# Undetected Genotyping Errors Cause Apparent Overtransmission of Common Alleles in the Transmission/Disequilibrium Test

Adele A. Mitchell, David J. Cutler, and Aravinda Chakravarti

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore

The transmission/disequilibrium test (TDT), a family-based test of linkage and association, is a popular and intuitive statistical test for studies of complex inheritance, as it is nonparametric and robust to population stratification. We carried out a literature search and located 79 significant TDT-derived associations between a microsatellite marker allele and a disease. Among these, there were 31 (39%) in which the most common allele was found to exhibit distorted transmission to affected offspring, implying that the allele may be associated with either susceptibility to or protection from a disease. In 27 of these 31 studies (87%), the most common allele appeared to be overtransmitted to affected offspring (a risk factor), and, in the remaining 4 studies, the most common allele appeared to be undertransmitted (a protective factor). In a second literature search, we identified 92 case-control studies in which a microsatellite marker allele was found to have significantly different frequencies in case and control groups. Of these, there were 37 instances (40%) in which the most common allele was involved. In 12 of these 37 studies (32%), the most common allele was enriched in cases relative to controls (a risk factor), and, in the remaining 25 studies, the most common allele was enriched in controls (a protective factor). Thus, the most common allele appears to be a risk factor when identified through the TDT, and it appears to be protective when identified through case-control analysis. To understand this phenomenon, we incorporated an error model into the calculation of the TDT statistic. We show that undetected genotyping error can cause apparent transmission distortion at markers with alleles of unequal frequency. We demonstrate that this distortion is in the direction of overtransmission for common alleles. Therefore, we conclude that undetected genotyping errors may be contributing to an inflated false-positive rate among reported TDT-derived associations and that genotyping fidelity must be increased.

## Introduction

Much attention has recently been paid to methods for predicting genotyping error rates, studying types of errors that can be expected with various technologies, detecting the presence of such errors, predicting the effect of undetected errors on genetic analysis, and developing analytical methods that are robust to such errors. Inconsistencies in genotype data can occur for many reasons, including sample mix-up, pedigree errors such as nonpaternity or incorrectly specified relationships, technician error, and technology failure. In the context of a genome scan or in the analysis of many unlinked candidate genes or of haplotypes, misspecified relationships can usually be detected. In this analysis, we focus only on random errors that have occurred because of imperfect genotyping technologies and manual mistakes that result in an incorrect genotype call at a single marker.

Different types of errors are encountered with different types of markers. For example, with single-nucleotide polymorphisms (SNPs), heterozygotes are believed to be more difficult to call than homozygotes, leading to allelic dropout (Cutler et al. 2001). This is because SNP genotyping technology often involves carrying out two reactions per locus. If an individual is heterozygous and one of the two reactions fails, the individual will appear homozygous. Microsatellite genotypes, on the other hand, are prone to both call errors and missed alleles (Ewen et al. 2000; Sobel et al. 2002). Call errors can result from errors in sample loading, low fluorescence, bleedthrough fluorescence from other markers run in the same lane, or inconsistencies in genotype scoring, particularly when alleles are called by eye. In addition, errors can be introduced by repeat expansion during PCR (Clarke et al. 2001), particularly for dinucleotide repeats.

A fraction of genotyping errors can be detected by checking for the presence of deviations from Mendelian inheritance in families. Rates of genotyping error detection through Mendelian inheritance vary with pedigree structure, the number of alleles at a marker, and the frequencies of the alleles; errors can be detected more easily in highly polymorphic microsatellite mark-

ers than in SNPs (Douglas et al. 2002). Through simulation, true error rates in SNPs have been shown to be approximately three to four times the Mendelian-detectable error rate (Gordon et al. 1999). Additional errors can be detected by investigating apparent double recombinants within small regions when multipoint data are available and haplotypes can be constructed. However, extended pedigrees are often not available, making haplotype construction difficult and reducing the chances of detecting an error through Mendelian analysis, as well. Even with highly polymorphic microsatellite loci, many errors will be undetectable—particularly errors in parental genotypes in trios and errors in any family member when data are only available from one parent (Douglas et al. 2002). As trends in the study of complex diseases are shifting away from the collection of large family pedigrees and toward the study of smaller family units such as sibling pairs or even toward unrelated individuals (Risch and Merikangas 1996), Mendelian checking is becoming less useful. Consequently, in such analyses, more undetected errors remain in the data. One way of handling such invisible errors is to treat each genotype as uncertain, making use of multipoint sibling pair data to compute the posterior probability of a genotyping error at each marker for each sib pair (Douglas et al. 2000). Another method involves the construction of all possible haplotypes and computation of the posterior probability for each pedigree, depending on user-specified allele frequencies, map distances, and penetrances (Sobel et al. 2002). This method is more powerful than the traditional selection of a single, best-fitting haplotype in the identification of potential double recombinants. Finally, an error model can be directly incorporated into a likelihood-based LOD score calculation (Lincoln and Lander 1992) or linkage disequilibrium (LD) analysis (Gordon et al. 2001).

Even if undetected errors are random, their effects are often nonrandom. In parametric linkage analysis, the presence of undetected genotyping errors inflates the apparent recombination fraction, resulting in increased apparent map distance (Buetow 1991; Shields et al. 1991), incorrect marker ordering, loss of power, and increased false-negative rate. Multipoint linkage analysis is affected by undetected errors to a greater degree than two-point linkage analysis. A significant two-point LOD score that becomes nonsignificant under multipoint linkage is often treated as a two-point false-positive result, when it may in fact be a multipoint false negative (Goring and Terwilliger 2000).

