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.

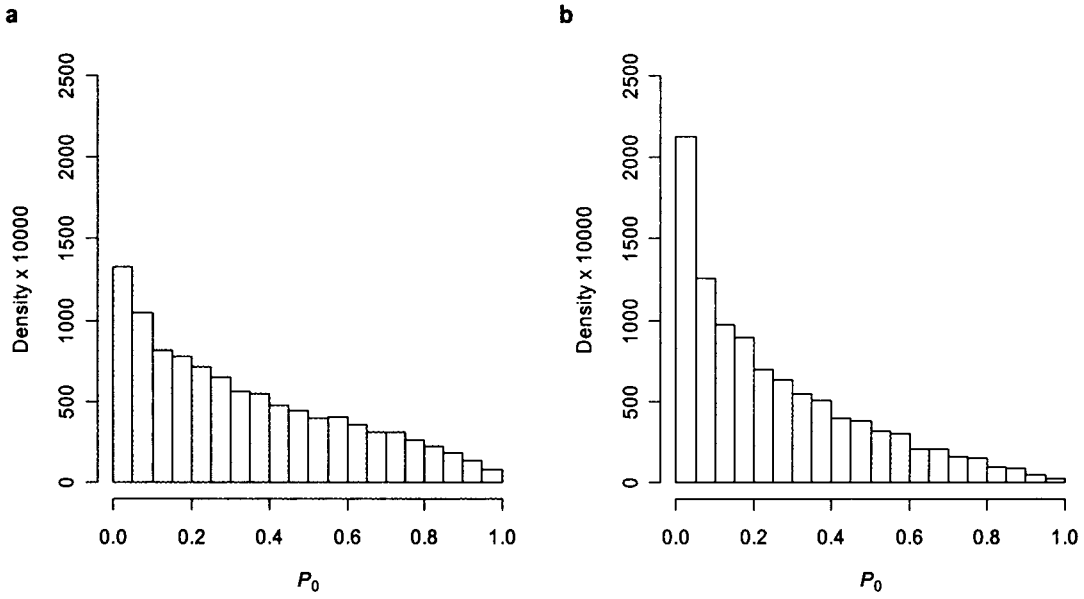

Figure 1 Density of global p values for the multiple-comparisons procedure used by Bugawan et al. (2003) under the global null hypothesis for two independent tests (*a*) and three independent tests (*b*). In panel a, $P_0 \equiv F_0(P_{(1)},P_{(2)})$, where P_1 and P_2 are independently uniform on (0,1) 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

the value we give corresponds to the Sidak-adjusted threshold ($1 - [1 - \alpha]^{1/k}$). As such, this example nicely illustrates that permutation testing, for two independent tests, yields familiar and contextually appropriate results.

It should also be noted that multiple-testing methods that rely on raw Bonferroni-type inequalities fail to incorporate correlation structures between tests. Therefore, although such methods (e.g., Simes 1986; Hochberg 1988; Rom 1990) provide control of FWE, they nevertheless are expected to be less powerful than methods that account for such dependencies. Indeed, these methods may be made more precise through resampling-based approaches (Westfall and Young 1993). In particular, the data from which the tests in table 7 (Bugawan et al. 2003) were derived are strongly correlated, and, therefore, tests that assume independence are not expected to be the most powerful. Moreover, Kraft fails to take into account the nonindependence of genotype distributions between chromosome 5 and chromosome 16 SNPs presented in table 6 (Bugawan et al. 2003). Applying the Simes correction suggested by the author for 10 comparisons (two sets: patients and controls, and five SNPs), the independence between IL4-524 and IL4R patient genotypes would be rejected with $P < .01$, supporting our conclusion of an interaction between chromosome 5 and chromosome 16 in T1D susceptibility.

In conclusion, what is needed, from a methodological perspective, are statistical procedures that adequately protect against false claims of significance while simultaneously addressing the correlated nature of multiple testing. The various methods discussed by Kraft address the former but do not address the latter. Having said this, whatever the statistical approach, the strongest test of the significance of any reported genetic interaction lies neither in initial-discovery *P* values nor in biologic plausibility—which we believe is high in this case—but in the ability to reproduce observations in independent cohorts.

