analysis of 117 kb of human DNA containing the gene BRCA1. Genome Res 6:1029–1049

Stoppa-Lyonnet D, Laurent-Puig P, Essioux L, Pagès S, Ithier G, Ligot L, Fourquet A, et al (1997) BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Am J Hum Genet 60:1021–1030

Address for correspondence and reprints: Dr. Mary-Claire King, Departments of Genetics and Medicine, University of Washington, Box 357720, Seattle, WA 98195-7720. E-mail: mcking@u.washington.edu

0002-9297/99/6502-0038$02.00

Am. J. Hum. Genet. 65:578–580, 1999

Power Comparisons of the Transmission/Disequilibrium Test and Sib–Transmission/Disequilibrium-Test Statistics

To the Editor:

Several recent papers have considered the extension of the transmission/disequilibrium test (TDT) to families in which parental DNA is not available but in which unaffected siblings can be sampled. Each of these tests compares the alleles in the affected offspring with those in the unaffected offspring. The tests differ both in the precise statistics used and in the numbers of affected and unaffected offspring included. Spielman and Ewens (1998) have developed the sib TDT (S-TDT) for families with an arbitrary number of affected and unaffected members (including at least one of each). Curtis (1997) has used families with a single affected offspring and an arbitrary number of unaffected offspring but has analyzed only that unaffected offspring who has the genotype most different from that of the affected offspring. Boehnke and Langefeld (1998) have used a discordant-sib-pair approach. The S-TDT is a test of linkage, but it is also valid as a test of allelic association in which precisely one affected sibling and one unaffected sibling are used, as is the case in the tests that have been described by Curtis (1997) and Boehnke and Langefeld (1998).

These authors have considered power in different contexts—for example, across offspring genotype configurations (Spielman and Ewens 1998) and across genetic models (Boehnke and Langefeld 1998)—but none of the approaches used was intended to provide an overall assessment of the power of a sibling-based TDT statistic compared with that of the original formulation of the TDT. Here we derive a relationship between power for the S-TDT and the TDT, which shows that, to achieve similar power, considerably more genotyping is required for the S-TDT than for the TDT. This is intuitively clear, for the following reason. For both tests, a family is informative only if at least one parent is heterozygous. The S-TDT requires an additional condition to be true: both alleles from the heterozygous parent must be present in the offspring. This implies that the informativeness of the S-TDT statistic increases with the number of siblings genotyped. Because of the variation associated with the alleles inherited by the n unaffected siblings, we expect that, for finite n, the S-TDT will be less powerful than the TDT, with the power of the S-TDT tending toward that of the TDT as $n \to \infty$. Below we formalize this argument. Our results extend the power calculations of Spielman and Ewens (1998): in table 5 of their paper, they give the power of both the S-TDT and the TDT, for families with one heterozygous and one homozygous parent, a single affected child, and two to four unaffected children. Their power calculations are conditional on both alleles from the heterozygous parent being present in the offspring, which, as the authors acknowledge, covers only a small proportion of possible family genotype configurations. This conditioning on the offspring genotypes implies that all families are informative for the S-TDT, and therefore it crucially affects the power of the S-TDT. With this conditioning, the power of the S-TDT is almost as great as that of the TDT; without it, the power of the S-TDT may be considerably reduced.

For the sake of simplicity, we consider a sample of k families, assuming that in each family there are a single affected offspring and n unaffected offspring. All individuals have been genotyped at a diallelic marker locus with alleles M and m; let the numbers of M alleles in the offspring in the ith family be X_i for the affected sib and Y_{ij}, $j \in \{1,2,\ldots,n\}$, for the unaffected sibs. We condition on the parental genotypes in the sample and compare the TDT and S-TDT for this sample. The difference between the two statistics can be summarized as follows. The TDT compares $X_{\cdot} = \Sigma_{i=1}^{k} X_i$, with $\mathrm{E}(X_{\cdot}|H_0)$, where this expected value is calculated from the parental marker information, under the assumption that the null hypothesis is true—that is, either of the two alleles in a heterozygous parent is equally likely to be transmitted to an affected child. The S-TDT, however, is designed for use when this parental information is unavailable; instead, $X_{\cdot}$ is compared with $Y_{\cdot\cdot}/n$, where $Y_{\cdot\cdot} = \Sigma_{i=1}^{k}\Sigma_{j=1}^{n} Y_{ij}$ is the total number of M alleles in the unaffected offspring.