Less has been written about the effects of undetected genotyping error on association and LD analyses than about its effects on linkage analysis. In traditional case-control analysis, a loss of power can result from the presence of undetected genotyping errors (Gordon and

Ott 2001). Apparent background LD rates are also affected by such errors. The direction and magnitude of the effect depends on allele frequencies, true haplotype frequencies, the measure of LD employed, and the genotyping error rate. Error rates in the range of 3% can have substantial effects on two-point measures of background LD between SNPs (Akey et al. 2001). Most relevant to the current study, some recent attention has been focused on inflated transmission/disequilibrium test (TDT) false-positive rates resulting from undetected genotyping errors (Gordon et al. 2001; Heath 1998).

The TDT is a family-based test for linkage in the presence of LD (Spielman et al. 1993). Its appeal lies in its simplicity and its robustness to population stratification (Ewens and Spielman 1995). In its most simple form, the TDT is carried out by counting the number of times a particular allele is transmitted and the number of times it is not transmitted from heterozygous parents to affected children. The null hypothesis is that there is no association between the allele and the disease under study, in which case transmission and non-transmission should be equally likely. Significant deviation from Mendelian transmission is assessed via the McNemar $\chi^2$ test statistic:

$$T = \frac{(t - u)^2}{t + u} \sim \chi_1^2 \, ,$$

where $t$ and $u$ are the transmitted and untransmitted counts, respectively, for the allele under investigation. We have determined that the presence of undetected genotyping errors can lead to substantial inflation of the false-positive rate of this basic TDT and that the bias leads to a particular phenomenon, namely the apparent overtransmission of common alleles.

How could undetected genotyping error lead to inflation of the TDT statistic? If the true genotypes of a father, mother, and child in a trio are $A_1A_1$, $A_1A_2$, and $A_1A_1$, respectively, the father will not be included in the TDT calculation, as he is homozygous for allele $A_1$. However, if his genotype is miscalled as $A_1A_2$, the error is Mendelian consistent and he will be included in the analysis, inflating the transmitted count for allele $A_1$ and the untransmitted count for allele $A_2$. If all families are trios, as is common in studies using the TDT, a homozygous parent miscalled as a heterozygote will never be detected through Mendelian checking. Genotype errors in children are easier to detect than errors in parents (Douglas et al. 2002), but complete detection is by no means possible.

We believe that genotyping error rates are high enough to substantially affect the TDT. Although the overall error rate for a particular data set may be low, the errors may not be distributed evenly across all markers. Because the TDT is performed for each marker in-

dividually, the mean error rate across all markers is not the critical statistic. Rather, the variance in error rates between markers must be considered. For example, a global error rate of 1% across 10 markers could be the result of a 1% error rate at each marker or nine markers could be error-free and one marker could have a 10% error rate. We explore this issue and present evidence that, in practice, the latter scenario is more likely. We assert that in any set of markers typed it is possible that there exists a subset of markers with genotyping error rates high enough to produce false-positive TDT results.

Undetected genotyping errors have been shown, via simulation, to lead to inflated TDT false-positive rates for biallelic markers (Gordon et al. 2001). Here, we present a mathematical model for quantifying the effect of the presence of undetected genotyping errors on the false-positive rate of the TDT. The model can be applied to biallelic loci, as well as to highly polymorphic microsatellites and it allows for the specification of a variety of error models, as described by Sobel et al. (2002). Consistent with the simulations of Gordon et al. (2001), our model predicts several characteristics of the effects of the errors. (1) In the presence of undetected genotyping errors, alleles of unequal frequency appear to be transmitted in a distorted manner—that is, there is an apparent deviation from Mendelian transmission that is not due to linkage to a disease-causing allele. (2) The apparent bias in transmission is in a specific direction: common alleles appear to be overtransmitted to affected children, which seems to increase susceptibility to the disease, and rare or midfrequency alleles appear to be undertransmitted, which seems to be protective. (3) Increasing the number of trios included in the analysis linearly increases the magnitude of the apparent distortion. (4) Only a fraction of genotyping errors are detectable by Mendelian checking, with only 25%–33% detectable in SNPs and 40%–60% detectable in highly polymorphic microsatellites. These characteristics suggest that many reported TDT-derived associations between diseases and marker alleles may be false positives.

## Methods

In this section, we present three of many possible genotyping error models and develop a method for estimating the size and direction of the effects of undetected errors on the TDT. This is done by calculating expected transmitted and untransmitted counts of each allele at a given locus from heterozygous parents to affected offspring in the presence of undetected errors. Expected transmitted and untransmitted counts are obtained through several simple manipulations of four matrices. Two of these four matrices delineate the genotyping error model by specifying the conditional probability of observing a particular genotype, given a particular true

genotype. This is done for both parents and children. It is in these two matrices that our method provides the flexibility to explore any desired error model. The third matrix specifies the relationship between true parental genotypes and true child genotypes—that is, Mendelian inheritance. The fourth matrix gives the expected frequency of each possible combination of parental genotypes under the assumption of Hardy-Weinberg equilibrium. The four matrices and the error models are described in greater detail below.

### Error Models

In our analyses, we maintain distinct allele order within genotypes—that is, genotype $A_1A_2 \neq$ genotype $A_2A_1$. This simplifies computer generation of matrices, particularly in the multiallelic case, as permutations of genotypes do not need to be considered within a single matrix element.