Ana Maria Valdes, Brian Rhees, and Henry Erlich

Roche Molecular Systems
Alameda, CA

References

Bugawan TL, Mirel DB, Valdes AM, Panelo A, Pozzilli 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

Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142: 285–294

Good P (1994) Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer-Verlag, New York

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802

Kraft P (2004) Multiple comparisons in studies of gene × gene, gene × environment interaction. Am J Hum Genet 74: 582–584 (in this issue)

Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I, Williams NM, et al (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: schizophrenia. Am J Hum Genet 73:34–48

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associated, Sunderland, MA, pp 441–442

Mirel DB, Valdes AM, Lazzeroni LC, Reynolds RL, Erlich HA, Noble JA (2002) Association of IL4R haplotypes with type 1 diabetes. Diabetes 51:3336–3341

Rom DM (1990) A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 77:663–665

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

Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for *P*-value adjustment. John Wiley & Sons, New York

Address for correspondence and reprints: Dr. Brian K. Rhees, Roche Molecular Systems, Department of Human Genetics, 1145 Atlantic Avenue, Alameda, CA 94501. E-mail: brian.rhees@roche.com

0002-9297/2004/7403-0028$15.00

Am. J. Hum. Genet. 74:585–588, 2004

Revisiting the Clinical Validity of Multiplex Genetic Testing in Complex Diseases

To the Editor:

The usefulness of genetic testing to identify high-risk patients for common multifactorial diseases is subject to debate. Optimism about the public health opportunities is counterbalanced with skepticism, since genetic factors appear to play a role in only a minority of patients with complex diseases, the number of genes involved is large, and their penetrance is incomplete (Holtzman and Marteau 2000; Vineis et al. 2001).

In last March's issue of the *Journal,* Yang and colleagues addressed the question of whether prediction of disease is improved by multiplex genetic testing (Yang et al. 2003). At first sight, their results seem promising. In a simulation study, they considered five genetic tests (g_1–g_5), which each could have a positive ($g_i = 1$) or negative result ($g_i = 0$). Yang et al. used the likelihood ratio to indicate the magnitude of change in disease probability before and after genetic testing. Positive test

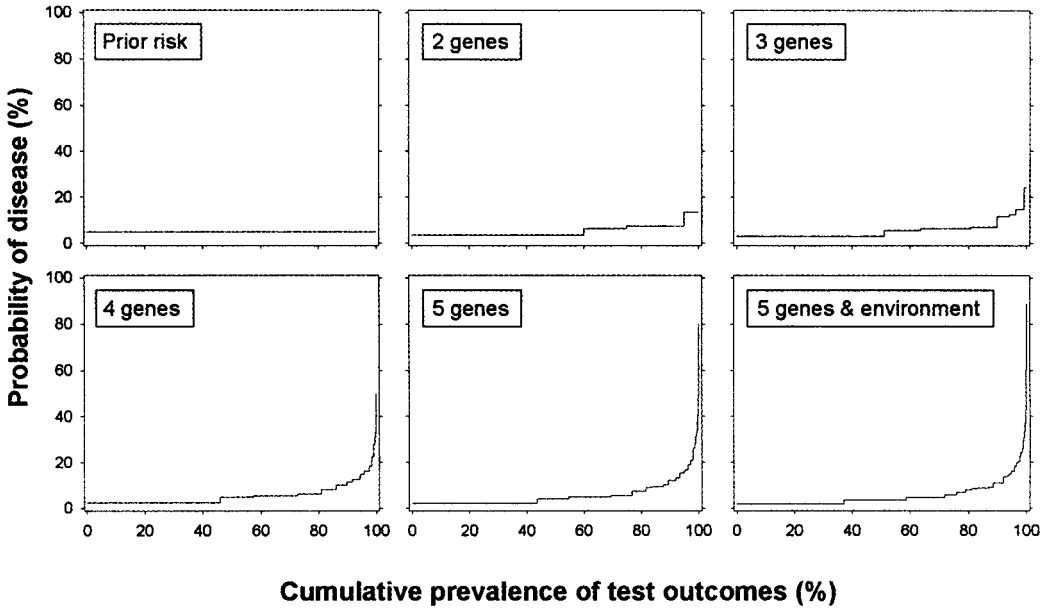

Figure 1 Probability of disease before and after testing for multiple genes and environmental exposure. The two-gene test has 4 (2^2) possible test results, the three-gene test has 8 (2^3) results, and so on. The posterior probability of disease for each combination of test results is obtained from the regression equations in table 1 of Yang et al. (2003).The prevalence of each combination is calculated by multiplying the probabilities of positive (p) and negative ($1 - p$) test results of each single test. For example, for the two-gene test we calculate that 60% ([1−0.25] × [1−0.20] × 100) of the individuals will have negative results on both tests and 15% ([1−0.25] × 0.20 × 100) will have a negative result on test 1 and a positive result on test 2. To facilitate presentation of all results, a cumulative prevalence (X-*axis*) was calculated, which was obtained by summing the prevalences after ranking the outcomes on their posterior probability.

results have a likelihood ratio >1, which means that the posterior disease probability is higher than the prior probability. Negative test results have a likelihood ratio <1. The combined likelihood ratio of several *independent* test results can be obtained by multiplying their individual likelihood ratios. Using these principles, Yang et al. showed that combining information on five genetic factors and one environmental exposure in one multiplex test may increase a 5% baseline risk to 88.9%, which was considerably higher than the posterior probabilities obtained by testing for the single genes (7.8%–16.4%). In addition, they demonstrated using empirical data from a study on deep venous thrombosis that the posterior probability of venous thrombosis was substantially higher when three genes, factor V Leiden, G20210A prothrombine, and protein C deficiency, were considered simultaneously (61.6%), rather than each gene alone (1.2%–3.1%). These estimates are correct, but they do not demonstrate the clinical validity of multiplex genetic testing, as the authors concluded. There are four reasons for this.

