= 50, 100, and 1,000). Each population was sampled, and D' was calculated. This sampling was repeated 1,000 ×, and the mean and variance for D' were calculated. Table 1 shows the ratio between the theoretical values of $V(\hat{D}')$ and the variances in the computer simulation. It is clear that, in general, the approximation of the theoretical $V(\hat{D}')$ is quite satisfactory and that the ratio approximates to 1 quite well, even for samples as small as $n = 100$. As expected from asymptotic theory, most of the significant differences between the two variances, detected by the $F_{\max}$-statistic test (Sokal and Rohlf 1995, p. 397), occur for $n = 50$, especially for extreme allele frequencies. From the results, use of $V(\hat{D}')$ for experimental sample sizes equal to or higher than $n = 100$ can be recommended.

Carlos Zapata,[1] Gonzalo Alvarez,[1]
and Carmen Carollo[2]
[1]*Departmento de Biología Fundamental and*
[2]*Departamento de Estadística e I. O., Universidad de Santiago, Santiago de Compostela, Spain*

Acknowledgments

This work was supported by Xunta de Galicia (of Spain) grant XUGA-20002B95 (to C.Z).

References

Brown AHD (1975) Sample sizes required to detect linkage disequilibrium between two or three loci. Theor Popul Biol 8:184–201

Chakraborty R (1984) Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. Genetics 108:719–731

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1–16

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

——— (1988) Inference of recombinational hotspots using gametic disequilibrium values. Heredity 60:435–438

Hedrick PW, Jain S, Holden L (1978) Multilocus systems in evolution. Evol Biol 11:101–184

Hedrick PW, Thomson G (1986) A two-locus neutrality test: applications to humans, *E. coli* and lodgepole pine. Genetics 112:135–156

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229–239

Kendall M, Stuart A (1977) The advanced theory of statistics. Vol 1. Charles Griffin, London

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49–67

——— (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Sokal RR, Rohlf FJ (1995) Biometry. WH Freeman, New York

Zapata C, Alvarez G (1992) The detection of gametic disequilibrium between allozyme loci in natural populations of *Drosophila.* Evolution 46:1900–1917

——— (1993) On the detection of nonrandom associations between DNA polymorphisms in natural populations of *Drosophila.* Mol Biol Evol 10:823–841

——— (1997*a*) Testing for homogeneity of gametic disequilibrium among populations. Evolution 51:606–607

——— (1997*b*) On Fisher's exact test for detecting gametic disequilibrium between DNA polymorphisms. Ann Hum Genet 61:71–77

Zapata C, Visedo G (1995) Gametic disequilibrium and physical distance. Am J Hum Genet 57:190–191

Address for correspondence and reprints: Dr. Carlos Zapata, Departamento de Biología Fundamental, Area de Genética, Facultad de Biología, Universidad de Santiago, Santiago de Compostela, Spain. E-mail: bfcazaba@usc.es

0002-9297/97/6103-0037$02.00

Am. J. Hum. Genet. 61:774–778, 1997

Transmission/Disequilibrium Tests for Multiallelic Loci

To the Editor:

Kaplan et al. (1997) address the interesting question of how the biallelic transmission/disequilibrium test (TDT) should be extended to multiallele loci. Four recently proposed test statistics were described, and their properties were investigated by simulation studies. Here, I would like to point out some defects of the Monte Carlo-T_m test and the χ^2-$T_{m\text{het}}$ test that were not revealed by these simulation studies.

All four test statistics are based on the square contingency table of the counts of allele transmission, as set out in table 1. The cell count n_{ij} is the number of parents

Table 1

Counts of Allele Transmission and Nontransmission

Transmitted Allele	Nontransmitted Allele 1	2	..	m	Total
1	n_{11}	n_{12}	..	n_{1m}	$n_{1.}$
2	n_{21}	n_{22}	..	n_{2m}	$n_{2.}$
.	..	..	..	..	
m	$\underline{n_{m1}}$	$\underline{n_{m2}}$	$\underline{..}$	$\underline{n_{mm}}$	$\underline{n_{..}}$
Total	$n_{.1}$	$n_{.2}$	..	$n_{.m}$	$n_{..}$

with genotype ij who transmitted allele i to an affected offspring. The T_m-test statistic is defined as follows

$$T_m = \sum_{i=1}^{m} \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i}} \, .$$