We consider three simple genotyping error models. The first is a random error model (Gordon et al. 2001; Sobel et al. 2002), under which each allele has an equal probability, $e$, of being miscalled and the call is equally likely to be any of the other alleles. The second model is a proportional error model (Douglas et al. 2002; Sobel et al. 2002), under which each allele has an equal probability, $e$, of being miscalled and the call is made in proportion to the frequencies of the other alleles at that marker—that is, allele $i$ is miscalled with probability $e$ and, given that allele $i$ has been miscalled, it will be called as allele $j$ with probability $p_j/(1-p_i)$. The third model is an allelic dropout model, in which the only source of error is miscalling a heterozygote as either homozygote.

### Matrix Descriptions

**M**, the "true-parent/true-child" or Mendelian inheritance matrix, is of size $s^4 \times s^2$, where $s$ is the number of alleles at the marker under observation. The rows in **M** represent all possible parental genotypes ($A_iA_j \times A_kA_l$; $1 \leq i,j,k,l \leq s$) and the columns in **M** represent all possible child genotypes ($A_mA_n$; $1 \leq m, n \leq s$). Element $M_{xy}$ contains the conditional probability that a child produced by parents with true mating type in row $x$ has the true genotype specified in column $y$. For example, row $A_1A_2 \times A_2A_2$ contains elements 0, 1/2, 0, 1/2, corresponding to the probability of parents $A_1A_2$ and $A_2A_2$ having a child with genotype $A_1A_1$, $A_1A_2$, $A_2A_1$ or $A_2A_2$, respectively.

**C**, the "true-child/observed-child" genotype matrix, is of size $s^2 \times s^2$. The rows of the matrix represent all possible true child genotypes and the columns represent all possible observed child genotypes. Element $C_{xy}$ gives the conditional probability of observing a child's genotype as that of column $y$, given that the child's true genotype is that of row $x$. Under the random error

model, elements of $C$ will be of the form $e^N(1 - e)^{2-N}$, where $N$ is the number of alleles that have been miscalled.

$P$, the "true-parent/observed-parent" matrix, is of size $s^4 \times s^4$. The rows in $P$ represent all possible true parental genotypes, and the columns in $P$ represent all possible observed parental genotypes. Each element in the matrix, $P_{xy}$, is the conditional probability of observing the parental genotypes in column $y$, given the true parental genotypes in row $x$. Under the random error model, elements of $P$ will be of the form $e^N(1 - e)^{4-N}$. For example, the element of $P$ that corresponds to row $A_1A_1 \times A_1A_2$ and column $A_1A_1 \times A_2A_2$ would be $e(1 - e)^3$, as three alleles have been called correctly and one has been miscalled. The error parameters in $P$ and $C$ can be tailored to describe any desired error model. This allows the specification of a model that is specific to the genotyping technology employed and the type of marker utilized (Sobel et al. 2002).

Let $H$, the Hardy-Weinberg proportion matrix, be a diagonal $s^4 \times s^4$ matrix. Each element, $H_{xy}$, where $x = y$, is the expected population frequency of the true parental genotypes represented by row $x$, under Hardy-Weinberg equilibrium. For example, with a biallelic marker, the nonzero element in row $A_1A_1 \times A_1A_2$ would be $p^3(1 - p)$, where $p$ is the frequency of the $A_1$ allele. Because genotypes $A_1A_2$ and $A_2A_1$ are considered as separate entities, permutations do not need to be included within a single matrix element. Thus, the frequency of $A_1A_2$ in this example is $p(1 - p)$, rather than $2p(1 - p)$. Note that deviations from Hardy-Weinberg due to inbreeding or other causes can be incorporated into $H$.

### Calculations

Taking the product of $P$, the true-parent/observed-parent genotype matrix, with $H$, the Hardy-Weinberg proportion matrix, produces an $s^4 \times s^4$ matrix, $F$, containing elements giving the frequency of each possible true-parent/observed-parent genotype pair under the assumption of Hardy-Weinberg equilibrium:

$$P \times H = F \qquad (1)$$

Transposing $F$ and taking the product with $M$, the Mendelian inheritance matrix relating true parent and true child genotypes, yields an $s^4 \times s^2$ matrix, $T$, containing the expected frequency of each possible observed-parent/true-child genotype triple:

$$F^T \times M = T \qquad (2)$$

Taking the product of $T$, the observed-parent/true-child genotype matrix, with $C$, the true-child/observed-child genotype matrix, yields $D$, an $s^4 \times s^2$ matrix containing the expected frequencies of all possible observed-parent/

observed-child triples. In other words, $D$ contains the data that would be expected experimentally under the error model specified in $P$ and $C$, with population allele frequencies specified in $H$:

$$T \times C = D \qquad (3)$$

Matrix $D$ has the same layout as $M$, except that the rows in $D$ represent observed parental genotype pairs and columns represent observed child genotypes, rather than true parent and true child genotypes, as found in $M$. Since $D$ contains the expected frequencies of all possible observed parent-child trio genotypes, it can be used to compute a predicted bias in the TDT statistic under any genotyping error model and to estimate the fraction of all genotyping errors that would be detectable by checking for Mendelian inconsistencies in trios. The cells containing zeros in $M$ contain the frequencies of Mendelian detectable errors in $D$, as they represent parent-child genotype trios that are inconsistent with the rules of Mendelian inheritance. The sum of these elements in $D$ is the expected frequency of Mendelian-detectable genotyping errors, $E$.