First, Yang et al. based their conclusion on only one outcome of the composite test—that is, the combination of positive results on all individual tests. Although Yang et al. acknowledged in their discussion that this concerns only a small proportion of the population, they did not quantify the size of the proportion. From multiplication of the prevalences of the test results, we calculate that the 18-fold increase in probability of disease in the simulated data was found in 0.0006% (6 per million) of all subjects and the 100-fold increase in the risk of venous thrombosis in only 0.0004% (4 per million). This low prevalence of high-risk combinations of genes may limit the clinical usefulness of genetic testing.

The second point is related to this issue. Yang et al. presented disease probabilities for subjects who had positive results on all single tests, but they did not report the probabilities for subjects who had combinations of both positive and negative results. The posterior probabilities and prevalences of all test result combinations are presented in figure 1. This figure demonstrates that the probabilities that Yang et al. had reported are the highest points in each of the graphs. Although these probabilities increase when genes are added, the probabilities of all other test result combinations do not rise accordingly. This is explained by the fact that positive results on each single test increase the combined likelihood ratio. This implies that the posterior probabilities reported by Yang et al. increase *by definition* when tests are added. In all other combinations with one or more negative test results, the likelihood ratios of negative results on the single tests will decrease the overall likelihood ratio. For the majority of subjects, the benefits of multiplex genetic testing in terms of the difference

between the prior and posterior probability are less profound.

A third point is that each genetic test that was added by Yang et al. was a stronger predictor of disease than those already considered in the multiplex test. The relative risks of the positive test results increased from 1.5 to 3.5, with likelihood ratios ranging from 1.6 to 3.7. This implies that the increase in the likelihood ratio of the composite test results may not only be due to the addition of tests but probably also to their higher predictive values. If the likelihood ratio of each single test had been 1.7, similar to the first test, then the combined likelihood ratio for subjects who had positive results on all five tests would have been 14.2, much lower than the 77.6 reported by Yang et al. This demonstrates that the substantial increase in the likelihood ratio was largely explained by the increasing predictive value of the single genes. In general, the added value of expanding a multiplex test will depend on the predictive value of each individual genetic test.

The fourth point concerns the most important conclusion of the authors that multiplex genetic testing has the potential to improve the clinical validity of predictive testing for common multifactorial diseases. This conclusion was based on the substantial increase in the probability of disease of individuals who had positive results on all single tests. However, the clinical validity of a test does not depend on the posterior probability for a few subjects, but on its ability to discriminate between the probability of disease in subjects who will develop the disease and those who will not. The discriminative ability of a test is commonly evaluated by its sensitivity and specificity. The sensitivity of a test is the percentage of positive test results among subjects who will develop the disease, and the specificity is the percentage of negative test results among subjects who will *not* develop the disease. On a perfect, or "gold-standard," test, all subjects who will develop the disease have a positive test result (sensitivity = 1), and all subjects who will not develop the disease have a negative result (specificity = 1). For composite tests, positive and negative results are defined by a cutoff value of the disease probability. The sensitivity and specificity of a composite test may differ, depending on the cutoff probability that is chosen. Therefore, the sensitivity and specificity are calculated for each possible cutoff value of the probability and plotted in a so-called receiver-operating–characteristic (ROC) curve (Hanley and McNeil 1982). The area under the ROC curve (AUC) indicates the discriminative ability of a composite test. The discriminative ability is perfect if the AUC is 1, whereas an AUC of 0.50 indicates a total lack of discrimination (Hanley and McNeil 1982). If one is interested in whether genetic tests can improve the accuracy of prediction above and beyond certain minimum levels of sensitivity or specificity, one may also consider analyses of a partial AUC (e.g., Thompson and Zucchini 1989). The ROC curves for the composite tests considered by Yang et al. are presented in figure 2. The total AUC increases from 0.59 for the two-gene test to 0.70 for the five-gene test, which means that adding genes improves the discriminative ability of the multiplex genetic test. Also here, one may question whether this increase was due to the addition of genes or to their increasing predictive values. To examine this, we considered the relative risks in equal steps from 1.5 to 1.7, rather than from 1.5 to 3.5, which is more realistic for genetic factors in common diseases. With these lower relative risks, the AUC of the two-gene test was 0.57 and that of the five-gene test was 0.61. This difference between the AUCs was smaller than that obtained from the data from Yang et al., which implies that also the increase in the discriminative ability of their multiplex tests is largely explained by the increasing predictive value of the added tests.