Our test statistics for the TDT and the S-TDT (T_{TDT} and $T_{\mathrm{S\text{-}TDT}}$, respectively) are obtained by the method described, by Spielman and Ewens (1998), as the Z-score procedure: test statistics are standardized to mean 0 and variance 1 and are assumed to follow a standard normal distribution. This gives $T_{\mathrm{TDT}} = (X_{\cdot} - \mu_0)/\sigma_0$, where μ_0 and σ_0^2 are, respectively, the mean and variance of $X_{\cdot}$, under the null hypothesis of no linkage.

The S-TDT permutation statistic compares X_i with a permutation of genotypes from the affected and unaffected individuals and then sums the resulting statistic over families. This is equivalent to a comparison of $X_{\cdot}$ with a pool of $X_{\cdot}$ and $Y_{\cdot\cdot}$, giving

$$T_{\text{S-TDT}} = \frac{X_{\cdot} - \frac{X_{\cdot}+Y_{\cdot\cdot}}{n+1}}{\sqrt{\mathrm{Var}\left(X_{\cdot} - \frac{X_{\cdot}+Y_{\cdot\cdot}}{n+1}\,|H_0\right)}} = \frac{nX_{\cdot} - Y_{\cdot\cdot}}{\sqrt{\mathrm{Var}(nX_{\cdot} - Y_{\cdot\cdot}\,|H_0)}} = \frac{nX_{\cdot} - Y_{\cdot\cdot}}{\sqrt{n(n+1)}\sigma_0}\ ,$$

since, under the null hypothesis, the random variables X_i and Y_{ij} are independent and identically distributed for each i, so $\mathrm{Var}(Y_{\cdot\cdot}|H_0) = n\mathrm{Var}(X_{\cdot}|H_0)$ and

$$\mathrm{Var}(nX_{\cdot} - Y_{\cdot\cdot}|H_0) = n^2\mathrm{Var}(X_{\cdot}|H_0) + n\mathrm{Var}(X_{\cdot}|H_0) = n(n+1)\sigma_0^2\ .$$

Note that, whereas Spielman and Ewens (1998) exclude from the S-TDT sibships when all sibs have the same genotype, we include them. This does not affect the value of the test statistic, because such families have 0 mean and 0 variance, but it does facilitate comparisons of the TDT and S-TDT, because both test statistics now use the same set of families.

We can define a second TDT statistic in these families, looking at inheritance of M alleles from heterozygous parents to unaffected children, giving $T'_{\text{TDT}} = (Y_{\cdot\cdot} - n\mu_0)/\sqrt{n}\sigma_0$. Then, using the expression for T_{TDT} above, we can write $T_{\text{S-TDT}}$ as

$$T_{\text{S-TDT}} = \frac{1}{\sqrt{n+1}}(\sqrt{n}T_{\text{TDT}} - T'_{\text{TDT}})\ .$$

Power comparisons of T_{TDT} and $T_{\text{S-TDT}}$ can be obtained through the expected values and variances of these statistics. For the models appropriate for many complex diseases, the probability that an M allele is transmitted from a heterozygous parent to an unaffected sib is very close to .5 (Spielman and Ewens 1998), so the genotypes of unaffected offspring can be treated as random observations from the parental genotypes. Then $\mathbf{E}(Y_{\cdot\cdot}) \approx n\mu_0$, and $\mathbf{E}(T'_{\text{TDT}}) \approx 0$, giving $\mathbf{E}(T_{\text{S-TDT}}) \approx \sqrt{n/(n+1)}\,\mathbf{E}(T_{\text{TDT}})$.

We now show that the variances of T_{TDT} and $T_{\text{S-TDT}}$ are approximately equal. We define γ_{A} and γ_{N} as the probabilities that an M allele is transmitted from a heterozygous parent to, respectively, an affected or unaffected sibling. Good approximations to the sampling distributions of $Y_{\cdot\cdot}$ and $X_{\cdot}$ are $Y_{\cdot\cdot} - c_{\text{N}} \sim \mathrm{Bi}(nh,\gamma_{\text{N}})$ and $X_{\cdot} - c_{\text{A}} \sim \mathrm{Bi}(h,\gamma_{\text{A}})$, respectively, where h is the number of heterozygous parents in the sample and c_{N} and c_{A} are constants determined by the number of MM parents in the sample. The approximation arises because the alleles transmitted from parents to a particular child are not independent conditional on the disease status of the child (Bickeböller and Clerget-Darpoux 1995), but it is adequate for most complex diseases and is exactly true for multiplicative disease models (e.g., see Whittaker et al. 1998). Thus,

$$\mathrm{Var}(X_{\cdot}) = h\gamma_{\text{A}}(1-\gamma_{\text{A}}) = h\left[\frac{1}{4} - \left(\gamma_{\text{A}} - \frac{1}{2}\right)^2\right]$$

and

$$\mathrm{Var}(Y_{\cdot\cdot}) = nh\gamma_{\text{N}}(1-\gamma_{\text{N}}) = nh\left[\frac{1}{4} - \left(\gamma_{\text{N}} - \frac{1}{2}\right)^2\right]\ .$$