To determine the expected fraction of errors that would be detectable through Mendelian checking, the total expected frequency of trios with genotyping errors can be calculated and compared to $E$, the expected frequency of Mendelian-detectable errors. Under the random error model, the probability that a trio contains no genotyping errors is $(1-e)^6$. The probability that a trio contains at least one error is then $1-(1-e)^6$. Thus, the expected fraction of errors that could be detected through Mendelian checking is:

$$\frac{E}{1 - (1 - e)^6} \cdot \qquad (4)$$

### The TDT Statistic

Since each element in $D$ gives the probability of observing a particular parent-child genotype combination, $D$ can be multiplied by $n$, the total number of trios in the data set, to give expected counts for each type of trio. It is then possible to determine the number of trios that would contribute to the TDT statistic and their relative contributions. All Mendelian-inconsistent genotypes typically are removed from a data set prior to any analysis. Under this protocol, the trios represented in the calculation of $E$ (trios with Mendelian errors) would be removed and only the remaining Mendelian-consistent trios would be considered in the calculation of the TDT statistic. Depending upon the genotype of the child, trios with one parent heterozygous for allele $A_1$ would each contribute one count to the transmitted

or untransmitted tally of $A_1$, while trios with two parents heterozygous for $A_1$ would contribute either two transmitted counts, two untransmitted counts or one of each. Using these expected transmitted and untransmitted counts for $A_1$, it is possible to compute the expected bias in the TDT statistic under the error model employed and subsequently to calculate the false-positive rate under the specified model.

Let $h$ be the number of observed heterozygous $A_1A_2$ parents included in the computation of the TDT statistic and let $t$ and $u$ be the observed transmitted and untransmitted counts, respectively, of the $A_1$ allele from these heterozygous parents to affected children, such that $t + u = h$. Under the null hypothesis ($H_0$), there is assumed to be no linkage, no LD, no genotyping error, and, consequently, no bias in transmission. Thus, under $H_0$, the probability of transmission of the $A_1$ allele, $p_0$, is equal to 0.5. If $t \sim \text{Binomial}(h,p_0)$, its expected value is $hp_0 = h/2$, with variance $hp_0(1 - p_0) = h/4$. Then,

$$T = \frac{(t - hp_0)^2}{hp_0(1 - p_0)} = \frac{(t - u)^2}{t + u} \sim \chi_1^2 \ ,$$

the familiar McNemar $\chi^2$ test statistic employed in the TDT.

Under the alternative hypothesis ($H_A$), the assumptions of no linkage and no LD still hold, but genotyping error is allowed. The probability of transmission of the $A_1$ allele under $H_A$ is

$$p_A = p_0 + \varepsilon \ ,$$

where $\varepsilon$ is the deviation of $p_0$ from 0.5, or the bias in transmission of allele $A_1$. This leads to $t \sim \text{Binomial}(h,p_A)$, with expected value $E\{t\} = hp_A = h(1/2 + \varepsilon)$ and variance $\text{Var}\{t\} = hp_A(1 - p_A) = h(1/2 + \varepsilon)(1 - 1/2 - \varepsilon)$.

To obtain the distribution of the TDT statistic under the alternative hypothesis, consider the transformation

$$Z = \frac{t - h/2}{\sqrt{h/4}} \ .$$

Under $H_A$,

$$E\{Z\} = \frac{E\{t\} - h/2}{\sqrt{h/4}}$$

$$= \frac{h/2 + h\varepsilon - h/2}{\sqrt{h/4}} = 2\varepsilon\sqrt{h}$$

and

$$\text{Var}\{Z\} = \frac{\text{Var}\{t\}}{[\sqrt{h/4}]^2}$$

$$= \frac{h(1/4 - \varepsilon^2)}{h/4} \to 1, \text{ as } \varepsilon \to 0 \ .$$

So, $Z \sim \text{N}(2\varepsilon\sqrt{h},1)$, when $\varepsilon$ is small. Therefore, $Z^2 \sim$ Noncentral $\chi_1^2$ with noncentrality parameter $\lambda = [E\{Z\}]^2 = 4\varepsilon^2 h$. The McNemar $\chi^2$ test statistic, derived above, is equivalent to $Z^2$, as shown below:

$$T = \frac{(t - u)^2}{t + u} = \frac{(2t - h)^2}{h} = \left[\frac{2t - h}{\sqrt{h}}\right]^2$$

$$= \left[\frac{t - h/2}{\sqrt{h/4}}\right]^2 = Z^2 \ .$$

Therefore, under $H_A$, the TDT statistic follows a noncentral $\chi_1^2$ distribution with noncentrality parameter $\lambda = 4\varepsilon^2 h$.

Note that if the observed $\varepsilon$ is assumed to equal its expected value, $\frac{t}{t+u} - \frac{1}{2}$, $\lambda$ can be expressed in a more intuitive manner. Substituting $t + u$ for $h$ and $\frac{t}{t+u} - \frac{1}{2}$ for $\varepsilon$ in $\lambda = 4\varepsilon^2 h$ yields

$$\lambda = 4\left[\frac{t}{t + u} - \frac{1}{2}\right]^2 (t + u)$$

$$= \left[\frac{2t - (t + u)}{t + u}\right]^2 (t + u) = \frac{(t - u)^2}{t + u} \ ,$$

which is the TDT test statistic using observed values of $t$ and $u$. When $t = u$, $\lambda = 0$ and when $t \neq u$, $\lambda$ represents the magnitude of the bias introduced into the TDT statistic by the presence of undetected genotyping errors.

We used **D** to determine expected transmitted and untransmitted counts for varying allele frequencies, number of alleles, sample sizes and error rates. Using these expected counts as $t$ and $u$, we calculated $\lambda$, the noncentrality parameter, and then used a noncentral $\chi_1^2$ distribution to compute the false-positive rates of the TDT statistic under the same parameters. This is analogous to the most common application of the noncentral $\chi^2$ distribution, as we have used it to compute the power to detect a shift from $\lambda = 0$ (under $H_0$) to a nonzero $\lambda$ (under $H_A$). For multiallelic markers, each allele was taken individually against all other alleles grouped together.

