Am. J. Hum. Genet. 69:371–380, 2001

A Transmission/Disequilibrium Test That Allows for Genotyping Errors in the Analysis of Single-Nucleotide Polymorphism Data

Derek Gordon,[1] Simon C. Heath,[2] Xin Liu,[1] and Jürg Ott[1]

[1]Laboratory of Statistical Genetics, Rockefeller University, and [2]Laboratory of Mathematical Genetics, Memorial Sloan-Kettering Cancer Center, New York

The present study assesses the effects of genotyping errors on the type I error rate of a particular transmission/disequilibrium test (TDT_{std}), which assumes that data are errorless, and introduces a new transmission/disequilibrium test (TDT_{ae}) that allows for random genotyping errors. We evaluate the type I error rate and power of the TDT_{ae} under a variety of simulations and perform a power comparison between the TDT_{std} and the TDT_{ae}, for errorless data. Both the TDT_{std} and the TDT_{ae} statistics are computed as two times a log-likelihood difference, and both are asymptotically distributed as χ^2 with 1 df. Genotype data for trios are simulated under a null hypothesis and under an alternative (power) hypothesis. For each simulation, errors are introduced randomly via a computer algorithm with different probabilities (called "allelic error rates"). The TDT_{std} statistic is computed on all trios that show Mendelian consistency, whereas the TDT_{ae} statistic is computed on all trios. The results indicate that TDT_{std} shows a significant increase in type I error when applied to data in which inconsistent trios are removed. This type I error increases both with an increase in sample size and with an increase in the allelic error rates. TDT_{ae} always maintains correct type I error rates for the simulations considered. Factors affecting the power of the TDT_{ae} are discussed. Finally, the power of TDT_{std} is at least that of TDT_{ae} for simulations with errorless data. Because data are rarely error free, we recommend that researchers use methods, such as the TDT_{ae}, that allow for errors in genotype data.

Introduction

There is growing interest in the use of single-nucleotide polymorphisms (SNPs) for the genetic dissection of complex human diseases (Collins et al. 1998). Some reasons include the following: (1) SNPs are significantly more abundant than microsatellite polymorphisms (~1 SNP for every 500–1,000 base pairs [Chakravarti 1999]) and therefore are potentially more powerful in detecting linkage in the presence of linkage disequilibrium (LD) around disease loci (Risch and Merikangas 1996); (2) genotyping of SNPs is easier to automate, leading to higher throughput; (3) some SNP mutations may be causative of disease phenotypes; and (4) the completion of the human genome reference sequence should pave the way for discovery of many of the common polymorphisms (Collins et al. 1998).

To take advantage of the greater LD that is expected between SNP loci and disease loci, population-based tests of LD (case-control studies) and family-based tests of linkage and LD (transmission/disequilibrium tests [TDTs]) are being considered for data analysis (Risch and Merikangas 1996; Schork et al. 2001). In the present study, we focus on family-based tests. Much work has been done to determine the statistical properties of such tests, including the determination of type I error and power under different genetic models of disease (Schaid 1996; Sham 1998; Xiong and Guo 1998). However, it is almost always assumed in these analyses that the genetic data are without errors. By "errors," we mean any miscoding of a person's correct marker genotype. Sources of error include nonpaternity, sample swaps in the lab, or genotyping errors. In this work, we focus on random genotyping errors.

Whereas much has been written about methods of error detection (Lincoln and Lander 1992; Brzustowicz et al. 1993; Ott 1993; Lunetta et al. 1995; Ehm et al. 1996; Stringham and Boehnke 1996; Ghosh et al. 1997; O'Connell and Weeks 1998, 1999; Broman 1999; Douglas et al. 2000; Ewen et al 2000; Giordano et al. 2001), there are only a few recent papers (Göring and Terwilliger 2000*a,* 2000*b,* 2000*c,* 2000*d;* Gordon and Ott 2001) that consider methodology allowing for errors in linkage and/or LD analysis, even though it is well known that errors in genetic data can have significant effects on linkage analyses. Such effects include an increase in the estimated recombination fraction be-

Received April 12, 2001; accepted for publication June 13, 2001; electronically published July 5, 2001.

Address for correspondence and reprints: Dr. Derek Gordon, Laboratory of Statistical Genetics, Rockefeller University, Box 192, 1230 York Avenue, New York, NY 10021. E-mail: gordon@linkage.rockefeller.edu

0002-9297/2001/6902-0013$02.00

Table 1

List of All Possible Values for Function *TrP*

True Recoded Trio (G)	$TrP(G, t)$
(*0, 1, 0*)	t
(*1, 0, 0*)	t
(*1, 2, 1*)	t
(*2, 1, 1*)	t
(*0, 1, 1*)	$1 - t$
(*1, 0, 1*)	$1 - t$
(*1, 2, 2*)	$1 - t$
(*2, 1, 2*)	$1 - t$
(*1, 1, 0*)	t^2
(*1, 1, 2*)	$(1 - t)^2$
(*1, 1, 1*)	$2t(1 - t)$
(*0, 0, 0*)	1
(*0, 2, 1*)	1
(*2, 0, 1*)	1
(*2, 2, 2*)	1

tween markers or between marker and disease (more generally, an inflation of the map distance for multiple markers), an increase in type I error rate, a decrease in power (Ott 1977; Terwilliger et al. 1990; Buetow 1991; Shields et al. 1991; Goldstein et al. 1997; Heath 1998; Gordon et al. 1999*b*), and an incorrect estimation of background LD (Akey et al. 2001). The purposes of the present study, therefore, include (1) the introduction of a new TDT (hereafter known as the "TDT_{ae}," a TDT allowing for errors) that allows for errors in the analysis, (2) the assessment of the effect of random genotyping errors on the type I error rate (rejection of a true null hypothesis) and power (rejection of a false null hypothesis) of the TDT_{ae}, and (3) a comparison between the performance of the TDT_{ae} and a standard TDT (hereafter referred to as "TDT_{std}").

