Am. J. Hum. Genet. 66:2009–2012, 2000

Report

Unbiased Application of the Transmission/Disequilibrium Test to Multilocus Haplotypes

Frank Dudbridge,[1] Bobby P. C. Koeleman,[3] John A. Todd,[1] and David G. Clayton[2]

[1]Wellcome Trust Centre for Molecular Mechanisms in Disease, University of Cambridge, and [2]MRC Biostatistics Unit, Cambridge; and [3]Department of Immunohematology and Bloodbank, University of Leiden, Leiden

When the transmission/disequilibrium test (TDT) is applied to multilocus haplotypes, a bias may be introduced in some families for which both parents have the same heterozygous genotype at some locus. The bias occurs because haplotypes can only be deduced from certain offspring, with the result that the transmissions of the two parental haplotypes are not independent. We obtain an unbiased TDT for individual haplotypes by calculating the correct variance for the transmission count within a family, using information from multiple siblings if they are available. An existing correction for dependence between siblings in the presence of linkage is retained. To obtain an unbiased multihaplotype TDT, we must either count transmissions from one randomly chosen parent or count all transmissions and estimate the significance level empirically. Alternatively, we may use missing-data techniques to estimate uncertain haplotypes, but these methods are not robust to population stratification. An illustration using data from the insulin-gene region in type 1 diabetes shows that the validity and power of the TDT may vary by an order of magnitude, depending on the method of analysis.

The transmission/disequilibrium test (TDT) detects the simultaneous presence of linkage and association between a marker and a susceptibility locus and is robust to population stratification (Spielman et al. 1993). The test, as originally proposed, considers the transmission of an allele from a heterozygous parent to an affected offspring, comparing the total transmission count to that expected under Mendelian segregation; that is, if T denotes the number of times that the allele is transmitted, and if U denotes the number of times that it is not transmitted, then $(T - U)^2/(T + U)$ has a χ^2 distribution on 1 df, under the null hypothesis of no linkage or association.

One may wish to apply the TDT to haplotypes of two or more loci if, for example, the informativity of the markers is low or variable (Kruglyak 1999) or if haplotype-specific associations are suspected (Cucca and Todd 1996). However, haplotype information is not usually available, and the parental gametic haplotypes must be deduced from genotype information. In some cases, the haplotypes are ambiguous, and the family must be discarded from the analysis. A necessary condition for haplotype ambiguity is that there is a locus for which both parents and offspring have the same heterozygous genotype and another locus for which both parents and offspring do not have the same homozygous genotype. This condition is not always sufficient, since, in some families, the specific haplotype analyzed may not be ambiguous, but we prefer this condition, since it allows the most general analysis.

As the number of loci increases, the information loss due to haplotype ambiguity increases rapidly (Hodge et al. 1999). Furthermore, when the TDT is used, the problem is not limited to loss of information from these families. We show here that, in some families for which haplotypes are known, a potentially serious bias is introduced into the TDT if the loss of information from families with ambiguous haplotypes is not taken into account.

The bias applies both to the TDT applied to an individual haplotype and to its extensions, which test all haplotypes simultaneously. For single-haplotype TDT, we give a modification to the scoring that gives an unbiased χ^2 statistic. The correction for multihaplotype

Received January 25, 2000; accepted March 30, 2000; electronically published April 13, 2000.

Address for correspondence and reprints: Dr. Frank Dudbridge, Wellcome Trust/MRC Building, Addenbrookes Hospital, Hills Road, Cambridge CB2 2XY, United Kingdom. E-mail: frank.dudbridge@cimr.cam.ac.uk

0002-9297/2000/6606-0034$02.00

tests is not as straightforward, and we give some alternative strategies. We first give an example of a situation in which bias is introduced into the TDT when the analysis includes every case in which parental haplotypes are known.

Consider two diallelic loci, with alleles *A* and *a* at the first locus and alleles *B* and *b* at the second locus. Suppose that we observe one parent with genotypes *AA* and *Bb,* the other parent with genotypes *Aa* and *Bb,* and the offspring with genotypes *AA* and *BB*. Then we know that the *A-B* haplotype has been transmitted twice; however, the two transmissions are not independent, conditional on deduction of this gametic haplotype in both parents. To see this, note that there are four possible offspring consistent with both parents having an *A-B* haplotype. These have genotypes *AA-BB, Aa-Bb, AA-bB,* and *Aa-bb*. But *A-B* can only be deduced in both parents when the first and fourth of these offspring are observed, so the transmission count of *A-B* can only be 2 or 0.