What can we learn from the ROC curve about the clinical validity of genetic testing? The aim of genetic screening is often to select high-risk subjects for preventive treatment or intensified surveillance programs. For this purpose, the sensitivity of the test should be high so that most (future) patients are identified by a positive test result. A high specificity of the test is desired to increase the efficiency of screening, because then the number of subjects who are unnecessarily selected for preventive interventions is minimized. From figure 2 it follows that a sensitivity of 0.80, which means that still 20% of the patients are missed by the screening program, is accompanied by a specificity of 0.45. The latter

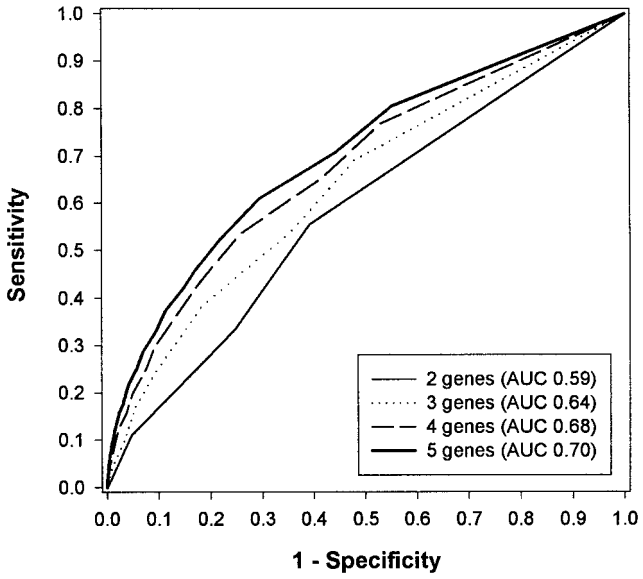

Figure 2 ROC curves for the multiplex genetic tests of Yang et al. (2003).

means that 55% of all subjects who will not develop the disease will be classified falsely. In a population in which 95% of the individuals will not develop the disease, as in the study of Yang et al., this means that 52% will undergo unnecessary preventive treatment. When a sensitivity of 0.90 is chosen, the percentage of all subjects who are unnecessarily selected is 73%. In comparison, the sensitivity and specificity of mammography in a large population–based breast cancer screening program were 0.75 and 0.92, respectively (Carney et al. 2003). Thus, the multiplex genetic tests of Yang et al. are by no means efficient screening strategies.

In conclusion, the clinical usefulness of genetic testing should be evaluated by ROC analysis. Using this approach for the data of Yang et al., we found that the discriminative ability of the multiplex genetic test increased by the addition of more genes but that its performance for use as a screening instrument was rather inefficient. It remains to be investigated whether these results are representative of the prediction of common disease by multiplex genetic tests that include genetic factors with low mutation prevalence and low relative risks. In that case, alternative statistical strategies are needed to increase the potential clinical application of selective genetic testing.

Acknowledgments

The study was financially supported by the Netherlands Organization for Scientific Research (NWO Pioneer and ZonMW; grant number 945-10-039) and the Center for Medical Systems Biology (CMSB).

A. Cecile J. W. Janssens,[1] M. Carolina Pardo,[2] Ewout W. Steyerberg,[1] and Cornelia M. van Duijn[2]

[1]*Department of Public Health and* [2]*Department of Epidemiology and Biostatistics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands*

References

Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, Geller BM, Abraham LA, Taplin SH, Dignan M, Cutter G, Ballard-Barbash R (2003) Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 138:168–175

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36

Holtzman NA, Marteau TM (2000) Will genetics revolutionize medicine? N Engl J Med 343:141–144

Thompson ML, Zucchini W (1989) On the statistical analysis of ROC curves. Stat Med 8:1277–1290

Vineis P, Schulte P, McMichael AJ (2001) Misconceptions about the use of genetic tests in populations. Lancet 357: 709–712

Yang Q, Khoury MJ, Botto L, Friedman JM, Flanders WD (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. Am J Hum Genet 72:636–649

Address for correspondence and reprints: Dr. Cecile Janssens, Center for Clinical Decision Sciences, Department of Public Health, Erasmus MC, P.O. Box 1738, 3000 DR The Netherlands. E-mail: a.janssens@erasmusmc.nl

0002-9297/2004/7403-0029$15.00

Am. J. Hum. Genet. 74:588–589, 2004

Revisiting the Clinical Validity of Multiplex Genetic Testing in Complex Diseases: Reply to Janssens et al.