### Literature Searches

Using PubMed searches for "TDT + association," "TDT + susceptibility," "transmission/disequilibrium,"

**Table 1**

**Expected Transmitted and Untransmitted Counts for Major Allele at a Biallelic Marker and Expected TDT False-Positive Rate for $P < .05$**

| $e$, $n$, and $p$ | $t$ | $u$ | $\lambda$ | $\alpha$ |
|---|---|---|---|---|
| .005: | | | | |
| 100: | | | | |
| .60 | 48.0 | 47.6 | .00 | .05 |
| .75 | 38.1 | 37.1 | .01 | .05 |
| .85 | 26.5 | 25.1 | .04 | .05 |
| .90 | 19.3 | 17.8 | .07 | .06 |
| .95 | 11.1 | 9.4 | .15 | .07 |
| .99 | 3.9 | 2.0 | .63 | .12 |
| 200: | | | | |
| .60 | 96.0 | 95.2 | .00 | .05 |
| .75 | 76.1 | 74.1 | .03 | .05 |
| .85 | 53.0 | 50.2 | .07 | .06 |
| .90 | 38.7 | 35.5 | .13 | .07 |
| .95 | 22.3 | 18.7 | .30 | .09 |
| .99 | 7.8 | 3.9 | 1.26 | .20 |
| 500: | | | | |
| .60 | 240.0 | 238.0 | .01 | .05 |
| .75 | 190.3 | 185.4 | .06 | .06 |
| .85 | 132.5 | 125.5 | .15 | .07 |
| .90 | 96.6 | 88.8 | .33 | .09 |
| .95 | 55.7 | 46.9 | .76 | .14 |
| .99 | 19.4 | 9.8 | 3.16 | .43 |
| .01: | | | | |
| 100: | | | | |
| .60 | 48.0 | 47.2 | .01 | .05 |
| .75 | 38.6 | 36.7 | .05 | .06 |
| .85 | 27.6 | 24.9 | .14 | .07 |
| .90 | 20.6 | 17.5 | .24 | .08 |
| .95 | 12.7 | 9.3 | .54 | .11 |
| .99 | 5.7 | 2.0 | 1.84 | .27 |
| 200: | | | | |
| .60 | 96.0 | 94.4 | .01 | .05 |
| .75 | 77.2 | 73.3 | .10 | .06 |
| .85 | 55.2 | 49.8 | .27 | .08 |
| .90 | 41.2 | 35.1 | .50 | .11 |
| .95 | 25.4 | 18.5 | 1.09 | .18 |
| .99 | 11.5 | 3.9 | 3.69 | .48 |
| 500: | | | | |
| .60 | 239.9 | 236.1 | .03 | .05 |
| .75 | 192.9 | 183.3 | .25 | .08 |
| .85 | 138.0 | 124.5 | 1.10 | .18 |
| .90 | 103.1 | 87.7 | 1.24 | .20 |
| .95 | 63.6 | 46.3 | 2.72 | .38 |
| .99 | 28.6 | 9.8 | 9.22 | .86 |
| .02: | | | | |
| 100: | | | | |
| .60 | 48.0 | 46.5 | .02 | .05 |
| .75 | 39.6 | 35.9 | .18 | .07 |
| .85 | 29.5 | 24.3 | .50 | .11 |
| .90 | 23.1 | 17.2 | .87 | .15 |
| .95 | 15.7 | 9.1 | 1.77 | .27 |
| .99 | 9.2 | 2.0 | 4.64 | .58 |

*(continued)*

**Table 1 (continued)**

| $e$, $n$, and $p$ | $t$ | $u$ | $\lambda$ | $\alpha$ |
|---|---|---|---|---|
| 200: | | | | |
| .60 | 95.9 | 93.0 | .05 | .06 |
| .75 | 79.1 | 71.76 | .36 | .09 |
| .85 | 59.0 | 28.6 | .99 | .17 |
| .90 | 46.1 | 34.3 | 1.73 | .26 |
| .95 | 31.5 | 18.2 | 3.55 | .47 |
| .99 | 18.5 | 4.0 | 9.28 | .86 |
| 500: | | | | |
| .60 | 239.8 | 232.4 | .12 | .06 |
| .75 | 197.9 | 179.4 | .90 | .16 |
| .85 | 147.5 | 121.5 | 2.48 | .35 |
| .90 | 115.3 | 85.8 | 4.33 | .55 |
| .95 | 78.7 | 45.5 | 8.87 | .85 |
| .99 | 46.2 | 10.1 | 23.20 | .99 |
| .05: | | | | |
| 100: | | | | |
| .60 | 47.7 | 44.4 | .12 | .06 |
| .75 | 42.1 | 33.9 | .87 | .15 |
| .85 | 34.4 | 23.0 | 2.26 | .32 |
| .90 | 29.5 | 16.4 | 3.70 | .49 |
| .95 | 23.7 | 9.0 | 6.56 | .73 |
| .99 | 18.6 | 2.6 | 12.05 | .93 |
| 200: | | | | |
| .60 | 95.4 | 88.9 | .23 | .08 |
| .75 | 84.1 | 67.8 | 1.75 | .26 |
| .85 | 68.8 | 46.0 | 4.52 | .57 |
| .90 | 58.9 | 32.8 | 7.40 | .78 |
| .95 | 47.4 | 18.1 | 13.12 | .95 |
| .99 | 37.1 | 5.2 | 24.09 | .99 |
| 500: | | | | |
| .60 | 238.6 | 222.3 | .58 | .12 |
| .75 | 210.3 | 169.5 | 4.37 | .55 |
| .85 | 172.0 | 115.0 | 11.30 | .92 |
| .90 | 147.3 | 82.1 | 18.50 | .99 |
| .95 | 118.5 | 45.2 | 32.80 | .99 |
| .99 | 92.8 | 13.0 | 60.23 | .99 |