The closely related $T_{m\text{het}}$-test statistic can be obtained by redefining the marginal totals to exclude the diagonal elements, $n_{i.}^* = n_{i.} - n_{ii}$ and $n_{.i}^* = n_{.i} - n_{ii}$. These "adjusted" marginal totals, based entirely on the off-diagonal elements, are then combined to give the $T_{m\text{het}}$-test statistic

$$T_{m\text{het}} = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(n_{i.}^* - n_{.i}^*)^2}{n_{i.}^* + n_{.i}^*} \, .$$

T_m and $T_{m\text{het}}$ can be regarded as weighted sums of the squared discrepancies between the row totals and the column totals of the m alleles, where the weights for the squared marginal discrepancy $d_i^2 = (n_{i.} - n_{.i})^2 = (n_{.i}^* - n_{.i}^*)^2$ are $w_i = 1/(n_{i.} + n_{.i})$ and $w_i^* = [(m - 1)/m]/[1/(n_{i.}^* - n_{.i}^*)]$, for T_m and $T_{m\text{het}}$, respectively. The crucial difference between the weighting schemes for T_m and $T_{m\text{het}}$ is that the former but not the latter depends on the diagonal elements of the contingency table. In calculating T_m, the higher the frequency of the homozygous genotype of an allele, the lower is the weight given to the squared marginal discrepancy of that allele. This feature of T_m can have an adverse effect on power under population stratification, as can be demonstrated by a triallelic locus in a population with three strata, for which the patterns of allele transmission are as shown in table 2A. These strata combine to give the overall pattern of allele transmission in the population, as shown in table 2B. For this combined table, the T_m statistic is only 3.58, because the largest marginal discrepancies (alleles 1 and 3) are "weighted down" by the correspondingly large diagonal elements (homozygous 11 and 33 genotypes). In contrast, the $T_{m\text{het}}$ statistic is invariant to the diagonal elements and takes the value of 14.58 for the same contingency table. Under these circumstances, tests based on the T_m statistic will be less powerful than those based on the $T_{m\text{het}}$ statistic, *even if* null distributions are determined by Monte Carlo simulation. Conversely, if the largest marginal discrepancies occur for alleles with disproportionately small homozygous frequencies, then the power of tests based on the T_m statistic will be greater than that of tests based on the $T_{m\text{het}}$ statistic. The dependence of the power of the T_m test on homozygous parental genotype frequencies is an unattractive feature that violates one of the basic principles of the original biallelic TDT.

Although the $T_{m\text{het}}$ statistic has the desirable property of being invariant to the frequencies of homozygous parental genotypes, its asymptotic distribution is not χ^2 with $m - 1$ df, as claimed by Spielman and Ewens (1996). Let the count of parents with heterozygous genotype ij be N_{ij} ($i < j$)—that is, $N_{ij} = n_{ij} + n_{ji}$, and let the count of heterozygous parents possessing allele i be N_i; that is,

$$N_i = \sum_{j=1}^{i-1} N_{ji} + \sum_{j=i+1}^{m} N_{ij} = n_{i.}^* + n_{.i}^* \, ;$$

then the asymptotic variance of $T_{m\text{het}}$ can be shown to be

$$\mathrm{V}_{m\text{het}} = \left(\frac{m-1}{m}\right)^2 \left(2m + 4 \sum_{j>i} \frac{N_{ij}^2}{N_i N_j}\right) ,$$

with the summation being over all $m(m - 1)/2$ possible heterozygous parental genotypes ij (for derivation, see the appendix). For $m = 2$ (i.e., the case in which $T_{m\text{het}}$ reduces to the original biallelic TDT), this formula gives the asymptotic variance as $(1/2)^2(4 + 4) = 2$, in accordance with a χ^2 distribution with 1 df. For $m > 2$, the asymptotic variance can be shown to be at a minimum when the frequencies of all parental heterozygous genotypes are equal; that is, $N_{ij} = N_{kl}$ for all ij and kl. In this case, $\mathrm{V}_{m\text{het}} = 2(m - 1)$, which is consistent with $T_{m\text{het}}$ having a χ^2 distribution with $m - 1$ df. When the frequencies of the parental heterozygous genotypes are not equal, however, $\mathrm{V}_{m\text{het}}$ will exceed $2(m - 1)$, so that a χ^2 test based on $T_{m\text{het}}$ will tend to be anticonservative.

