*Am. J. Hum. Genet. 74:582–584, 2004*

## Multiple Comparisons in Studies of Gene × Gene and Gene × Environment Interaction

*To the Editor:*

$(d \log h_{it}/d \log W_{it}) \mid_{\lambda_{it}} = 0$

Complex diseases are (by definition) influenced by multiple genes, environmental factors, and their interactions. There is currently a strong interest in studies testing for association between combinations of these factors and disease, in part because genes that affect the risk of disease only in the presence of another genetic variant or particular environment may not be detected in a marginal (gene-by-gene) analysis (Culverhouse et al. 2002). Such studies raise the problem of multiple comparisons. Even when a small number of candidate genes and environmental factors is examined, a large number of possible interactions may need to be tested, as illustrated by a recent article in *The American Journal of Human Genetics* (Bugawan et al. 2003).

Bugawan et al. (2003) investigated potential interaction between the IL4R locus and five tightly linked SNPs in the IL4 and IL13 loci on chromosome 5, through use of a sample of 90 patients with type I diabetes and 94 population-based controls. They independently tested each of the chromosome 5 SNPs for interaction with IL4R, through use of logistic regression (cf. their table 7), and corrected for multiple comparisons through use of a permutation procedure. They concluded that there is statistically significant evidence for an epistatic interaction between at least one of the chromosome 5 SNPs and the IL4R locus. However, the authors' permutation procedure does not have the desired statistical property—that is, it rejects the global null hypothesis of no interaction too often when none of the estimated interaction parameters differ from their null value. In this letter, I discuss why their procedure fails, present several alternatives, and compare the performance of these alternatives in a small simulation study.

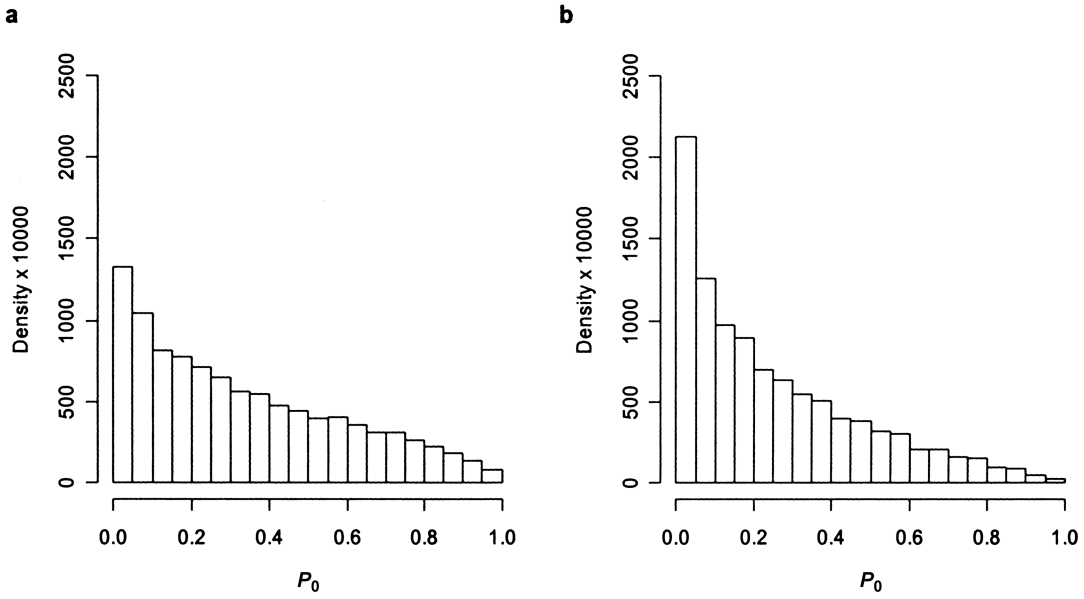


**Figure 1** Density of global $p$ values for the multiple-comparisons procedure used by Bugawan et al. (2003) under the global null hypothesis for two independent tests (*a*) and three independent tests (*b*). In panel a, $P_0 \equiv F_0(P_{(1)},P_{(2)})$, where $P_1$ and $P_2$ are independently uniform on (0,1) and $F_0$ is the cumulative distribution function of the order statistics, as discussed in the text. In panel b, $P_0 \equiv F_0(P_{(1)},P_{(2)},P_{(3)})$, where $P_1$, $P_2$, and $P_3$ are independently uniform on (0,1). Densities are estimated from 10,000 Monte Carlo replicates.

The procedure presented by Bugawan et al. (2003) amounts to plugging the order statistics for the observed $p$ values, $p_{(1)},\ldots,p_{(5)}$, into their joint cumulative distribution function under the null: $p = F_0(p_{(1)},\ldots,p_{(5)}) = \Pr(P_{(1)} \leq p_{(1)},\ldots,P_{(5)} \leq p_{(5)})$. (Here, italicized uppercase letters refer to random variables, and lowercase letters refer to observed values of the corresponding variables. This differs from the notation in the Bugawan et al. [2003] article.) The authors estimate $F_0$ by permuting case-control labels 200 times and calculating the ordered $p$ values for each permutation.