For complex disease models, γ_{A} and γ_{N} will be sufficiently close to $\frac{1}{2}$ that $\mathrm{Var}(Y_{\cdot\cdot}) \approx n\mathrm{Var}(X_{\cdot})$ and

$$\mathrm{Var}(T_{\text{TDT}}) = \frac{1}{\sigma^2}\mathrm{Var}(X_{\cdot}) \approx \frac{1}{n\sigma^2}\mathrm{Var}(Y_{\cdot\cdot}) = \mathrm{Var}(T'_{\text{TDT}})\ .$$

Conditional on parental genotypes, $X_{\cdot}$ and $Y_{\cdot\cdot}$ are independent—and, hence, T_{TDT} and T'_{TDT} are independent—and

$$\mathrm{Var}(T_{\text{S-TDT}}) = \frac{1}{n+1}[n\mathrm{Var}(T_{\text{TDT}}) + \mathrm{Var}(T'_{\text{TDT}})] \approx \mathrm{Var}(T_{\text{TDT}})\ ,$$

as required.

We have shown that $\mathbf{E}(T_{\text{S-TDT}}) \approx \sqrt{n/(n+1)}\,\mathbf{E}(T_{\text{TDT}})$ and that T_{TDT} and $T_{\text{S-TDT}}$ have approximately equal variances. When the standard formula for power (e.g., see Risch and Merikangas 1996) is used, it follows that, if the two tests are to have the same power, then, for the TDT, we require $n/(n+1)$ as many families with a single affected and n unaffected offspring as are required for the S-TDT.

The S-TDT is required only when parental genotypes are missing—and, hence, when σ_0 is unknown and must be estimated from the sib data. Spielman and Ewens (1998) have provided an estimator based on their permutation procedure; an alternative would be to use the sample SD of $X_i - (X_i + Y_{i\cdot})/(n + 1)$. For large sample sizes the distribution of $T_{\text{S-TDT}}$ is well approximated by the standard normal, whereas for small sample sizes exact tests should be used. The results given above depend on the assumption that the probability that an M allele is transmitted from a heterozygous parent to an unaffected sib is .5. This probability is actually slightly <.5, so that $\mathbf{E}(Y_{\cdot\cdot})$ is slightly less than $n\mu_0$, and the formula above slightly understates the power of the S-TDT; but, for complex diseases, the discrepancy is insufficient to be of practical importance.

These results allow us to evaluate the optimum number of unaffected sibs to genotype if multiple unaffected sibs are available. Using only one unaffected sibling ($n = 1$) will require twice the number of families as is required for the TDT. Two unaffected siblings ($n = 2$) give the same genotyping load per family as is given for the TDT but require 50% more families to achieve the same power. These results can also be used to consider the trade-off between genotyping costs and power. For example, for a specific number of genotypes, maximum power is obtained for the S-TDT by inclusion of only one unaffected sibling per family.

The loss of power in the S-TDT may be severe, particularly if only a single unaffected sib is available. Also, of course, families with no unaffected sibs cannot be used in the S-TDT. However, the extension of TDT to sibling-based sampling will allow family-based association testing to be performed for late-onset diseases when parental samples are not available. In this case, the loss of power will be outweighed by the utility of the study design. An overall assessment of design of a study can be made, allowing for the availability of different family members and for costs of family ascertainment, phenotype testing and genotyping.

Acknowledgments

This work was supported in part by the Special Trustees of Guy's Hospital. We would like to thank Robert Curnow and David Balding for their helpful comments on an earlier draft of this article.

John. C. Whittaker[1] and Cathryn M. Lewis[2]
[1]Department of Applied Statistics, University of Reading, Reading, United Kingdom; and [2]Division of Medical and Molecular Genetics, The Guy's, King's and St. Thomas' School of Medicine, King's College London, London

References

Bickeböller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidemiol 12:577–582

Boehnke M , Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

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

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:516–1517

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Whittaker JC, Morris A, Curnow RN (1998) Using information from both parents when testing for association between marker and disease loci. Genet Epidemiol 15:193–200

Address for correspondence and reprints: Dr. Cathryn M. Lewis, Division of Medical and Molecular Genetics, 8th Floor, Guy's Tower, Guy's Hospital, London SE1 9RT, United Kingdom. E-mail: cathryn.lewis@kcl.ac.uk

0002-9297/99/6502-0039$02.00