The T_l statistic proposed by Sham and Curtis (1995) does not have the undesirable properties of T_m and $T_{m\text{het}}$. Like the original biallelic TDT, it is invariant to the frequencies of homozygous parental genotypes and has an asymptotic χ^2 distribution under the null hypothesis. The statistic is based on a well-established statistical model for square contingency tables proposed by Bradley and Terry (1952). For allele-transmission data, the model specifies that the logarithm of the odds of transmitting allele i, given parental genotype ij, is given by $\ln(p_{ij}/p_{ji}) = b_i - b_j$, where b_i and b_j are parameters associated with alleles i and j, respectively. This model has $m - 1$ independent parameters, with b_m being arbitrarily set at 0. The null hypothesis is that $b_1 = b_2 = \ldots = b_{m-1} = 0$, so that the two alleles of every heterozygous parent are equally likely to be transmitted to an affected offspring. The log-likelihood function of the model is

$$\ln L = \sum \left[n_{ij}\ln(p_{ij}) + n_{ji}\ln(p_{ji})\right] ,$$

where the summation is over all $m(m - 1)/2$ heterozygous genotypes. The T_l test uses the likelihood-ratio statistic $T_1 = 2(\ln \mathrm{L}_1 - \ln \mathrm{L}_0)$, where $\ln \mathrm{L}_1$ is the value of the

Table 2

Transmission Tables for a Stratified Population

A. Transmission Table for Three Population Strata

	Nontransmitted								
	Stratum 1			Stratum 2			Stratum 3		
Transmitted	1	2	3	1	2	3	1	2	3
1	27	36	27	200	0	0	0	0	0
2	24	32	24	0	0	0	0	0	0
3	9	12	9	0	0	0	0	0	200

B. Transmission Table for Entire Population

	Nontransmitted			
Transmitted	1	2	3	Total
1	227	36	27	290
2	24	32	24	80
3	9	12	209	230
Total	260	80	260	600

log likelihood maximized with respect to the $m - 1$ model parameters, and $\ln L_0$ is the value of the log likelihood at the null hypothesis.

The Bradley-Terry model gives rise to other statistics, including the score test proposed by Stuart (1955). Let $\mathbf{d} = (d_1, d_2, \ldots, d_{m-1})^T$ be a vector of the marginal discrepancies for alleles 1 to $m - 1$. Let the covariance matrix of $\mathbf{d}$ (under the null hypothesis) be denoted by $\mathbf{V}$, with diagonal elements $v_{ii} = N_i$ and off-diagonal elements $v_{ij} = -N_{ij.}$ Stuart's score test, T_s, is given by $T_s = \mathbf{d}^T\mathbf{V}^{-1}\mathbf{d}$. The similarity of T_l and T_s, as well as the difference between these two statistics and $T_{m\text{het}}$, can be illustrated by table 3. The evidence for asymmetry of this table derives entirely from the differential transmissions of alleles 1 and 3 from parents with genotype 13. The χ^2 statistic for this difference is $(120 - 80)^2/(120 + 80) = 8$. This value can be regarded as an approximate upper bound for any reasonable χ^2 statistic. The T_l and T_s statistics, being 8.01 and 7.96, respectively, are therefore not unreasonable. In contrast, the $T_{m\text{het}}$ statistic is too large at 10.56, which reflects the inflated variance of the statistic and gives rise to an anticonservative test.

Table 3

Transmission Table with Highly Discrepant Parental Genotype Frequencies

	Nontransmitted			
Transmitted	1	2	3	Total
1	...	1	120	121
2	1	...	1	2
3	80	1	...	81
Total	81	2	121	204

Several other multiallele extensions of the TDT, including conditional logistic regression and the associated likelihood-ratio and score tests (Harley et al. 1995; Schaid 1996), as well as the weighted least-squares and the associated Wald tests (Duffy 1995; Rice et al. 1995), are also closely related to the Bradley-Terry model. Likelihood-ratio tests, score tests, and Wald tests are asymptotically equivalent in terms of power. Likelihood-ratio tests are more convenient for testing a system of nested hypotheses. Wilson (1997) has, for example, extended T_l to multiple loci. Although score tests and Wald tests are sometimes easier to compute than likelihood-ratio tests, the maximization of the log likelihood for the Bradley-Terry model can be achieved very efficiently by use of the iteratively reweighted least-squares algorithm for generalized linear models. This algorithm has been implemented for calculating T_l in the ETDT software, which uses LINKAGE-format pedigree and locus files (Sham and Curtis 1995). A Monte Carlo procedure (similar to that described by Kaplan et al.) has also been implemented in ETDT, which can be used to obtain empirical P values (based on T_l) for sparse tables.