Methods

Error Model

For all our analyses, we assume that the SNP locus has two alleles, coded as "*1*" and "*2*." We also assume that each *1* allele has a constant probability ε_1 of being incorrectly coded as a *2* allele, and, likewise, each *2* allele has a constant probability ε_2 of being incorrectly coded as a *1* allele. We choose this error model because it is straightforward, it easily allows for the computation of the probability Pr(observed genotype|true genotype) for any person's true and observed genotypes, and it has been studied elsewhere for population-based tests of LD (Gordon and Ott 2001). In addition, it is reasonable to expect that high-throughput automated SNP genotyping technology will have such random errors, as is the case with a number of automated processes (Box et al. 1978; Wang et al. 1998). Finally, we note that it is straightforward to compute our new TDT statistic through use of this error model.

Statistical Tests

The statistics considered for the null simulations are (1) a likelihood-based version of the TDT (TDT_{std}) (Terwilliger and Ott 1992; Spielman et al. 1993) and (2) the TDT allowing for errors (TDT_{ae}). Matise (1995) showed that the TDT_{std} performed in a way equivalent to the TDT proposed by Spielman et al. (1993), in terms of power and type I error for simulated genotype data from multiallelic loci. The sampling frame for this test is a trio of individuals (father, mother, and child). We present the TDT_{ae} statistic first, since the TDT_{std} statistic is just a special case of the TDT_{ae} statistic, in which the error rates ε_1 and ε_2 are each set to 0.

For notational simplicity, let "*0,*" "*1,*" and "*2*" represent the genotypes *1/1, 1/2,* and *2/2,* respectively. We shall hereafter refer to these values as the "recoded genotypes." Let the symbols O_{ijk} and T_{ijk} represent the observed and true trio, respectively, of genotypes in which the father has genotype i, the mother has genotype j, and the affected child has genotype k. For example, O_{001} is an observed trio in which father and mother both have recoded genotype *0* (i.e., genotype *1/1*), and the child has recoded genotype *1* (i.e., genotype *1/2*). Similarly, T_{112} is a true trio in which the father and mother both have recoded genotype *1* (i.e., genotype *1/2*) and the child has a recoded genotype *2* (i.e., genotype *2/2*). Since we are allowing for errors, there are $3 \times 3 \times 3 = 27$ possible sets of subscripts for O, but each set of subscripts for T must be consistent with Mendel's laws, so there are only 15 possible configurations for T. The complete list of 15 configurations may be found in table 1. From this point forward, we shall use the terms "consistent" and "consistency" to mean, respectively, consistent with Mendel's laws and a trio that is consistent with Mendel's laws.

Let $P_{ij}(\varepsilon_1,\varepsilon_2) = \Pr(\text{observing } i \text{ recoded genotype} \mid \text{true recoded genotype} = j)$. Note that P_{ij} is a function of the error rates. P_{ij} is often referred to as a "penetrance function." Also note that there are $3 \times 3 = 9$ possible values

Table 2

Probabilities (or Penetrances) P_{ij}, for All Pairs of Observed Recoded Genotypes and True Recoded Genotypes

Observed Recoded Genotype	True Recoded Genotype		
	0	*1*	*2*
0	$(1 - \varepsilon_1)^2$	$\varepsilon_2(1 - \varepsilon_1)$	ε_2^2
1	$2\varepsilon_1(1 - \varepsilon_1)$	$\varepsilon_1\varepsilon_2 + (1 - \varepsilon_1)(1 - \varepsilon_2)$	$2\varepsilon_2(1 - \varepsilon_2)$
2	ε_1^2	$\varepsilon_1(1 - \varepsilon_2)$	$(1 - \varepsilon_2)^2$

for P_{ij}, and these values are listed in table 2. Furthermore, let I_{ijk} be the Mendelian indicator function, so that

$$I_{ijk} = \begin{cases} 1, & \text{if trio } ijk \text{ shows consistency} \\ 0, & \text{if trio } ijk \text{ does not show consistency} \end{cases} .$$

The genotype-frequency function, $GF(i, p_{11}, p_{12})$, is defined by

$$GF(i, p_{11}, p_{12}) = \begin{cases} p_{11}, & i = 0 \\ p_{12}, & i = 1 \\ 1 - p_{11} - p_{12}, & i = 2 \end{cases} .$$

In this function, i represents a recoded genotype, p_{11} represents the population frequency of the genotype *1/1*, and, likewise, p_{12} represents the population frequency of the genotype *1/2*. Finally, let $TrP(i, j, k, t)$ represent the probability that parents with recoded genotypes i and j transmit a *1* allele to a child with recoded genotype k, where t = Pr(heterozygous parent transmits a *1* allele to child). For example, $TrP(1, 0, 1, t) = 1 - t$. A list of all values of the function TrP is given in table 1.

Given these definitions, we now compute the likelihood of an observed trio of recoded genotypes (*i,j,k*) as a function of the parameters t, p_{11}, p_{12}, ε_1, and ε_2. The likelihood is given by

$$\begin{aligned} L_{ijk}(t, p_{11}, p_{12}, \varepsilon_1, \varepsilon_2) = \sum_{x=0}^{2} \sum_{y=0}^{2} \sum_{z=0}^{2} & P_{ix}(\varepsilon_1, \varepsilon_2) \\ & P_{jy}(\varepsilon_1, \varepsilon_2) P_{kz}(\varepsilon_1, \varepsilon_2) I_{xyz} \\ & GF(x, p_{11}, p_{12}) GF(y, p_{11}, p_{12}) \\ & TrP(x, y, z, t) . \end{aligned} \quad (1)$$

It is important to note that although equation (1) appears to sum over all 27 possible combinations of sets of recoded genotypes, because of the indicator function I_{xyz}, only those sets of recoded genotypes that are consistent are added to the likelihood.

