n = 200$.

where α_i is the significance level, $1 - \beta_i$ is the power at the ith marker whose calculation formula for different genetic models is provided in Equation 1, and $\Pr(H_0^{(i)}) \cdot [\Pr(H_1^{(i)})]$ is the prior probability of the null (alternative) hypothesis being true for the ith test. Fernando *et al.* (2004) show that if for each test the PFP is controlled by γ, then for all M tests it is controlled still by γ.

We now show that for a specified scenario, *i.e.*, for a given number K of truly disease-associated markers, genetic models, and population allele frequencies, controlling FWER is substantially equivalent to controlling PFP. In fact, for any given PFP level of γ, we can find an overall significance level of α such that this PFP level can be obtained: First, we choose the type I error rate α_i for the ith test, which satisfies

$$\frac{\alpha_i \Pr(H_0^{(i)})}{\alpha_i \Pr(H_0^{(i)}) + (1 - \beta_i)\Pr(H_1^{(i)})} = \gamma \qquad (3)$$

(or Equation 2). Then we take $\alpha = 1 - \prod_{i=1}^{M}(1 - \alpha_i)$. Clearly, such a choice of the overall significance level (correspondingly, the significance level for ith test is taken as α_i, which can be regarded as a generalized Bonferroni correction) will lead to the PFP level of γ. On the other hand, for any given overall significance level of α, we can find a PFP level of γ such that this overall significance level is attained: First, we calculate

$$\frac{\alpha_i \Pr(H_0^{(i)})}{\alpha_i \Pr(H_0^{(i)}) + (1 - \beta_i)\Pr(H_1^{(i)})} = \gamma_i$$

with $\alpha_i = \alpha/M$. Then we take $\gamma = M/\sum_{i=1}^{M}(1/\gamma_i)$ [here we assume that the prior probability $\Pr(H_0^{(i)}) = 1 - K/M$]. Obviously, $0 < \gamma < 1$. Such a choice of the PFP level (correspondingly, the PFP level for ith test is taken as γ_i) can lead to the overall type I error rate of α.

Note that in practical genetic studies, the true number of disease-associated markers and the genetic models are unknown to us; the equivalence should be approximate by using their estimates.

RESULTS

We have seen that controlling FWER is in fact equivalent to controlling PFP by setting their different levels. Here we provide examples that show that controlling FWER can lead to higher powers than controlling PFP, even though their levels are set to be the same.

For comparing the powers based on Bonferroni correction and based on controlling PFP, we assume that the overall significance level α for M tests is the same as the PFP level γ ($=\alpha$) for these tests and consider $M = 1000$. Then for each test, the significance level $\alpha_i = \alpha/M$ if we use Bonferroni correction, and PFP $\gamma_i = \gamma = \alpha$ if we control PFP. The power results are summarized in Table 1, where the true number of disease-associated markers K is set to be 1, 2, and 5, and the prior probability is assumed to be $\Pr(H_0^{(i)}) = 1 - K/M$.

From Table 1, we observe that for a small allele frequency difference (5%) between the case and control groups, using Bonferroni correction for multiple tests often leads to larger power, especially for the case of small sample size (say, 200) and small number of disease-associated markers. This is true even for larger allele frequency difference when the sample size is smaller. On the other hand, controlling PFP can lead to higher power for large allele frequency difference (10%) between the cases and controls, especially for large sample sizes (1000) and a large number of disease-associated markers. This is true for smaller allele frequency difference when the sample size is larger. Note that the number of disease-associated markers is unknown in practical genetic association studies. Thus, overall, to detect disease markers with small allele frequency difference (such as 5%), using

TABLE 2

Sample sizes to attain the desired PFP level of 0.05 and power of 80% for various allele frequency differences and population allele frequencies under four genetic models

	$p = 0.05$	$p = 0.2$	$p = 0.7$
$p_A - p_U = 3\%$			
Dominant	1878	5233	6303
Recessive	1949	5308	6371
Multiplicative	1884	5252	6354
Additive	1882	5250	6352
$p_A - p_U = 5\%$			
Dominant	767	1941	2212
Recessive	809	1979	2249
Multiplicative	772	1952	2239
Additive	770	1949	2237
$p_A - p_U = 7\%$			
Dominant	436	1017	1113
Recessive	466	1048	1123
Multiplicative	440	1025	1117
Additive	438	1023	1115
$p_A - p_U = 10\%$			
Dominant	245	516	561
Recessive	266	538	531
Multiplicative	248	523	528
Additive	246	521	527

Bonferroni correction for sample sizes that are not large (such as 500 and 200) can give a higher level of power; to detect disease markers with large allele frequency difference (such as 10%), using PFP for sample sizes that are not small (such as 1000 and 500) will lead to a better result in power.