and "transmission disequilibrium," we compiled a list of studies that used the basic TDT to demonstrate significant $(P < .05)$ association between a marker allele and a disease. Studies that considered haplotypes rather than single markers and studies that applied a global TDT (i.e., those that considered all alleles simultaneously, rather than each allele individually against all others as a group) were excluded. The collection was then narrowed to include only studies available online or in the Johns Hopkins University School of Medicine library. Finally, only those that provided allele frequencies for the marker in question or those for which marker allele frequencies could be located for an ethnically matched control population were analyzed. Given allele frequencies in both the study and a control population, the control population frequencies were chosen. Case-

**Table 2**

**Expected TDT False-Positive Rates for *P* < .05 at an Eight-Allele Locus, Testing One Allele Against All Others as a Group**

| | VALUE FOR | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Highest Frequency Allele | | | Midfrequency Allele | | | Lowest Frequency Alleles | | |
| *e* AND *n* | $p_1^a$ | $\alpha_r^b$ | $\alpha_p^c$ | $p_2^a$ | $\alpha_r^b$ | $\alpha_p^c$ | $p_3^a$ | $\alpha_r^b$ | $\alpha_p^c$ |
| .01: | | | | | | | | | |
| 200 | .30 | .05 | .05 | .25 | .05 | .05 | .075 | .05 | .05 |
| | .40 | .05 | .05 | .30 | .05 | .05 | .050 | .05 | .05 |
| | .60 | .06 | .05 | .30 | .05 | .05 | .017 | .06 | .05 |
| | .80 | .08 | .07 | .14 | .05 | .06 | .010 | .06 | .05 |
| 500 | .30 | .05 | .05 | .25 | .05 | .05 | .075 | .05 | .05 |
| | .40 | .05 | .05 | .30 | .05 | .05 | .050 | .05 | .05 |
| | .60 | .07 | .06 | .30 | .05 | .05 | .017 | .07 | .05 |
| | .80 | .12 | .11 | .14 | .05 | .08 | .010 | .08 | .05 |
| .02: | | | | | | | | | |
| 200 | .30 | .05 | .05 | .25 | .05 | .05 | .075 | .05 | .05 |
| | .40 | .06 | .05 | .30 | .05 | .05 | .050 | .05 | .05 |
| | .60 | .08 | .06 | .30 | .05 | .06 | .017 | .07 | .05 |
| | .80 | .15 | .14 | .14 | .05 | .10 | .010 | .09 | .05 |
| 500 | .30 | .06 | .05 | .25 | .05 | .05 | .075 | .05 | .05 |
| | .40 | .07 | .06 | .30 | .06 | .05 | .050 | .06 | .05 |
| | .60 | .12 | .08 | .30 | .06 | .06 | .017 | .09 | .05 |
| | .80 | .31 | .27 | .14 | .05 | .18 | .010 | .14 | .06 |
| .04: | | | | | | | | | |
| 200 | .30 | .06 | .05 | .25 | .06 | .05 | .075 | .05 | .05 |
| | .40 | .08 | .06 | .30 | .06 | .05 | .050 | .06 | .05 |
| | .60 | .15 | .10 | .30 | .06 | .07 | .017 | .11 | .05 |
| | .80 | .41 | .35 | .14 | .05 | .23 | .010 | .17 | .06 |
| 500 | .30 | .08 | .06 | .25 | .06 | .05 | .075 | .06 | .05 |
| | .40 | .12 | .07 | .30 | .08 | .05 | .050 | .07 | .05 |
| | .60 | .31 | .17 | .30 | .07 | .10 | .017 | .20 | .06 |
| | .80 | .78 | .71 | .14 | .05 | .48 | .010 | .34 | .09 |
| .06: | | | | | | | | | |
| 200 | .30 | .07 | .06 | .25 | .06 | .05 | .075 | .06 | .05 |
| | .40 | .11 | .07 | .30 | .07 | .05 | .050 | .07 | .05 |
| | .60 | .26 | .15 | .30 | .07 | .09 | .017 | .16 | .06 |
| | .80 | .68 | .59 | .14 | .05 | .39 | .010 | .26 | .08 |
| 500 | .30 | .11 | .07 | .25 | .08 | .05 | .075 | .06 | .05 |
| | .40 | .20 | .10 | .30 | .11 | .05 | .050 | .10 | .06 |
| | .60 | .56 | .30 | .30 | .09 | .14 | .017 | .33 | .07 |
| | .80 | .97 | .93 | .14 | .05 | .75 | .010 | .55 | .13 |

[a] Frequencies are given for the most common allele, a midfrequency allele, and the six remaining, equally frequent alleles.
[b] $\alpha_r$ = expected false-positive rate under the random error model.
[c] $\alpha_p$ = expected false-positive rate under the proportional error model.

control studies were compiled by searching PubMed for "association + case + control + genetic" and "microsatellite + case + control." Only analyses of microsatellite or SNP marker alleles were compiled. Articles not available online were excluded. A complete list of all TDT and case-control studies can be found in appendix A (online only).