If N_{ijk} represents the number of trios observed in our data set to have recoded genotypes (*i, j, k*), and if ln is the $\log_e$ function, then the overall log-likelihood for an observed data set as a function of the parameters t, p_{11}, p_{12}, ε_1, and ε_2 is

$$\begin{aligned} \ln [L(t, p_{11}, p_{12}, \varepsilon_1, \varepsilon_2)] = \sum_{i=0}^{2} \sum_{j=0}^{2} \sum_{k=0}^{2} & N_{ijk} \\ & \ln [L_{ijk}(t, p_{11}, p_{12}, \varepsilon_1, \varepsilon_2)] . \end{aligned} \quad (2)$$

To compute the TDT_{ae}, we first maximize the log-likelihood equation (2) over all five parameters. In our simulations, we maximize the three parameters t, p_{11}, and p_{12}, over the closed interval [0,1], in increments of .125, discarding any sets of parameters in which $p_{11} + p_{12} > 1$. Also, the error parameters ε_1 and ε_2 are maximized over the closed interval [0, .1], in increments of .0125. Each log-likelihood equation (2) is therefore maximized over $9^5 = 59{,}049$ values. Let the notation $\hat{z}$ represent the maximum-likelihood estimates (MLEs) of any of the five parameters in equation (2)—that is, the estimates that jointly maximize that equation. Next, fix $t = .5$, maximize the log-likelihood equation (2) over the other four parameters, and let the notation $\hat{\hat{z}}$represent those MLEs. Then the TDT_{ae} statistic is given by the formula

$$\begin{aligned} 2 \times \{ & \ln [L(\hat{t}, \hat{\varepsilon}_1, \hat{\varepsilon}_2, \hat{p}_{11}, \hat{p}_{12})] \\ & - \ln [L(.5, \hat{\hat{\varepsilon}}_1, \hat{\hat{\varepsilon}}_2, \hat{\hat{p}}_{11}, \hat{\hat{p}}_{12})] \} . \end{aligned} \quad (3)$$

According to likelihood-ratio theory (Kendall et al. 1991), under the null hypothesis, TDT_{ae} is asymptotically distributed as χ_1^2 (a χ^2 distribution with 1 df). It is important to note that the TDT_{ae} does not require estimates of the error parameters ε_1 and ε_2 to calculate the statistic; rather, it provides estimates of these parameters under the null hypothesis ($t = .5$) and the alternative hypothesis (t maximized jointly over interval [0.0–1.0] with the other four parameters).

For the TDT_{std} statistic, we assume that there are no errors, so that $\varepsilon_1 = \varepsilon_2 = 0$. In this case, equation (1) reduces to

$$\begin{aligned} \tilde{L}_{ijk}(t, p_{11}, p_{12}) = & GF(i, p_{11}, p_{12}) \\ & GF(j, p_{11}, p_{12}) TrP(i, j, k, t) . \end{aligned} \quad (1a)$$

The symbol $\tilde{L}$ in equation (1a) is used to distinguish the likelihoods for TDT_{std} from the likelihoods for TDT_{ae}. Because we assume that there are no errors, the recoded genotypes (*i,j,k*) are all consistent (see table 1). An important consequence of this simplification is that the log likelihood of equation (1a) reduces to

$$\begin{aligned} \ln [\tilde{L}_{ijk}(t, p_{11}, p_{12})] = & \ln [GF(i, p_{11}, p_{12})] \\ & + \ln [GF(j, p_{11}, p_{12})] \\ & + \ln [TrP(i, j, k, t)] . \end{aligned} \quad (1b)$$

Through examination of equation (1b), we note that maximizing over the parameter t is independent of maximizing over the parameters p_{11} and p_{12}. Therefore, when considering the overall log likelihood of the data set, which is given by the formula

$$\ln[\tilde{L}(t, p_{11}, p_{12})] = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{2} N_{ijk} I_{ijk} \ln[\tilde{L}_{ijk}(t, p_{11}, p_{12})] , \quad (2a)$$

and considering the difference of log likelihoods,

$$2 \times \{\ln[\tilde{L}(\hat{t}, \hat{p}_{11}, \hat{p}_{12})] - \ln[\tilde{L}(.5, \hat{\hat{p}}_{11}, \hat{\hat{p}}_{12})]\} , \quad (3a)$$

it follows from equation (1b) that $\hat{p}_{11} = \hat{\hat{p}}_{11}, \hat{p}_{12} = \hat{\hat{p}}_{12}$ and, in fact, the difference (eq. [3a]) is actually independent of the parameters p_{11} and p_{12}. With this understanding, we may rewrite equation (3a) as

$$2 \times \{\ln[\tilde{L}(\hat{t})] - \ln[\tilde{L}(.5)]\} . \quad (3b)$$

We shall refer to the value of equation (3b) as TDT_{std}. Through use of standard calculus techniques, it is possible to solve for the MLE $\hat{t}$ in terms of the number of different observed trios N_{ijk}. Let

$$x_1 = N_{010} + N_{100} + N_{121} + N_{211} + 2N_{110} + N_{111} ,$$
$$x_2 = N_{011} + N_{101} + N_{122} + N_{212} + 2N_{112} + N_{111} .$$

Through use of this notation, the value of t that maximizes the log likelihood $\tilde{L}(t)$ in equation (3b) is

$$\hat{t} = \frac{x_1}{(x_1 + x_2)} . \quad (4)$$

This value of t is used when computing the test statistic TDT_{std}.