To calculate the sample size for attaining a desired PFP level γ and a desired power level $1 - \beta$ that a disease-associated marker is detected, we assume that the prior probability of the marker tested being truly disease associated is 0.0001 and use Equation 3. We find that the sample size required depends on the genetic model and population allele frequency substantially through the allele frequency difference $p_A - p_U$ between the cases and controls and population allele frequency p (the results for the PFP level of 0.05 under dominant, recessive, multiplicative, and additive models with $f_0 = 0.01$ are presented in Table 2; a similar conclusion for the power of the two-stage design can be found in Y. Zuo, G. Zou and H. Zhao, unpublished results). On the basis of this, we consider a recessive genetic model and let $f_0 = 0.01$ in our calculation for simplicity when the PFP levels of 0.20 and 0.50 are used. Tables 2–4 present the sample sizes to attain the PFP levels $\gamma = 0.05$, 0.20, and 0.50 and the power level $1 - \beta = 80\%$ for various allele frequency differences between the cases and controls and population allele frequencies, respectively. Interestingly, it can be seen from Tables 2–4 that reducing the level of PFP will not necessarily lead to a great increase in sample size required. Therefore, we can use the sample sizes derived by setting PFP at a small level. This will not significantly increase the cost of the experiment.

TABLE 3

Sample sizes to attain the desired PFP level of 0.20 and power of 80% for various allele frequency differences and population allele frequencies

	$p = 0.05$	$p = 0.2$	$p = 0.7$
$p_A - p_U = 3\%$	1706	4639	5565
$p_A - p_U = 5\%$	709	1730	1965
$p_A - p_U = 7\%$	409	916	981
$p_A - p_U = 10\%$	233	471	464

TABLE 4

Sample sizes to attain the desired PFP level of 0.50 and power of 80% for various allele frequency differences and population allele frequencies

	$p = 0.05$	$p = 0.2$	$p = 0.7$
$p_A - p_U = 3\%$	1488	4040	4400
$p_A - p_U = 5\%$	619	1507	1711
$p_A - p_U = 7\%$	357	799	854
$p_A - p_U = 10\%$	204	410	404

DISCUSSION

In multiple tests, there is an increasing trend to use FDR, pFDR, and PFP as measures of global error instead of using overall type I error rate. This article gives the examples on the power comparison between controlling FWER and PFP when their levels are set to be the same (as is usually done in the literature), which show that using Bonferroni correction does not necessarily lead to a lower power. This article also shows that controlling FWER and controlling PFP, seemingly two different approaches, based on prior and posterior probabilities, respectively, are actually intrinsically equivalent.

Note that controlling only PFP does not necessarily lead to a desired level of power. We work out the sample size to attain the desired power that a disease-associated marker is detected under various population allele frequencies and various allele frequency differences between the cases and controls when PFP is controlled. Our results reveal that lowering the PFP level alone will not give rise to much increase in sample size required to attain a desired power level. Therefore, taking a small PFP level may be appropriate in general in multiple case-control association tests. Further, as we have seen, FDR and especially pFDR are often close to PFP. Combining this and the above fact that the effect of PFP level on the sample sizes required is not large, we see that the

sample sizes we obtained should be useful when FDR and especially pFDR are controlled.

Finally, we remark that the sample size calculation in this article is done for unrelated individual data. We note that family-based data are also often used in genetic epidemiological studies. The design issue for such data, which is not pursued here, is no doubt an interesting topic when PFP (or FDR or pFDR) is controlled.

The authors are grateful to the associate editor Rebecca Doerge and to the two reviewers for their constructive comments and suggestions that led to substantial improvements of the original manuscript. This work is supported in part by grants DMS0234078 from the National Science Foundation (to Y. Zuo) and 70221001 and 10471043 from National Natural Science Foundation of China (to G. Zou).

LITERATURE CITED

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

Fernando, R. L., D. Nettleton, B. R. Southey, J. C. Dekkers, M. F. Rothschild *et al.*, 2004 Controlling the proportion of false positives in multiple dependent tests. Genetics **166:** 611–619.

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

Southey, B. R., and R. L. Fernando, 1998 Controlling the proportion of false positives among significant results in QTL detection. Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia, Vol. 26, pp. 221–224.

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

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

Westfall, P. H., W. O. Johnson and J. M. Utts, 1997 A Bayesian perspective on the Bonferroni adjustment. Biometrika **84:** 419–427.

Zaykin, D. V., S. S. Young and P. H. Westfall, 2000 Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller *et al.* Genetics **154:** 1917–1918.

Zou, G., and H. Zhao, 2004 The impacts of errors in individual genotyping and DNA pooling on association studies. Genet. Epidemiol. **26:** 1–10.

Communicating editor: R. W. Doerge