To the Editor:

We appreciate the comments by Janssens and her associates (2004 [in this issue]) regarding our study on the use of likelihood ratios to improve the prediction of complex diseases by testing for multiple-susceptibility genes (Yang et al. 2003). As Janssens et al. correctly point out, our study considers only the predicted probability of disease for subjects who have all positive testing results, and this is likely to be an infrequent occurrence. We think that the suggestion made by Janssens et al. to use receiver-operating–characteristic (ROC) curves to assess multiple genetic testing is very useful. The ROC curves provide a valuable way of evaluating the accuracy and discriminatory ability of diagnostic tests (Hanley 1989). Janssens et al. use the ROC curves to assess the classification of patients into a disease group, but multiplex genetic testing is likely also to be of value in identifying people who are at lower-than-average risk for developing a particular disease. This might allow them to put off receiving a more expensive intervention for some time—for example, to defer mammography for breast cancer detection for 10 years (Fletcher 1997) or to avoid screening for prostate cancer until $\geqslant$60 years of age (Harris and Lohr 2002).

The predictive value of combining tests obviously does depend on the relative risk associated with each component test, with a bigger effect resulting from tests that make larger independent contributions. Janssens et al. suggest that an odds ratio of 1.5–1.7 for each test is more likely than an odds ratio of 3.5. This might be true, but we do not yet know what the relative frequency of genes of larger or smaller effect will turn out to be for any common multifactorial disease. We used five genetic tests and an environmental factor as a simplified illustration in our analysis, but, in the near future, 50

or 100 genetic tests might be available for many common diseases. If there are numerous predisposing alleles and each has an independent odds ratio of only 1.5–1.7, the overall effect would still be substantial. We simulated models of 10, 15, and 20 genes with a risk of 1.5–1.7 each and found the areas under the ROC curves (AUCs) to be 0.70, 0.74, and 0.77, respectively. The discriminatory ability of 20 gene tests, each with an odds ratio of 1.5–1.7, is comparable with the test of total cholesterol level for prediction of coronary heart disease (Wilson et al. 1998). The effect would be even greater if only 5% or 10% of all alleles tested had odds ratios in the range of 2.5–3.5 or if we could identify combinations of a few genes and/or gene-environment interactions that are strong predictors of the disease.

The comments of Janssens et al. also raise several interesting points regarding different perspectives on multiple genetic testing. Epidemiologic studies, including those on the utility of ROC curves for screening, provide a useful population perspective. In contrast, clinicians usually focus on individual patients rather than on the population as a whole, and this focus will be enhanced by the development of personalized genomic medicine (Roses 2000; Jain 2002). It is true that no more than a few people per million might turn out to have a very high risk defined by positive results for multiple genetic tests for a particular disease. However, it might be very important to these few people to know that they are at high risk if an intervention is available to prevent the disease. Our likelihood-ratio–based method provides an approach that is useful for individual patients and their physicians in predicting the probability of developing disease.

QUANHE YANG,[1] MUIN J. KHOURY,[2] LORENZO BOTTO,[1] J. M. FRIEDMAN,[4] AND W. DANA FLANDERS[3]

[1]*National Center on Birth Defects and Developmental Disabilities and* [2]*Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention (CDC), and* [3]*Department of Epidemiology, School of Public Health, Emory University, Atlanta; and* [4]*Department of Medical Genetics, University of British Columbia, Vancouver*

References

Fletcher SW (1997) Whither scientific deliberation in health policy recommendations? Alice in the Wonderland of breast-cancer screening. N Engl J Med 336:1180–1183

Hanley JA (1989) Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 29:307–335

Harris R, Lohr KN (2002) Screening for prostate cancer: an update of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 137:917–929

Jain KK (2002) Personalized medicine. Curr Opin Mol Ther 4:548–558

Janssens ACJW, Pardo MC, Steyerberg EW, van Duijn CM (2004) Revisiting the clinical validity of multiplex genetic testing in complex diseases. Am J Hum Genet 74:585–588 (in this issue)

Roses AD (2000) Pharmacogenetics and the practice of medicine. Nature 405:857–865

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97:1837–4187

Yang Q, Khoury MJ, Botto L, Friedman JM, Flanders WD (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. Am J Hum Genet 72:636–649

Address for correspondence and reprints: Dr. Quanhe Yang, National Center on Birth Defects and Developmental Disabilities (CDC), 1600 Clifton Road, MS E-86, Atlanta, GA 30333. E-mail: qay0@cdc.gov

0002-9297/2004/7403-0030$15.00

Am. J. Hum. Genet. 74:589–591, 2004

Impact of Genotyping Errors on Type I Error Rate of the Haplotype-Sharing Transmission/Disequilibrium Test (HS-TDT)

To the Editor:

In a recent issue of the *Journal,* Zhang et al. (2003) proposed a haplotype-sharing transmission/disequilibrium test (HS-TDT) for the null hypothesis of no linkage or no association between a disease and a chromosomal region in which several tightly linked markers have been typed. Their method is applicable to data of nuclear families without phase information. The general idea of their approach is to compare the similarity of the transmitted haplotypes with the similarity of the nontransmitted haplotypes. If the chromosomal region contains a susceptibility locus, it is expected that the haplotypes being transmitted to affected children are more similar than parental haplotypes that have not been transmitted. This reasoning seems intuitively appealing. However, it may be supposed that a larger observed similarity for transmitted than for nontransmitted haplotypes is not necessarily due to the presence of a disease-susceptibility locus but can be a consequence of undetected genotyping errors. The proportion of genotyping errors that result in a Mendelian inconsistency (MI) is relatively small for family trios (Gordon et al. 1999). More important, in the context of HS-TDT, is the fact that the chance to detect a genotyping error differs for transmitted and nontransmitted haplotypes. Obviously, mistyping of an allele on a nontransmitted parental haplotype can never

Table 1

Estimates (Based on 1,000 Replicated Samples) of the True Type I Error Rate of HS-TDT in the Absence of Genotyping Errors for Nominal Type I Error α and for Two Different AOs to Handle Ambiguous Families

	Estimated True Type I Error Rate			
	AO1		AO2	
N	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
100	.049	.009	.045	.006
200	.061	.009	.061	.009
1,000	.045	.010	.045	.010

result in a MI and, therefore, cannot become prominent. Another way to understand this problem is to imagine that transmitted haplotypes are partially checked for their integrity, whereas there is no such checking at all for the nontransmitted haplotypes. A single error occurring at one locus of a haplotype, however, can have a tremendous effect on the measure of similarity of this haplotype with all other haplotypes. Thus, nontransmitted haplotypes can appear less similar than transmitted haplotypes as a result of undetected genotyping errors. In statistical terms, genotyping errors may lead to an inflated type I error rate for the HS-TDT.

To quantify the magnitude of this inflation, we performed a simulation study. Our simulation study assumes that, for 19 tightly linked and equidistant diallelic marker loci, only 29 different haplotypes occur in the population. This set of haplotypes and the corresponding frequencies are shown in table A of the online-only supplemental material. For all family trios, we generate the parents' genotypes according to this haplotype distribution. The haplotype pair in the child is obtained by randomly selecting one of the two haplotypes in each parent. Next, genotyping errors are introduced independently into the alleles according to the stochastic error model, for which ε denotes the probability that, at each marker locus, the allele is changed. We consider the cases for $\varepsilon = 0$ (no genotyping errors), $\varepsilon = 0.001$, $\varepsilon = 0.005$, and $\varepsilon = 0.01$. Sometimes, a genotyping error becomes visible by leading to MI. We consider three different error options (EOs) as strategies for responding to such an inconsistency: (EO1) genotypes of a marker locus with MI are considered to be unknown in all individuals of the family; (EO2) in the presence of MI for at least one marker locus, the whole family is discarded from the analysis; and (EO3) a marker locus showing MI is typed again, and it is assumed that the retyping results in error-free genotypes for this marker locus. The number of family trios in a sample is denoted by N, and we let $N = 100$, 200, or 1,000. Note that for EO2, the number of families used for statistical analysis is generally smaller than N. For each combination of ε, EO, and N, 1,000 samples are generated.

To analyze a simulated sample by the HS-TDT proposed by Zhang et al. (2003), we discard the phase information. The first step for statistical analysis is to obtain haplotype estimates. This is achieved by the program FAMHAP (Becker and Knapp, in press), which applies a locus-iterative mode of the expectation-maximization (EM) algorithm (Dempster et al. 1977) to obtain maximum-likelihood estimates of haplotype frequencies in general nuclear families. Zhang et al. (2003) discussed two different analysis options (AOs) to make use of estimated haplotype frequencies in case of ambiguous phase information in the families of the sample: (AO1) each possible haplotype explanation of an ambiguous family is weighted by its relative likelihood and (AO2) each ambiguous family is assigned its most likely haplotype explanation. Our simulated samples are analyzed by both of these AOs. The HS-TDT requires a permutation procedure to obtain the P value of the test. For each sample, we estimate the P value by 10,000 permutations. The true type I error rate at nominal error rate α is estimated by the fraction of the 1,000 replicated samples resulting in a P value $\leqslant\alpha$.

The results are shown in tables 1 and 2. If there are no genotyping errors (i.e., $\varepsilon = 0$), table 1 reveals a good agreement between nominal and true type I error rate, irrespective of the AO used to handle ambiguous families. (Note that if there are no genotyping errors, no MIs can occur, and, therefore, the EO is irrelevant.) Table 2 gives estimated type I error rates for the three EOs

Table 2

Estimates (Based on 1,000 Replicated Samples) of the True Type I Error Rate of HS-TDT (AO1) in the Presence of Genotyping Errors for Nominal Type I Error α and for Three Different Options to Handle MIs