Simulations

The data selected for use with the TDT tests consisted of an SNP locus that has two alleles in the population. These alleles were coded as "*1*" and "*2*." Each replicate of each simulation consisted of genotype data from a number of trios (father, mother, and affected child). We simulated genotype data under two models: the null model, in which there was neither linkage nor LD, and the power model, in which the SNP marker locus was linked to the disease locus and there was LD between the marker and the disease locus. We assumed that the disease locus also has two alleles. For the null simulations, we set a recombination fraction (θ) of .5 between marker and disease, and we assumed that all loci were in Hardy-Weinberg equilibrium, so that two-locus haplotype frequencies were just the products of the allele frequencies at each of the two loci. For the power simulations, two-locus haplotype frequencies were completely determined by the allele frequencies at each of the two loci and by an additional parameter, D'. The value D' is related to Lewontin's (1964) D, by the formula $D' = D/\min(p_+p_2, p_dp_1)$, where p_+ and p_d were the allele frequency of the wild-type allele and the disease allele, respectively, at the disease locus, and p_1 and p_2 were the allele frequencies of the *1* allele and *2* allele, respectively, at the SNP marker locus. The values of D' considered for these simulations were .5 and .8. From this point forward, the term "m% LD" (meaning "the %LD is m%") for some integer m and some power simulation means that $D' = .01 \times m$. For example, 20% LD means that $D' = .2$.

In all power simulations, we assumed that disease locus and marker locus were completely linked ($\theta = 0$). We assert that this assumption is reasonable, given the dense coverage that SNPs have throughout the human genome (Chakravarti 1999).

For both the null and power simulations, we considered sample sizes of 100 and 500 trios. Allele frequencies for the marker locus were set either at .5 each (equal allele frequencies) or at .25 for the *1* allele in all simulations. For the LD simulations, we considered allele frequencies of .001 (rare) and .2 (common) for the disease (non–wild type) allele. As above, we used the notations "+" and "*d*" to refer to the wild-type and disease alleles, respectively, at the disease locus.

Genotype data for the null simulations were created by SIMULATE (Terwilliger and Ott 1994; SIMULATE ftp site), and data for the power simulations were created by FASTSLINK (Ott 1989; Weeks et al. 1990; Statgen Software Web site). For the power simulations, there were two modes of inheritance for the disease locus: recessive (a fully penetrant recessive model with no phenocopies), and dominant (a reduced-penetrance dominant model, in which the penetrance of each of the genotypes [at the disease locus] +*d* and *dd* was .6, and the penetrance of the genotype ++ was .02). The recessive disease model was chosen because it has been shown (Terwilliger and Ott 1992) that the TDT statistic is most powerful for such a disease model. The dominant disease locus model was chosen to reflect a more "realistic" disease model for complex diseases.

The pairs of error rates (ε_1, ε_2) we assumed for the marker locus were (.01, .01), (.01, .05), (.05, .01), (.05, .05), (.05, .10), and (.10, .05). We chose these pairs to provide a sense of the performance of the test statistics under a broad range of error rates. Errors were introduced randomly and independently into the genotype data files, by means of a computer program. For each simulation, the proportion of trios that showed consistency was recorded. In table 3, we report the average proportion of trios (over 1,000 replicates) that showed consistency.

Throughout this article, the terms "type I error rate" and "power," at the α% level of significance for a particular statistic, mean the proportion of replicates for a particular null or power simulation, respectively, that

Table 3

Results of Null Simulations

No. of Trios	1-Allele Frequency	ε_1	ε_2	Average Proportion of Consistent Pedigrees	Type I Error Rate			
					TDT_{ae}		TDT_{std}[a]	
					5% Level	1% Level	5% Level	1% Level
100	.25	.01	.01	.984	.054	.006	.051	.011
100	.25	.01	.05	.941	.051	.010	***.102***	***.024***
100	.25	.05	.01	.969	.061	.005	***.085***	***.028***
100	.25	.05	.05	.929	.042	.009	***.112***	***.036***
100	.25	.05	.10	.887	.052	.011	***.174***	***.068***
100	.25	.10	.05	.913	.055	.014	***.185***	***.063***
100	.50	.01	.01	.985	.046	.009	***.065***	.012
100	.50	.01	.05	.958	.055	.009	.056	.007
100	.50	.05	.01	.958	.046	.015	.040	.006
100	.50	.05	.05	.935	.053	.009	.061	.009
100	.50	.05	.10	.906	.054	.015	.059	.014
100	.50	.10	.05	.906	.042	.009	.059	.016
500	.25	.01	.01	.984	.052	.013	***.078***	***.022***
500	.25	.01	.05	.942	.048	.008	***.268***	***.118***
500	.25	.05	.01	.969	.058	.012	***.163***	***.047***
500	.25	.05	.05	.928	.048	.008	***.411***	***.201***
500	.25	.05	.10	.886	.045	.009	***.606***	***.395***
500	.25	.10	.05	.912	.052	.011	***.631***	***.386***
500	.50	.01	.01	.985	.052	.012	***.067***	.010
500	.50	.01	.05	.958	.046	.010	.056	.014
500	.50	.05	.01	.959	.042	.008	.055	.006
500	.50	.05	.05	.932	.052	.014	.041	.008
500	.50	.05	.10	.904	.046	.010	.063	.010
500	.50	.10	.05	.905	.058	.012	.060	.011

[a] Values in boldface italics have 95% CIs that do not contain a set level of significance (5%, 1%), based on the method for establishing CIs (see Results section, Null Simulations).

exceed $\chi_1^2(.01 \times \alpha)$, where $\chi_1^2(.01 \times \alpha)$ refers to the (two-sided) cutoff for a χ^2 statistic with 1 df. The type I error rate at the 5% and 1% levels of significance are reported for null simulations in table 3, and the power at the 5% and 1% levels of significance are reported for power simulations in tables 4 and 5.