Estimated genotyping error rates for collected TDT studies were obtained by using the sample size and sig-

nificance level to compute the lowest value of λ that would be required to render the reported TDT statistic nonsignificant. Our random error model was then used to determine the magnitude of genotyping error that would generate the calculated λ value for the number of trios included in the analysis.

## Results

### Impact of Allele Frequency on Transmission Distortion

Table 1 shows λ, the magnitude of the expected bias in the TDT statistic, and α, the expected false-positive rate under the random error model for varying genotyping error rates, numbers of trios and allele frequencies for a biallelic marker. Results are identical for the proportional error model. Table 2 gives α for an eight-allele marker, with one common allele, one midfrequency allele, and six equally frequent rare alleles under the random error model and the proportional error model. For both types of markers and both error models, we found an increasing false-positive rate with increasing genotyping error rate, sample size, and difference in frequency between alleles at the marker. In all cases, the most common allele appears to be overtransmitted. Under the random error model, the rarest alleles at a multiallelic marker appear to be undertransmitted, whereas alleles of midfrequency are unaffected. Under the proportional error model, midfrequency alleles sometimes appear to be undertransmitted, and rare alleles are almost always unaffected.

Under the allelic dropout model, common alleles appear to be overtransmitted and rare alleles appear to be undertransmitted, but error rates must be approximate

**Table 3**

**Expected Fraction of Mendelian-Detectable Errors at an Eight-Allele Marker with One High-Frequency Allele, One Midfrequency Allele, and Six Equally Frequent Rare Alleles under the Random Error Model**

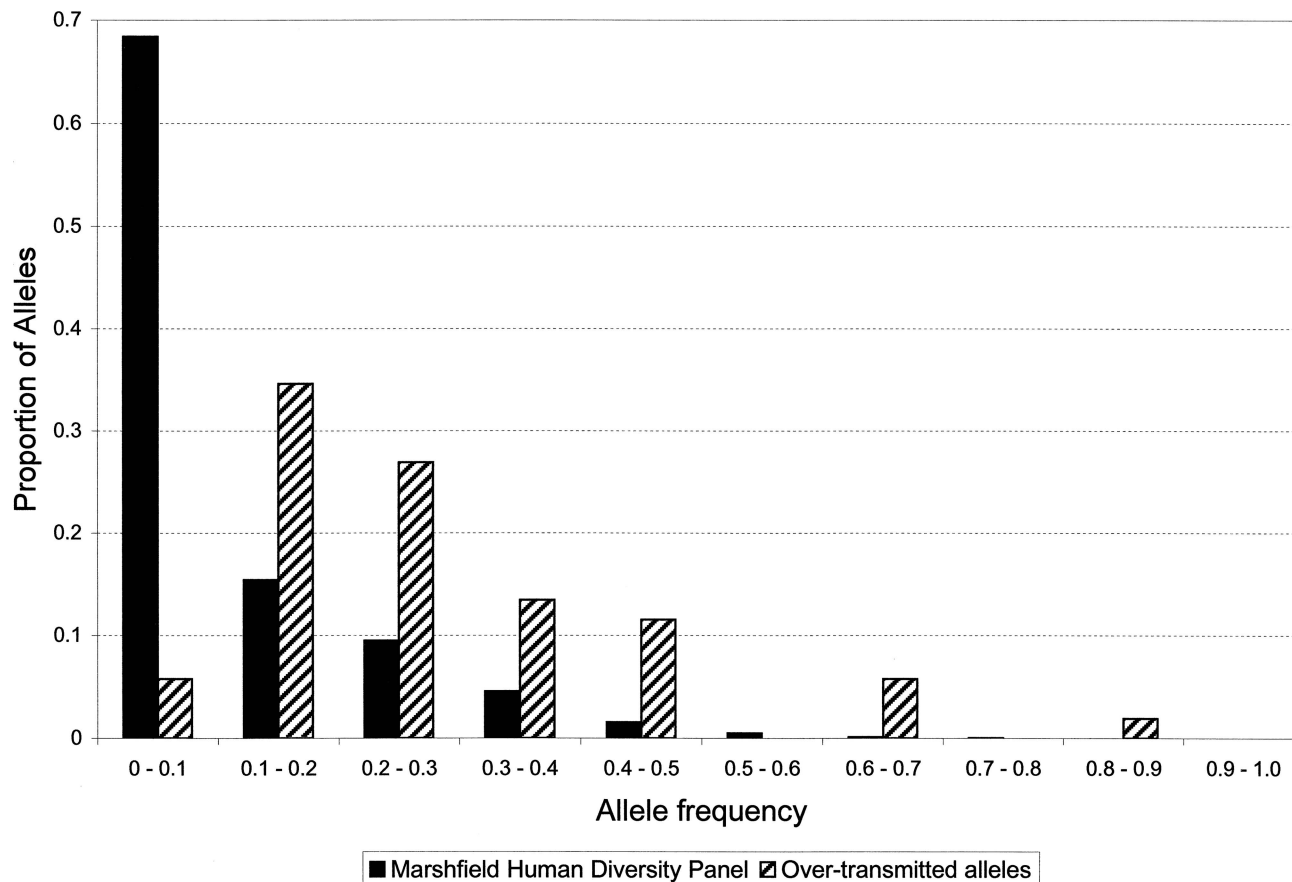| *e* | FRACTION OF TRIOS WITH ≥1 ERROR | FREQUENCY OF ALLELE(S) | | | FRACTION OF ERRORS DETECTABLE |
|---|---|---|---|---|---|
| | | 1 | 2 | 3–8 | |
| .02 | .11 | .30 | .25 | .075 | .56 |
| | | .40 | .30 | .050 | .53 |
| | | .60 | .30 | .017 | .47 |
| | | .80 | .14 | .010 | .43 |
| .04 | .22 | .30 | .25 | .075 | .57 |
| | | .40 | .30 | .050 | .54 |
| | | .60 | .30 | .017 | .49 |
| | | .80 | .14 | .010 | .44 |
| .06 | .31 | .30 | .25 | .075 | .58 |
| | | .40 | .30 | .050 | .56 |
| | | .60 | .30 | .017 | .50 |
| | | .80 | .14 | .010 | .46 |