		Estimated True Type I Error Rate					
		EO1		EO2		EO3	
N	ε	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
100	.01	.576	.297	.471	.228	.540	.264
100	.005	.215	.079	.219	.075	.208	.075
200	.005	.389	.164	.393	.146	.383	.164
1,000	.001	.146	.039	.167	.045	.147	.039

and AO1. Results are virtually identical when the most likely haplotype explanation is assigned to ambiguous families (see table B in the online-only supplemental material). As is obvious from table 2, the agreement between nominal and true type I error rate is disastrous in the presence of genotyping errors. Even quite small probabilities of genotyping errors lead to a dramatic inflation of the type I error. For fixed values of ε, the extent of this inflation increases with increasing sample size (N), as can be seen by comparing the second and third row in table 2. For a large sample size of $N = 1{,}000$ family trios, an error probability of $\varepsilon = 0.1\%$ is sufficient to falsely reject the null hypothesis at $\alpha = 0.05$ in almost every sixth study. For small values of N and large values of ε, the inflation of type I error is slightly less pronounced for EO2 than for EO1, which is explained by noting that EO2 leads to a decrease of the sample size used for the analysis. At first sight, it may be surprising that no essential decrease of the inflation of type I error is obtained by employing EO3. However, correcting genotypes leading to MIs does not affect errors in the nontransmitted haplotypes.

What are possible limitations of our simulation study? We assume a specific haplotype structure in the population, such that only 29 different haplotypes are present. Indeed, we conjecture that with larger haplotype diversity, the effect of genotyping errors on the type I error rate of the HS-TDT will be less pronounced than in the example considered here. On the other hand, however, it does not seem very realistic to expect that the HS-TDT will have substantial power to detect a disease locus in a region in which the markers are in complete or nearly complete linkage equilibrium in the population. Thus, although our example describes a specific situation, it does not seem to be unrealistic for the genetic structure of a region for which the HS-TDT may have a good chance of detecting a disease locus. A second possible limitation is that we employed a quite simple error model that assumes the independence of genotyping errors from factors such as marker locus, true allele, etc. However, we see no reason why the behavior of the type I error rate of the HS-TDT should be qualitatively different for more complex models of genotyping errors. Additionally, we are convinced that the range 0.1%–1% for the probability (ε) of a genotyping error considered here is not too pessimistic for currently available methods of high-throughput genotyping.

In summary, we have shown that the correctness of genotypes is crucial for obtaining meaningful results by the HS-TDT. We have also demonstrated that the retyping of only those marker loci that show MIs within a family is useless. A more extreme approach is to genotype all marker loci in all families in duplicate, which is very expensive and certainly not very popular with geneticists responsible for generating genotypes. However, unless extreme care is taken to guarantee the integrity of the data analyzed by the HS-TDT, this interesting and appealing method has the potential of becoming a mighty tool for the enlargement of the heap of false-positive association results in human genetics.

Acknowledgments

This work was supported by grant Kn378/1 (Project D1 of FOR 423) from the Deutsche Forschungsgemeinschaft.

MICHAEL KNAPP AND TIM BECKER
Institute for Medical Biometry, Informatics, and Epidemiology
University of Bonn
Bonn

Electronic-Database Information

The URL for data presented herein is as follows:

FAMHAP: Haplotype Frequency Estimation, http://www.uni-bonn.de/~umt70e/becker.html

References

Becker T, Knapp M. Maximum likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol (in press)

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–22

Gordon D, Heath SC, Ott J (1999) True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Hum Hered 49:65–70

Zhang S, Sha Q, Chen HS, Dong J, Jiang R (2003) Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 73:566–579

Address for correspondence and reprints: Dr. Michael Knapp, Institute for Medical Biometry, Informatics, and Epidemiology, University of Bonn, Sigmund-Freud-Straße 25, D-53105 Bonn, Germany. E-mail: knapp@uni-bonn.de

0002-9297/2004/7403-0031$15.00

Am. J. Hum. Genet. 74:591–593, 2004

Reply to Knapp and Becker

To the Editor:

Knapp and Becker (2004 [in this issue]) have argued that genotyping errors may lead to an inflated type I

Table 1

Parameters and Results of Simulation Study of Type I Error Rate of HS-TDT in the Presence of Genotyping Error

	Parameters			Type I Error Rate for			
				EO2		EO3	
Similarity Measure	No. of Children/ Family	No. of Nuclear Families/ Sample	Typing Error Rate (ϵ)	$\alpha = .05$[a]	$\alpha = .01$[a]	$\alpha = .05$[a]	$\alpha = .01$[a]
Original:	1	100	.01	.457	.227	.364	.147
	1	100	.005	.228	.08	.193	.075
	1	200	.005	.364	.147	.315	.120
	3	100	.01	.053	.012	.06	.016
	3	100	.005	.056	.013	.046	.011
	3	200	.005	.044	.008	.053	.006
New:	1	100	.01	.117	.037	.092	.019
	1	100	.005	.079	.016	.073	.014
	1	200	.005	.081	.016	.101	.029
	3	100	.01	.059	.016	.042	.006
	3	100	.005	.059	.015	.043	.009
	3	200	.005	.047	.010	.045	.003

Note.—The "original similarity measure" refers to the one used by Zhang et al. (2003). Simulation studies were based on 1,000 replicated samples.