Maximization over Three Parameters for Power Simulations

Although it is possible to maximize the likelihood equation (2) over all five parameters, the process is computationally intensive. Therefore, for our power simulations (tables 4 and 5), we assumed that we knew the values of the error rates ε_1 and ε_2 used to generate errors and only maximized the log-likelihood equation (2) over, at most, three parameters. It is true that this assumption has the potential effect of increasing the power of the TDT_{ae} for these simulations, but comparisons of the power from the TDT_{ae} maximizing over three parameters versus five parameters did not show an appreciable increase in power in favor of the three-parameter simulations (data not shown), whereas the reduction in computation time was appreciable (a factor of 9^2).

Power Comparison

The TDT_{ae} statistic has the advantage of allowing for errors in the analysis of SNP genotype data, but at the computational cost of maximizing over five parameters (all parameters but t are nuisance parameters), in contrast to the TDT_{std}, which maximizes over one parameter. In theory, both statistics are asymptotically disturbed as χ_1^2 and, given a dense enough grid search, there is no difference in power between the two methods. In practice, however, the exact maximum likelihood for the TDT_{ae} is most likely not achieved when maximizing over the five parameters, because of computational limitations. For the TDT_{std}, the maximum likelihood is always achieved through use of the value of $\hat{t}$ in equation (4).

To assess the effect that maximization over an additional four parameters has on the power of the TDT_{ae}, we performed power simulations in which there are no errors in the genotype data created. Each simulation is determined by two factors: the number of trios simulated (100, 200, 500, and 1000) and the *1*-allele frequency at the marker locus (.25, .50). In all simulations, the disease-allele frequency (p_d) was .20, θ between the disease and the marker locus was 0, and the %LD was 50. For

Table 4

Results of Power Simulations with TDT_{ae} for a Fully Penetrant Recessive Disease Model, 50% LD, and a .001 Disease-Allele Frequency

No. of Trios	*1*-Allele Frequency	ε_1	ε_2	Power	
				5% Level	1% Level
100	.25	.01	.01	.916	.742
100	.25	.01	.05	.768	.528
100	.25	.05	.01	.879	.681
100	.25	.05	.05	.749	.500
100	.25	.05	.10	.537	.302
100	.25	.10	.05	.660	.419
100	.50	.01	.01	1.000	.995
100	.50	.01	.05	.999	.981
100	.50	.05	.01	.999	.988
100	.50	.05	.05	.990	.956
100	.50	.05	.10	.950	.858
100	.50	.10	.05	.972	.904
500	.25	.01	.01	1.000	1.000
500	.25	.01	.05	1.000	.999
500	.25	.05	.01	1.000	1.000
500	.25	.05	.05	.999	.998
500	.25	.05	.10	.990	.958
500	.25	.10	.05	.999	.992
500	.50	.01	.01	1.000	1.000
500	.50	.01	.05	1.000	1.000
500	.50	.05	.01	1.000	1.000
500	.50	.05	.05	1.000	1.000
500	.50	.05	.10	1.000	1.000
500	.50	.10	.05	1.000	1.000

each simulated data set (replicate) in each simulation, the TDT_{std} and TDT_{ae} were computed as described above. A total of 250 replicates were created and analyzed for each simulation. Power curves for each method are presented in figure 1.

Results

Null Simulations

Table 3 presents a summary of the results for our null simulations. Each row records the number of trios considered, the error rates ε_1 and ε_2, the frequency of the *1* allele at the marker locus, the average proportion of consistent trios in each replicate (averaged over 1,000 replicates), and the type I error rates at the 5% and 1% levels of significance for the TDT_{ae} and TDT_{std} statistics. We indicate, in boldface italic type, those simulations for which the 95% confidence interval (CI) (Fisher 1960) of the type I error rate does not include the chosen significance level, indicating that the test statistic showed an inflation in type I error for this particular simulation.

From studying table 3, we see that the TDT_{ae} statistic maintains a correct type I error rate in all simulations, for each of the significance levels (5% and 1%). On the other hand, the TDT_{std} statistic shows an inflation in the type I error rate for a number of simulations. In fact, with the exception of the (.01, .01) pair of error rates, the TDT_{std} always shows inflation in type I error when the *1*-allele frequency is .25. As a way of comparing the increases in type I error across different levels of significance, we consider the ratios (type I error rate at 5% level)/.05 and (type I error rate at 1% level)/.01. The largest ratio occurs for 500 trios, a *1*-allele frequency of .25, the pair of error rates (.10, .05), and a significance level of 1%. Under these conditions, we see a ratio of 38.6, an ~40-fold increase in type I error.

For the equal allele frequency case (*1*-allele frequency = .5), the TDT_{std} statistic shows a small inflation in the type I error rate when the pair of error rates is (.01, .01), for both the 100- and the 500-trio case. However, it maintains a correct type I error rate for all other simulations of the equal allele frequency case. We hypothesize that, for our error model, TDT_{std} maintains a correct type I error rate only when marker-allele frequencies are equal.

Intuitively, the reason for an increase in type I error rate for the TDT_{std} test statistic for unequal allele frequencies seems clear. When the allele frequencies are more divergent—as opposed to more equal—there are more trios in which both parents are homozygous. When errors are introduced into trios in which both parents are homozygous and the resultant trio is consistent, the observed (and incorrect) trio of genotypes is counted in

Table 5

Results of Power Simulations with TDT_{ae} for a Reduced Penetrance Dominant Model, 80% LD, and a .001 Disease-Allele Frequency