We condition on deduction of the haplotypes because the problem is one of how to score the two observed transmissions. Therefore, we do not need to consider recombination between the loci (Lazzeroni and Lange 1998). The expected transmission count of *A-B* in this family is 1, and its variance is 1. But, for a single parent, the transmission count for a haplotype is 0 or 1, so the expected transmission count is $\frac{1}{2}$, with variance $\frac{1}{4}$. When the two parents are independent, the expected count is therefore 1, with variance $\frac{1}{2}$. Since, in this example, the variance is 1, we expect an increase in type 1 error when the analysis includes every case with known haplotypes.

Curtis and Sham (1995), Curtis (1997), and Knapp (1999) have shown that bias may be introduced into the TDT when genotype data are missing. Ambiguous haplotypes are another form of missing information that may bias the TDT. However, simply discarding the families that cause bias may result in considerable loss of data, particularly with diallelic marker loci. Following Knapp (1999), we construct an unbiased TDT including all transmissions of known haplotypes.

For each parent-offspring trio *i,* let T_i be the observed number of transmissions of the haplotype in question. Let e_i and v_i be the expectation and variance of T_i under the null hypothesis, conditional on deduction of the parental haplotypes. Then, $\sum (T_i - e_i)/\sqrt{V}$ has approximately the standard normal distribution under the null hypothesis, where $V = \sum v_i$. If T and U denote, respectively, the total transmission and nontransmission counts for the haplotype, then it follows that $(T - U)^2/4V$ has approximately a χ^2 distribution on 1 df.

If there are no loci for which the parents are doubly heterozygous, then haplotypes can always be deduced, and each heterozygous parent contributes $\frac{1}{4}$ to *V*. When this is the case for all parent-offspring trios, the usual TDT statistic is obtained. When there is a locus for which the parents are doubly heterozygous and there is another for which they are not doubly homozygous, the haplotypes may not be deduced from some offspring, but, if we assume that there is no recombination between the loci, then it may be possible to deduce them from additional offspring. If we assume that transmissions to all offspring are independent, then we can calculate v_i as follows.

Let n be the number of offspring in the family, so that there are 4^n possible sibships. In general, there are 2^n sibships for which parental haplotypes cannot be deduced from any offspring, so there are $4^n - 2^n$ cases in which the haplotypes can be deduced. We first consider a haplotype present in both parents. The expected transmission count to one offspring is 1. The observed count T_i is 2 when the offspring is homozygous for this haplotype, which occurs in 4^{n-1} of the possible sibships; $T_i = 0$ when the haplotype is not transmitted by either parent, which also occurs in 4^{n-1} sibships; otherwise the observed count is 1. Thus, the variance v_i for this trio is $2 \cdot 4^{n-1}/(4^n - 2^n)$.

For a haplotype present in only one parent, the expected transmission count is $\frac{1}{2}$. The observed count T_i is either 0 or 1 for each of the sibships with known haplotypes, so the variance v_i is $\frac{1}{4}$. This is the same variance as occurs when there is no haplotype ambiguity, so no bias would be introduced into the TDT in this case. Thus, for TDT applied to individual haplotypes, the only situation that biases the test is when (*a*) parents are doubly heterozygous at some locus and are not doubly homozygous at another and (*b*) the same haplotype is deduced in both parents.

To combine the TDT for individual haplotypes into a single test, Spielman and Ewens (1996) and Cleves et al. (1997) have proposed the test of marginal homogeneity (T_{mh}), calculated by summing the TDT statistic for each haplotype and scaling the sum by $(m - 1)/m$, where m is the number of haplotypes. The scaling factor accounts for dependence of the TDT between the different haplotypes, but, when there is possible haplotype ambiguity, an additional source of dependence is introduced. Since haplotypes can be deduced only when the offspring is homozygous at the locus for which the parents are doubly heterozygous, the transmissions of the two haplotypes are not independent, regardless of whether the same haplotype is deduced in both parents. Thus, although valid χ^2 statistics may be constructed for the individual haplotypes, T_{mh} will not necessarily be valid. Furthermore, the extended TDT (Sham and Curtis 1995*a*), which is another commonly used multiallelic TDT, is a likelihood-ratio test that cannot use our unbiased scoring. Thus, for the multihaplotype TDT, the test may be biased by any family with parents doubly heterozygous at some locus and not doubly homozygous

at another locus. In these cases, we may ensure validity of the χ^2 distribution by counting transmissions from one randomly chosen parent. Alternatively, we may count all known haplotype transmissions and use Monte Carlo procedures to estimate the significance level empirically (Sham and Curtis 1995*b;* Cleves et al. 1997; Kaplan et al. 1997).