Although the undesirable properties of T_m and $T_{m\text{het}}$ are only manifest in certain circumstances, the use of these statistics is unnecessary when tests without these defects—namely T_l, T_s, and other tests related to the

Bradley-Terry model—are available. These statistics can be computed rapidly by use of standard algorithms and are asymptotically χ^2, so that Monte Carlo methods for determining empirical P values become necessary only for sparse tables.

Finally, Kaplan et al. are incorrect in saying that Sham and Curtis (1995) recommended separate analyses for data from fathers and mothers. It is an inherent feature of the TDT that each "trio" (i.e., an affected offspring and the two parents) is divided into two separate pairs of observations: (1) a pair of transmitted and nontransmitted alleles from the father and (2) a pair of transmitted and nontransmitted alleles from the mother. These two pairs of observations are entered separately into a square contingency table. This procedure is fully justified only if the conditional probability of the genotype of an affected offspring, given the genotypes of the parents, is simply the product of the conditional probability of the paternally transmitted allele, given the paternal genotype, and the conditional probability of the maternally transmitted allele, given the maternal genotype. The violation of this independence assumption does not invalidate the TDT but can reduce its power (Schaid 1996).

Appendix

Let n_{ij} be a realization of the random variable X_{ij} for $i = 1, \ldots, m$, $j = i + 1, \ldots, m$. Conditioning on the parental heterozygous genotype frequencies dictates that $X_{ji} = N_{ij} - X_{ij}$. The adjusted row and column totals $n_{i.}^*$ and $n_{.i}^*$, which define $T_{m\text{het}}$, are then realizations of the random variables

$$X_{i.}^* = \sum_{j=1}^{m} X_{ij} - X_{ii}$$

and

$$X_{.i}^* = \sum_{j=1}^{m} X_{ji} - X_{ii} \, .$$

The variance of $T_{m\text{het}}$ is therefore

$$\mathrm{V}_{m\text{het}} = \left(\frac{m-1}{m}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{m} \frac{(X_{i.}^* - X_{.i}^*)^2}{X_{i.}^* + X_{.i}^*}\right] .$$

The random variable $(X_{i.}^* - X_{.i}^*)^2/(X_{i.}^* - X_{.i}^*)$ can be denoted as Y_i and rewritten as

$$Y_i = \frac{\left[\sum_{j\neq i} (2X_{ij} - N_{ij})\right]^2}{\sum_{j\neq i} N_{ij}} = \frac{\left[\sum_{j\neq i} 2X_{ij} - N_i\right]^2}{N_i} .$$

Under the null hypothesis, X_{ij} is binomial with parameters $(N_{ij}, 1/2)$, so that, for large samples, $2X_{ij}$ is asymptotically normal with mean and variance N_{ij}. Moreover, since X_{ij} and X_{ik} are independent for any $k \neq j$, $\Sigma_{j\neq i} 2X_{ij}$ is asymptotically normal with mean and variance N_i. It follows that the square root of Y_i is standard normal, so that Y_i is χ^2 with 1 df and $\mathrm{Var}(Y_i) = 2$. It also follows that the covariance between Y_i and Y_j is

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= \mathrm{Cov}\left[\frac{(2X_{ij} - N_{ij})^2}{N_i}, \frac{(2X_{ij} - N_{ij})^2}{N_j}\right] \\ &= \frac{N_{ij}^2}{N_i N_j} \mathrm{Var}\left[\frac{(2X_{ij} - N_{ij})^2}{N_{ij}}\right] \\ &= \frac{2N_{ij}^2}{N_i N_j} . \end{aligned}$$

The asymptotic variance of $T_{m\text{het}}$ is therefore

$$\begin{aligned} V_{m\text{het}} &= \left(\frac{m-1}{m}\right)^2 \left[\sum_i \mathrm{Var}(Y_i) + 2 \sum_{j>i} \mathrm{Cov}(Y_i, Y_j)\right] \\ &= \left(\frac{m-1}{m}\right)^2 \left(2m + 4 \sum_{j>i} \frac{N_{ij}^2}{N_i N_j}\right) . \end{aligned}$$

Acknowledgment

I thank the Wellcome Trust for University Award 037641 and Project Grant 043279.