A simple example shows that this approach is inappropriate. Consider the $p$ values from two independent tests, $P_1$ and $P_2$. If we assume a large enough sample size, $P_1$ and $P_2$ are independently uniform on (0,1) under the null, and, hence, the cumulative distribution function for the associated order statistics, $F_0(p_{(1)},p_{(2)})$, is $P_{(1)}(2p_{(2)} - p_{(1)})$ (Bickel and Doksum 1977). The distribution of $P = F_0(P_{(1)},P_{(2)})$ under the global null is shown in figure 1*a*. $P$ does not have a uniform distribution under the null, as we expect for a $p$ value. In this case, a test that rejects the global null hypothesis that both tests are null when $P < .05$ would have a type I error rate between 10% and 15%. As shown in figure 1*b*, the magnitude of the type I error rate increases as the number of independent tests increases.

There are several alternative, theoretically justified and simple procedures that correct for multiple comparisons, besides the notoriously conservative Bonferroni correction. Simes's test (Simes 1986), for example, controls the overall significance level (also known as the "familywise error rate") when the tests are independent or exhibit a special type of dependence (Sarkar 1998). Simes's test rejects the global null hypothesis that all $K$ test-specific null hypotheses are true if $p_{(k)} \leq \alpha k/K$ for any $k$ in $1,\ldots,K$. Simulation results reported in table 1 suggest that Simes's test has the appropriate false-positive rate, even when the tests are correlated.

Other approaches with particular appeal in the context of multiple-gene and multiple-environmental-factor studies aim to control the false-discovery rate—that is, the expected proportion of rejected null hypotheses that are falsely rejected. This approach is particularly useful when a portion of the null hypotheses can be assumed false, as in microarray studies. Devlin et al. (2003) recently proposed a variant of the Benjamini and Hochberg (1995) step-up procedure that controls the false-discovery rate when testing a large number of possible gene × gene interactions in multilocus association studies. The Benjamini and Hochberg procedure is related to Simes's test; setting $k^* = \max k$ such that $p(k) \leq \alpha k/K$, it rejects all $k^*$ null hypotheses corresponding to $p_{(1)},\ldots,p_{(k^*)}$. In fact, the Benjamini and Hochberg procedure reduces to Simes's test when all null hypotheses are true (Benjamini and Yekutieli 2001).

**Table 1**

**Observed False-Positive Rates (False-Discovery Rates) for Procedures with Nominal 5% Rates in the Context of Testing Five Possible Gene × Gene Interactions, Calculated from 500 Simulated Data Sets**

| | False-Positive Rate under Model | |
|---|---|---|
| Procedure[a] | Null I | Null II |
| CDF | .194 | .214 |
| Simes | .032 | .036 |
| RSimes | .048 | .058 |
| | False-Discovery Rate under Model | |
| | Null I | Null II |
| BHD | .014 | .014 |
| DRW | .050 | .070 |

Note.—Six SNPs were simulated for 100 cases and 100 controls. The first SNP had mutant-allele frequency of .2; the other five SNPs were generated independently of the first by sampling five-SNP haplotypes with frequencies similar to those given in table 5 of Bugawan et al. (2003). Under model Null I, none of the SNPs were associated with disease. Under Null II, each mutant allele for the first SNP doubles disease risk, but the remaining five SNPs are not associated with disease. The multiple-comparisons procedures are applied to the $p$ values from five Wald tests for interaction based on the logistic model $\Pr(\text{disease}) = \alpha + \beta_1\text{SNP}_1 + \beta\text{SNP}_i + \beta_{\text{int}}\text{SNP}_1^*\text{SNP}_i$, analogous to that of Bugawan et al. (2003).

[a] "CDF" denotes the cumulative distribution function procedure used by Bugawan et al. (2003); "Simes" is the standard Simes's test; "RSimes" is Simes's test applied to $p$ values calculated by comparing the observed $p$ values to the distribution of $p$ values generated by permuting the outcome variable 200 times; "BHD" is the Benjamini and Hochberg step-up procedure corrected for general dependency (Benjamini and Yekutieli 2001) (the usual step-up procedure is identical to Simes's test in this case); and "DRW" is the related procedure proposed by Devlin et al. (2003).

Devlin et al.'s (2003) proof for the validity of their false-discovery-rate procedure requires that the analyzed genes be statistically independent. This is not the case for the IL4 and IL13 SNPs studied by Bugawan et al. (2003), but the simulation results in table 1 suggest that Devlin et al.'s (2003) procedure controls the false-discovery rate even when the analyzed genes are correlated.