No. of Trios	*1*-Allele Frequency	ε_1	ε_2	Power	
				5% Level	1% Level
100	.25	.01	.01	.169	.075
100	.25	.01	.05	.116	.037
100	.25	.05	.01	.154	.060
100	.25	.05	.05	.129	.039
100	.25	.05	.10	.097	.033
100	.25	.10	.05	.117	.031
100	.50	.01	.01	.404	.190
100	.50	.01	.05	.335	.165
100	.50	.05	.01	.398	.193
100	.50	.05	.05	.287	.124
100	.50	.05	.10	.228	.097
100	.50	.10	.05	.243	.092
500	.25	.01	.01	.602	.301
500	.25	.01	.05	.377	.160
500	.25	.05	.01	.542	.282
500	.25	.05	.05	.356	.161
500	.25	.05	.10	.276	.107
500	.25	.10	.05	.330	.151
500	.50	.01	.01	.985	.927
500	.50	.01	.05	.962	.809
500	.50	.05	.01	.975	.854
500	.50	.05	.05	.928	.713
500	.50	.05	.10	.774	.577
500	.50	.10	.05	.834	.628

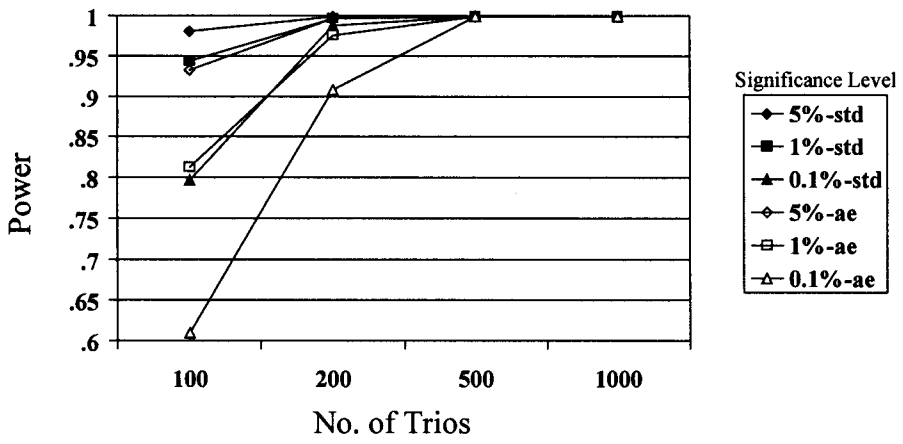

Figure 1 Average power for the TDT_{std} and TDT_{ae} statistics, for errorless data sets in which the *1*-allele frequency = .25, p_d = .2, and %LD = 50, and for which the disease-locus model is a fully penetrant recessive locus. Each average is computed over 250 replicates. The suffixes "-std" and "-ae" in the Significance Level legend indicate power for the TDT_{std} and TDT_{ae} statistics, respectively.

the estimation of the parameter *t*, introducing a bias in *t* away from its true value of .5. In addition, although it seems counterintuitive that the TDT_{std} type I error rates are so much greater for a larger sample than for a smaller one, we comment that, even though the *percentage* of trios that show Mendelian consistency is the same, on average, for the same error rates in the 100- or 500-trio cases, the actual *number* of trios that have errors and that show Mendelian consistency is approximately five times as large in the 500-trio case as in the 100-trio case. As mentioned above, when marker-allele frequencies are unequal, a significant number of these trios (specifically, the trios in which true homozygous parents are incorrectly coded as heterozygous) will be counted in the estimation of the parameter *t*. For the 500-trio case, we expect that five times as many such trios are counted as for the 100-trio case, thus increasing the type I error rate, as was observed in the simulation results.

Because of the significant and consistent increase in type I error rate of the TDT_{std} statistic for unequal allele frequencies, we conclude that it is not a useful test in the presence of errors. For our power simulations, we therefore focus on the TDT_{ae} statistic. Finally, we note that, for all simulations considered, the average proportion of trios that display consistency is $\geq$88%, and this average increases to 93% when error rates of $\leq$.05 are considered.

Power Simulations

Here we present tables for some of our simulations and discuss results for all of the simulations. The number of tables has been reduced, to conserve space. Table 4 provides simulation results for the fully penetrant recessive disease model with 50% LD and a .001 disease-allele frequency. Table 5 presents simulation results for the reduced-penetrance dominant disease with 80% LD and a .001 disease-allele frequency.

In studying these two tables, we notice some common patterns. First, we notice, for each sample size (100 or 500 trios) and each set of allele frequencies (*1*-allele frequency = .25 or .5), that, as the values of ε_1 and ε_2 increase, the power of the TDT_{ae} statistic decreases. This decrease due to larger error rates is to be expected, since an increase in the value of the error rates decreases the probability that any single true trio of genotypes is associated with a given observed trio of genotypes. With regard to specific error rates, it is interesting to note that power was reduced most when ε_2 was largest—that is, when error was introduced into the *2* allele, which is the allele in coupling with the disease allele *d*. A second observation is that, for each sample size and each pair of error rates, the power of the TDT_{ae} is greater when marker-allele frequencies are equal, as opposed to when the *1*-allele frequency is .25. A third observation is that, particularly for the dominant mode of inheritance, to achieve any kind of power with the TDT_{ae} (say, power >.60), one needs large sample sizes ($\geq$500 trios) and small error rates. In fact, for sample sizes of 500 trios with a dominant mode of inheritance (table 5), power is >.6 at the 5% level for 7/12 simulations and at the 1% level for 5/12 simulations.

As mentioned in the Methods section, power simulations were also performed for the recessive mode of inheritance, in which the disease-allele frequency p_d was .20 and LD was 50%, as well as for conditions of 80% LD with two sets of disease-allele frequencies (p_d = .001 or .20). For the case of 50% LD, p_d = .20, power at the 5% level was .69–1.00, and power at the 1% level was .46–1.00. With one exception (100 trios, ε_1 = .05 and ε_2 = .10), power at the 5% and 1% levels was

>.8 and >.62, respectively. When the sample size was 500 trios, all power (at the 5% and 1% levels) was >.99. For recessive simulations in which there was 80% LD, power at the 5% level was .93–1.00, and power at the 1% level was .78–1.00. As was the case for 50% LD and $p_d = .20$—with one exception (100 trios, $\varepsilon_1 = .05$, $\varepsilon_2 = .10$, $p_d = .001$)—power at the 5% and 1% levels was >.98 and >.93, respectively.