Pak Sham

Department of Psychological Medicine
Institute of Psychiatry
London

References

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika 39:324–345

Duffy DL (1995) Screening a 2 cM genetic map for allelic association: a simulated oligogenic trait. Genet Epidemiol 12:595–600

Harley JB, Moser KL, Neas BR (1995) Logistic transmission modelling of simulated data. Genet Epidemiol 12:607–612

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691–702

Rice JP, Neuman RJ, Hoshaw SL, Daw EW, Gu C (1995) TDT with covariates and genomic screens with Mod scores: their behavior on simulated data. Genet Epidemiol 12:659–664

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multiallele marker loci. Ann Hum Genet 59:323–336

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

Stuart A (1955) A test of homogeneity of the marginal distribution in a two-way classification. Biometrika 42:412–416

Wilson SR (1997) On extending the transmission/disequilibrium test (TDT). Ann Hum Genet 61:151–161

Address for correspondence and reprints: Dr. Pak Sham, Department of Psychological Medicine, Institute of Psychiatry, DeCrespigny Park, London SE5 8AF, United Kingdom. E-mail: p.sham@iop.bpmf.ac.uk

0002-9297/97/6103-0038$02.00

Am. J. Hum. Genet. 61:778, 1997

Reply to Sham

To the Editor:

We thank Dr. Sham for his thoughtful comments on our paper and regret our incorrect statement that Sham and Curtis (1995) recommended separate analyses for fathers and mothers. We agree that heterozygous parents can be treated independently under the hypothesis of no linkage or no association and that, in general, they are not independent when there is linkage and association. We agree further with Dr. Sham that we did not study the consequences of stratification in our simulations. As we mentioned in our Discussion, we were thinking more of admixture as a source of association when linkage is absent.

We differ from Dr. Sham in standing by our statements concerning the distribution of $T_{m\text{het}}$. We had noticed, as he has, that the variance of the statistic may be greater than that for a χ^2 variable, but our simulations focused on the whole distribution. The statements in our paper were therefore based on percentiles rather than just the variance. We made explicit mention of the significance level and power being well approximated by χ^2 theory in our simulations at that time.

We have now performed simulations for populations from which samples had the degree of sparseness and imbalance shown in the example of Dr. Sham. We have found that power levels for Monte Carlo (MC)–$T_{m\text{het}}$ were very similar to those obtained under the assumption of χ^2. We also found the power of MC-$T_{m\text{het}}$ to be very similar to that of the Sham and Curtis likelihood ratio test, and it may even be greater under some circumstances.

There is theoretical interest in the statistic T_m because power of the test can be predicted from a noncentral χ^2 distribution for which the noncentrality parameter may be estimated. However, we stress that we did not advocate use of the T_m statistic, even when it is used with a Monte-Carlo procedure.

Norman L. Kaplan,[1] E. R. Martin,[1,2] and
B. S. Weir[2]

[1]*Biostatistics Branch, National Institute for Environmental Health Science, National Institutes of Health, Research Triangle Park, North Carolina; and* [2]*Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh*

References

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multiallele marker loci. Ann Hum Genet 59:323-336

Address for correspondence and reprints: Dr. Norman Kaplan, Biostatistics Branch, National Institute for Environmental Health Science, National Institutes of Health, P.O. Box 12233, Research Triangle Park, NC 27709.

0002-9297/97/6103-0039$02.00

Am. J. Hum. Genet. 61:778–779, 1997

Family Cell Lines Available for Research—An Endangered Resource?

To the Editor:

Diabetes continues to be a major health problem that is continuing to grow not only in the United States, but worldwide, at an escalating cost to the patient as well as to society. The cost to the individual is tremendous, and a shortened life span is the outcome regardless of whether expert care to delay late complications is available. The genetic factors that control the insulin-dependent type of diabetes, type 1 diabetes, are still not understood. Genomewide scanning has confirmed HLA as a major genetic factor for type 1 diabetes and a number of potential loci for contributing genes (Davies et al. 1994; Todd et al. 1996). This task was in part accomplished and progress accelerated by investigator-supported initiatives to establish large collections of DNA and cell lines from multiplex type 1 diabetes families. Some 5 years ago, emerging new human genome technologies were available, but there was a shortage of families with type 1 diabetes to be analyzed for genetic linkage or association between the disease and polymorphic markers on human chromosomes.

In a letter to the editor (Lernmark et al. 1990), the availability of cell lines and DNA from the Human Biological Data Interchange (HBDI), a not-for-profit orga-