When the TDT is used as a test of linkage in the presence of association, all affected siblings may be regarded as independent, so haplotypes may be deduced from all available siblings; however, if the TDT is used as a test of association in the presence of linkage, the transmissions to multiple affected siblings are not independent (Spielman and Ewens 1996). Cleves et al. (1997) and Martin et al. (1997) have proposed a test (T_{sut}) in which transmissions are counted only from parents transmitting the same allele (here, haplotype) to all affected siblings. This approach may be retained here, with the aforementioned caveat for the multihaplotype test. If an additional affected sibling allows haplotypes to be deduced, he or she can have, at most, one haplotype in common with an ambiguous offspring. If that haplotype is transmitted to all affected siblings, then we count 1 transmission, with $v_i = \frac{1}{4}$. Unaffected siblings may be considered independently, as before.

Although these methods give an unbiased test for known haplotype transmissions, they do not deal with cases in which haplotypes are unknown; these data must still be discarded. An alternative approach for missing information has been given by Clayton (1999), who estimates parental haplotype frequencies and constructs a likelihood taking all possible solutions into account. This approach uses all the available information and is expected to have greater power than the methods given here. However, it can be very computationally intensive and forfeits an attraction of the TDT—namely, the freedom from population modeling. Consequently, the method is not robust to population stratification. If the strata are known, then we may estimate haplotype frequencies within each stratum, but one of the main reasons for preferring family-based association studies to the case-control design is that the structure of the population may be unknown (Falk and Rubinstein 1987). Nevertheless, when the issues of stratification and computation are not significant, the "Transmit" program of Clayton (1999) may be preferred to the present methods.

To illustrate these methods, we considered the transmission of the haplotype consisting of the "−" allele of the *INS*−23/*Hph*I SNP on chromosome 11p15 and the "Z" allele of the *HUMTH01* microsatellite located ~−9,000 bp from the *INS* ATG codon. Association of this protective haplotype with type 1 diabetes has been demonstrated in 198 U.K. families (Bennett et al. 1995), but the significance of the association varies according to the analysis used. The haplotype was certainly transmitted 41 times and was certainly not transmitted 103 times. The unbiased significance was $P < 3.4 \times 10^{-6}$, but the usual TDT calculation gives $P < 2.4 \times 10^{-7}$. This shows that incorrect scoring may affect the validity and, hence, the interpretation of the P value, by an order of magnitude. The standard TDT may be used if all families with doubly heterozygous parents are discarded; in this case, we obtained 35 transmissions and 61 nontransmissions, giving $P < 0.008$. On the other hand, the approach of Clayton (1999), which uses all the data, gives $P < 2 \times 10^{-6}$. Therefore, the power of the TDT may also be strongly affected by the method of analysis (software implementing unbiased TDT for haplotypes of arbitrary length is available from the authors, either at Internet site ftp://ftp-gene.cimr.cam.ac.uk/pub/software or via the following e-mail address: frank.dudbridge@cimr.cam.ac.uk).

Acknowledgments

We thank Heather Cordell, Iain Eaves, and Rebecca Twells for many useful comments. We thank the Wellcome Trust and MRC for support.

References

Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, et al (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9:284–292

Clayton DG (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177

Cleves MA, Olson JM, Jacobs KB (1997) Exact transmission/disequilibrium tests with multiallelic markers. Genet Epidemiol 14:337–347

Cucca F, Todd JA (1996) HLA susceptibility to type 1 diabetes: methods and mechanisms. In: Browning MJ, McMichael AJ (eds) HLA and MHCL genes, molecules and function. BIOS Scientific, Oxford, pp 383–406

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Curtis D, Sham PC (1995) A note on the application of the transmission/disequilibrium test when a parent is missing. Am J Hum Genet 56:811–812

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Hodge SE, Boehnke M, Spence MA (1999) Loss of information due to ambiguous haplotyping of SNPs. Nat Genet 21: 360–361

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

Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-com-

bined transmission/disequilibrium test. Am J Hum Genet 64: 861–870

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67–81

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61: 439–448

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

——— (1995*b*) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59:97–105

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

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516