For those power simulations with a dominant mode of inheritance for the disease locus that are not reported in table 5 (50% LD with $p_d = .001$ or .20, and 80% LD with $p_d = .20$), we report the following ranges of power observed: for the case of 50% LD ($p_d = .001$ or .20), power at the 5% level was .07–.99, and power at the 1% level was .02–.98. In general, for sample sizes of 100 trios, power was low. In fact, the largest power observed at the 5% level for 100 trios was .60. For simulations in which there was 80% LD and $p_d = .20$, power at the 5% level was .19–1.00, and power at the 1% level was .03–1.00. As mentioned above, in *all* cases in which other variables (%LD, allele frequencies) were fixed, power was lowest when the error rates were largest.

An overall observation that can be made about the TDT_{ae} statistic on the basis of these simulations is that the factors that influence the power of this statistic are sample size, mode of inheritance of the disease locus, marker- and disease-allele frequencies, %LD, and, for this analysis, error rates ε_1 and ε_2. The same factors affect the power of the TDT_{std} (Xiong and Guo 1998); however, the addition of errors into genotype data has the adverse effect of decreasing the power in comparison with the power in the errorless data situation.

Power Comparison

Figure 1 presents the results of the power-comparison simulations. The vertical axis is the power at the α% level of significance for the TDT_{std} and TDT_{ae} statistics for simulations, considering four different numbers of trios: 100, 200, 500, and 1,000. What we glean from this graph is that, for this simulation, the difference in power is dependent on both the number of trios considered and the level of significance. The greatest power difference is almost .20 when there are 100 trios, and the level of significance is 0.10%. Note that, at the 5% level of significance, the greatest power difference is .05, when the number of trios is 100. Another observation made from this graph is that, when the sample size is large enough ($\geq$500), there is essentially no difference in power between the TDT_{std} and TDT_{ae} statistics.

We also performed simulations in which the marker-allele frequencies were equal. The result of those simulations was that there was no difference in power between the TDT_{std} and TDT_{ae} statistics, for any number of trios or any level of significance. Both statistics had a power of 1.0 for all simulations.

Summary and Discussion

The purposes of the present study included the assessment of the effects of genotyping errors on a particular TDT (TDT_{std}) and on a new TDT (TDT_{ae}) that allows for random genotyping errors in the analysis, the evaluation of power for the TDT_{ae} under a variety of scenarios, and a power comparison between the TDT_{std} and the TDT_{ae}, when no errors are present in the genotype data. The results indicated that the TDT_{std}, when applied to data that have been "cleaned" (i.e., data in which inconsistent trios are removed), does not maintain the correct type I error rate and that the type I error increases both with an increase in sample size and with an increase in the error rates $\varepsilon_i (i \in \{1, 2\})$. In contrast to this, the TDT_{ae} statistic maintains a correct type I error rate for the simulations considered. The power of the TDT_{ae} is dependent on individual error rates, mode of inheritance at the disease locus, allele frequencies at the disease and marker locus, %LD, and sample size. Finally, we note that, for simulations in which no genotyping errors are introduced, the power of the TDT_{std} statistic is at least that of the TDT_{ae} statistic, but this difference in power decreases as the sample size increases.

It is important to note that the TDT_{ae} is designed for application to data sets in which inconsistent trios are observed. We comment that the TDT_{ae} probably does not maintain a correct type I error rate when applied to data in which inconsistent trios have been removed. Another way of saying this is that *the TDT_{ae} statistic should be applied only to data sets that are "raw"—that is, data sets in which inconsistent trios (if any exist) are not removed when computing the test statistic.*

An interesting result of the present study, although not its main focus, is that even when error rates are relatively high ($\varepsilon_1, \varepsilon_2 \geq .05$), most (>88%) trios will display consistency (table 3). This finding agrees with the analytic solutions of Gordon et al. (1999*a*, 1999*b*). For example, Gordon et al. (1999*a*) showed that, for $\varepsilon_1 = \varepsilon_2 = .05$, on average, >90% of trios will show consistency when marker-allele frequencies are equal, there is no linkage between marker and disease locus, and the marker locus is in Hardy-Weinberg equilibrium.

The error model assumed in the present study is based on an assumption of random errors. A question that arises is whether this assumption is reasonable. For SNP data, this question will be answered more conclusively as more SNP genotype data are created and analyzed. From a statistical viewpoint, however, the real question is whether statistics like the TDT_{ae} are robust to different error models. This research is work in progress.

Because of the potential increase in power that hap-

lotype methods have over single-locus methods (Dudbridge et al. 2000; Xiong et al. 2000) and because of the widespread use of microsatellite markers for linkage and LD analysis (e.g., Lee et al. 2001), a natural question to ask is whether the TDT_{ae} method can be extended to a test using multi-locus haplotypes and/or multi-allelic markers. Perhaps the main challenge of such extensions is the use of as few parameters for error rates as is possible. For example, with n alleles at a marker, the number of possible individual error rates ε_i is $n(n-1)$. We suspect that, for highly polymorphic loci, the best approach may be the one recommended by Schaid (1996) and Spielman and Ewens (1998), which involves down-coding of alleles and performance of multiple two-allele tests. We plan to pursue this research.

Finally, we note that, as is the case with all statistics applied to genotype data, a reassessment of power would be needed for whole-genome scans. We plan to make software available shortly that computes the TDT_{ae} statistic. The code will be freely available.

Acknowledgments

The authors gratefully acknowledge National Institutes of Health grants K01-HG00055-01 and MH59492. Also, Mark A. Levenstien is gratefully acknowledged for computer programs he wrote, which cross-validated the results in tables 3–5. Finally, the authors gratefully acknowledge anonymous reviewers for their helpful comments.