The $p$ values reported in table 7 of Bugawan et al. (2003) do not lead to any significant results at the .05 level when any of the alternative procedures discussed here are used.

Clearly, effective methods are needed for adjusting for multiple comparisons when testing for association between multiple factors and complex disease. On the one hand, blithely reporting any results marginally "significant" at the .05 level or relying on outdated and ill-performing stepwise model-building procedures (see, e.g., Burnham and Anderson [2002] and Devlin et al. [2003]) will lead to spurious results, expensive follow-up studies with little chance of replication, and confusion. On the other hand, overly conservative procedures will create missed opportunities. Although the proce-

dures discussed here are known to control the familywise error rate or false-discovery rate in particular situations (e.g., independent covariates), their performance in more general situations needs further investigation.

PETER KRAFT
*Departments of Epidemiology and Biostatistics*
*Harvard School of Public Health*
*Boston*

## References

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

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

Bickel PJ, Doksum KA (1977) Mathematical statistics: basic ideas and selected topics. Prentice Hall, Englewood Cliffs, New Jersey

Bugawan TL, Mirel DB, Valdes AM, Panelo A, Pozzili P, Erlich HA (2003) Association and interaction of the IL4R, IL4, and IL13 loci with Type 1 diabetes among Filipinos. Am J Hum Genet 72:1505–1514

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70:461–471

Devlin B, Roeder K, Wasserman L (2003) Analysis of multilocus models of association. Genet Epidemiol 25:36–47

Sarkar S (1998) Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. Ann Stat 26:494–504

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Address for correspondence and reprints: Dr. Peter Kraft, 665 Huntington Avenue, SBuilding 2, Room 109, Boston, MA 02115. E-mail: pkraft@hsph.harvard.edu


0002-9297/2004/7403-0027$15.00

---

*Am. J. Hum. Genet. 74:584–585, 2004*

## Reply to Kraft

*To the Editor:*

Our study (Bugawan et al. 2003) reported a negative association of a specific IL4-524 haplotype with type 1 diabetes (T1D), consistent with a previous report (Mirel et al. 2002), and presented evidence for a genetic interaction between IL4-524 and IL4R SNPs. To test the latter, we computed relevant *P* values by permuting multilocus genotypes separately in case and control groups.

The criticism raised by Kraft (2004 [in this issue]) is not directed at our implementation of permutation testing, per se, but at permutation testing in general. His argument is that permutation testing does not properly account for multiple comparisons, resulting in an increase in false claims of significance, or type I familywise error (FWE). In the place of permutation testing, Kraft advocates the use of the Simes method—an elaboration of the classic Bonferroni procedure. In response, we wish to show that permutation testing can be used to obtain a desired false-positive error rate (as, indeed, can be demonstrated using Kraft's example) and, moreover, that such an approach has the added advantage of providing additional protection against false claims of nonsignificance, or type II error.

It should be noted that permutation methods are well established as a robust approach for obtaining overall significance levels while minimizing type II error (e.g., Good 1994; Doerge and Churchill 1996; Lynch and Walsh 1998), that such methods are extensible to multiple-testing scenarios (Westfall and Young 1993), and that examples of their application to human genetics are not uncommon (e.g., Lewis et al. 2003). However, as with any statistical method, the validity is dependent on correct application. Kraft provides an analysis of the permutation testing by discussing the distribution of two *P* values obtained from hypothetically permuted distributions (i.e., independent and uniformly distributed under the null hypothesis). The joint cumulative distribution function (CDF) for these two *P* values is given as $F(P_{(1)},P_{(2)}) = P_{(1)}(2P_{(2)} - P_{(1)})$, where $P_{(1)}$ and $P_{(2)}$ are, respectively, the first- and second-ordered *P* values. As such, Kraft notes that the $\Pr(P < .05)$ for this joint distribution is ~0.1, indicating that we would expect to see the smaller *P* value, or $P_{(1)} < .05$, about 10% of the time. Kraft's argument, therefore, is that for independent tests, use of a critical value of .05 leads to a type I error rate of 10%.

In fact, the proper approach for permutation testing—adjusted or unadjusted for multiple comparisons—is to find the critical value corresponding to the desired type I error rate. Specifically, if we consider the simulations presented by Kraft as equivalent to the result of a permutation test, we would seek the value of $\boldsymbol{x}$ in the permuted distribution for which $\Pr(P < \boldsymbol{x})$ is actually $\leqslant\alpha$ and would use that value, not the .05 value as Kraft appears to suggest. For $P_{(1)}$, this critical value would be .0253, as can be shown either by simulation or by solving Kraft's joint CDF for $\alpha = 0.05$, given $P_{(2)} = 1$ (in effect, solving the marginal CDF for $P_{(1)}$). It is interesting to note that the first *P* value that Kraft gives (.10) corresponds to the Sidak multiple comparison–adjusted *P* value for observed $\alpha = 0.05$ and $k = 2$ tests, whereas