[a] α = nominal type I error rate.

error rate for the haplotype-sharing transmission/disequilibrium test (HS-TDT) that we proposed (Zhang et al. 2003). The reason is that transmitted haplotypes are partially checked for genotyping errors by Mendelian inconsistency (MI), whereas there is no such checking at all for nontransmitted haplotypes. As a result of the unbalanced checking for genotyping errors, nontransmitted haplotypes appear less similar than transmitted haplotypes, which may lead to an inflated type I error rate for the HS-TDT. This is especially true for cases in which there is only one child per nuclear family. As noted by Gordon et al. (2001), the original TDT also has this problem. The HS-TDT that we proposed is applicable to any size of nuclear family and to different traits. To quantify the magnitude of type I error inflation of HS-TDT, Knapp and Becker (2004) performed a simulation study of nuclear families with one child. In fact, the magnitude of the type I error inflation caused by the unbalanced checking of the genotyping errors depends on the genotyping error rate as well as the following factors:

1. The number of children. If there is more than one child in the nuclear family, the genotyping errors in the haplotypes that do not transmit to the first child may be still detectable because these haplotypes may transmit to the other children. So, the inclusion of families with more than one child can reduce the type I error inflation.
2. The allele frequencies. A smaller minor allele frequency will lead to a larger probability of homozygous genotypes and, therefore, a larger probability of detectable genotyping errors (MI). Consequently, it will lead to larger type I error inflation (see table 3 of Gordon et al. 2001). For HS-TDT, a marker with a small minor allele frequency in the middle part of the haplotype has a bigger effect than a marker with a small minor allele frequency in the edge part of the haplotype.
3. The haplotype similarity measure.

We believe that the reasons for the high type I error rate of HS-TDT in Knapp and Becker's simulation studies are the following: (1) only families with one child were used; (2) the minor allele frequencies are small for the markers in the middle part of the haplotypes (for the total 19 markers, the minor allele frequencies from marker 7 to marker 16 are 0.16, 0.125, 0.143, 0.143, 0.11, 0.268, 0.089, 0.143, 0.143, and 0.036, respectively); and (3) the haplotype similarity measure that we proposed in Zhang et al. 2003 is not robust to genotyping errors. To compare the different haplotype similarity measures, we propose another measure (called "new similarity measure") as follows. For two haplotypes, H and h, let H_i (h_i) denote the allele of the haplotype H (h) at marker i. To find the similarity measure of the two haplotypes around marker i, we compare alleles of the two haplotypes in the right-hand markers, beginning with marker $i + 1$, until marker $i + r$ satisfies $H_{i+r} \neq h_{i+r}$ and either $H_{i+r+1} \neq h_{i+r+1}$ or $H_{i+r+2} \neq h_{i+r+2}$. Then, similarly, we compare alleles of the two haplotypes in the left-hand markers, beginning with

marker $i-1$, until marker $i-l$ satisfies $H_{i-l} \neq h_{i-l}$ and either $H_{i-l-1} \neq h_{i-l-1}$ or $H_{i-l-2} \neq h_{i-l-2}$. The new similarity measure is defined as the distance between marker $i-l$ and marker $i+r$. Note that a genotyping error that occurs at one marker but does not occur at the nearby markers will not affect the new similarity measure. The probability that genotyping errors will occur in several consecutive markers is very small. To compare the effect of the number of children and different haplotype similarity measures, we performed simulation studies in which we used the data and the error options EO2 and EO3 given by Knapp and Becker (2003). We did not use EO1 because our program automatically deletes the families with MI genotyping errors. The simulation results are summarized in table 1. This table reveals that, if there are three children in each of the nuclear families, a good agreement between the nominal and estimated type I error rate is evident for all the simulated samples. In the case of one child per family, the inflation of the type I error rate is greatly reduced by using the new similarity measure. We currently are investigating methods that are more robust to genotyping errors.

SHUANGLIN ZHANG, QIUYING SHA, HUANN-SHENG CHEN, JIANPING DONG, AND RENFANG JIANG
Department of Mathematical Sciences
Michigan Technological University
Houghton, MI

References

Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 69:371–380

Knapp M, Becker T (2004) Impact of genotyping error on type I error rate of the haplotype-sharing transmission /disequilibrium test (HS-TDT). Am J Hum Genet 74:589–591 (in this issue)

Zhang S, Sha Q, Chen HS, Dong J, Jiang R (2003) Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 73:566–579

Address for correspondence and reprints: Dr. Shuanglin Zhang, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931. E-mail: shuzhang@mtu.edu

0002-9297/2004/7403-0032$15.00