Electronic-Database Information

The URLs for data in this article are as follows:

SIMULATE ftp site, ftp://linkage.rockefeller.edu/software/simulate/ (for SIMULATE software)

Statgen Software, http://watson.hgen.pitt.edu/register/soft_doc.html (for FASTSLINK software)

References

Akey JM, Zhang K, Xiong M, Doris P, Jin L (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet 68: 1447–1456

Box GEP, Hunter WG, Hunter JS (1978) Statistics for experimenters. John Wiley & Sons, New York

Broman KW (1999) Cleaning genotype data. Genet Epidemiol 17 Suppl 1:S79–S83

Brzustowicz LM, Merette C, Xie X, Townsend T, Gilliam C, Ott J (1993) Molecular and statistical approaches to the detection and correction of errors in genotype databases. Am J Hum Genet 53:1137–1145

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49: 985–994

Chakravarti A (1999) Population genetics: making sense out of sequence. Nat Genet 21 Suppl:56–60

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229–1231

Douglas JA, Boehnke M, Lange K (2000) A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am J Hum Genet 66:1287–1297

Dudbridge F, Koeleman BPC, Todd JA, Clayton DG (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 66:2009–2012

Ehm MG, Kimmel M, Cottingham RW Jr (1996) Error detection for pedigree data, using likelihood methods. Am J Hum Genet 58:225–234

Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW, Foote SJ (2000) Identification and analysis of error types in high-throughput genotyping. Am J Hum Genet 67: 727–736

Fisher RA (1960) The design of experiments. Oliver and Boyd, Edinburgh

Ghosh S, Karanjawala ZE, Hauser ER, Ally D, Knapp JI, Rayman JB, Musick A, Tannenbaum J, Te C, Shapiro S, Eldridge W, Musick T, Martin C, Smith JR, Carpten JD, Brownstein MJ, Powell JI, Whiten R, Chines P, Nylund SJ, Magnuson VL, Boehnke M, Collins FS (1997) Methods for precise sizing, automated binning of alleles, and reduction in large-scale genotyping using fluorescently labelled dinucleotide markers: FUSION (Finland-US Investigation of NIDDM Genetics) study group. Genome Res 7:165–178

Giordano M, Mellai M, Hoogendoorn B, Momigliano-Richiardi P (2001) Determination of SNP allele frequencies in pooled DNAs by primer extension genotyping and denaturing high-performance liquid chromatography. J Biochem Biophys Methods 47:101–110

Goldstein DR, Zhao H, Speed TP (1997) The effects of genotyping errors and interference on estimation of genetic distance. Hum Hered 47:86–100

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

Gordon D, Matise TC, Heath SC, Ott J (1999*b*) Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. Genet Epidemiol 17 Suppl 1: S587–S592

Gordon D, Ott J (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput 2001:18–29

Göring HHH, Terwilliger JD (2000*a*) Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. Am J Hum Genet 66: 1095–1106

——— (2000*b*) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. Am J Hum Genet 66:1107–1118

——— (2000*c*) Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. Am J Hum Genet 66:1298–1309

——— (2000*d*) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when

the mode of inheritance cannot be accurately specified. Am J Hum Genet 66:1310–1327

Heath SC (1998) A bias in TDT due to undetected genotyping errors. Am J Hum Genet Suppl 63:A292

Kendall MG, Stuart A, Ord JK (1991) Kendall's advanced theory of statistics, vol 2A, 2d ed. Oxford University Press, New York

Lee MH, Gordon D, Ott J, Lu K, Ose L, Miettinen T, Gylling H, Stalenhoef AF, Pandya A, Hidaka H, Brewer B Jr, Kojima H, Sakuma N, Pegoraro R, Salen G, Patel SB (2001) Fine mapping of a gene responsible for regulating dietary cholesterol absorption: founder effects underlie cases of phytosterolaemia in multiple communities. Eur J Hum Genet 9: 375–384

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

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14:604–610

Lunetta KL, Boehnke M, Lange K, Cox DR (1995) Experimental design and error detection for polyploid radiation hybrid mapping. Genome Res 5:151–163

Matise TC (1995) Genome scanning for complex disease genes using the transmission/disequilibrium test and haplotype-based haplotype relative risk. Genet Epidemiol 12:641–645

O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266

——— (1999) An optimal algorithm for automatic genotype elimination. Am J Hum Genet 65:1733–1740

Ott J (1977) Linkage analysis with misclassification at one locus. Clin Genet 12:110–124

——— (1989) Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 86:4175–4178

——— (1993) Detecting marker inconsistencies in human gene mapping. Hum Hered 43:25–30

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

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

Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob HJ, Cohen D (2001) The future of genetic case-control studies. Adv Genet 42:191–212

Sham P (1998) Statistics in human genetics. J Wiley & Sons, New York

Shields DC, Collins A, Buetow KH, Morton NE (1991) Error filtration, interference, and the human linkage map. Proc Natl Acad Sci USA 88:6501–6505

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

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

Stringham HM, Boehnke M (1996) Identifying marker typing incompatibilities in linkage analysis. Am J Hum Genet 59: 946–950

Terwilliger J, Ott J (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 42:337–346

——— (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Terwilliger JD, Weeks DE, Ott J (1990) Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimates of the recombination fraction. Am J Hum Genet Suppl 47:A201

Xiong M, Akey J, Jin L (2000) The haplotype linkage disequilibrium test for genome-wide screens: its power and study design. Pac Symp Biocomput 2000:675–686

Xiong M, Guo SW (1998) The power of linkage detection by the transmission disequilibrium tests. Hum Hered 48:295–312

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Weeks DE, Ott J, Lathrop GM (1990) SLINK: a general simulation program for linkage analysis. Am J Hum Genet Suppl 47:A204