**Figure 1** Distribution of frequencies of overtransmitted microsatellite alleles among controls in TDT studies from the literature (*striped bars*) and of all microsatellite alleles in the Marshfield Clinic Human Diversity Panel (*solid bars*). The difference between the distributions is highly significant ($P < 1.0 \times 10^{-11}$, Kolmogorov-Smirnov Test).

ly twice as high as with the random error model to see the same magnitude of apparent distortion (data not shown).

Table 3 gives the total number of expected errors and the expected rate of error detection in trios for varying allele frequencies and error rates under the random error model for multiallelic markers. The detectable error rate ranged from 25% to 33% for SNPs (data not shown), and from 40% to 60% for multiallelic markers, with increasing rates of detection with increasing major allele frequency and decreasing error rates.

### TDT in the Literature

PubMed searches yielded 189 studies that used the TDT to demonstrate significant LD between a single marker allele and a disease. Significance was assessed at the 0.05 level, without adjustment for multiple testing. Of these 189 studies, 23 used a likelihood-based or global TDT, 24 did not provide allele frequencies, 6 analyzed a quantitative trait, 32 were not available online

or in the Johns Hopkins University medical library, and 5 were not in English. In the remaining 99 studies, 79 microsatellite, 18 variable number of tandem repeats (VNTR), 8 human leukocyte antigen (HLA), and 29 SNP alleles and 4 microdeletions were found to exhibit significantly biased transmission to affected offspring. Table A (online only) provides allele frequencies, transmitted and untransmitted counts, and TDT scores for the SNPs and multiallelic markers in the 99 studies.

Figure 1 shows a histogram of frequencies of overtransmitted microsatellite alleles in controls from TDT studies in the literature, as compared to frequencies of all microsatellite alleles in the Marshfield Clinic Human Diversity Panel. The difference between the distributions is highly significant ($P < 1.0 \times 10^{-11}$, Kolmogorov-Smirnov Test).

### TDT versus Case-Control

Of the 79 identified transmission distortions of microsatellite alleles, 31 involved the most common allele

at the locus. Of these 31 transmission distortions of the most common allele, 27 (87%) found it to be overtransmitted to affected offspring, implying that the most common allele may be in LD with a disease-causing mutation. Of the 29 identified associations between SNP alleles and disease, 17 (59%) were between the most common allele and susceptibility to the disease. To assess the significance of the tendency of the TDT to identify the most common allele as associated with susceptibility to disease, rather than protection from disease, we compared these results to those of case-control analyses. Of 92 significant case-control associations involving microsatellite alleles, 37 implicated the most common allele as conferring either protection or risk. Of these, 12 (32%) found the most common allele to be enriched in cases relative to controls (a risk allele). Of 182 case-control studies involving SNPs, 57 (31%) found the common allele to be enriched in cases relative to controls. The difference between the TDT and case-control results, for both microsatellites and SNPs, is highly significant ($P = 3.6 \times 10^{-6}$, Fisher's exact test). These results are summarized in table 4 and figure 2.

### Error Rate Estimation

In addition to the 31 identified distortions of the most common allele at a microsatellite marker, we found 9 distortions of the most common allele at an HLA marker or VNTR, for a total of 40 associations between disease status and the most common allele at a multiallelic marker. For 29 of these 40 associations, frequencies were available or could be approximated for the remaining alleles at the marker. Using our random error model, we estimated the genotyping error rates that would be required to render these 29 results insignificant. We found that genotyping error, as we have modeled it, could not explain the significance of 10 of the 29 results. All 10 of these studies contained <200 trios. Of the results that could be explained by a random genotyping error model, four would require an allelewise error rate of 10% or higher, eight would require 4%–8%, and the remaining seven significant results could be explained by genotyping error rates of 3% or less. On average, an allelewise genotyping error rate of 5.5% would create a noncentrality parameter large enough to bring the true $P$ values above .05. Error rate estimation was also performed on the 17 associations between the most common SNP allele and disease susceptibility. Genotyping error, as we have modeled it here, could not account for four of the associations, and the remainder could be explained by an average allelewise error rate of 3.2%. These results are presented in table B (online only).

To understand the magnitude of error rates in practice, we compiled a list of microsatellite genotyping error rate

**Table 4**

**TDT versus Case-Control Studies in the Literature**

| ALLELE TYPE AND CATEGORY | No. (%) OF ALLELES IN STUDY TYPE | |
| --- | --- | --- |
| | TDT | Case-Control |
| Microsatellite:[a] | | |
| Protective | 4 (12.9) | 25 (67.6) |
| Risk | 27 (87.1) | 12 (32.4) |
| SNP:[b] | | |
| Protective | 12 (41.4) | 125 (68.7) |
| Risk | 17 (58.6) | 57 (31.3) |
| Combined:[c] | | |
| Protective | 16 (32.0) | 150 (68.5) |
| Risk | 34 (68.0) | 69 (31.5) |

NOTE.—Given significant association ($P < .05$) between the most common allele and disease status, results are broken down into protective (TDT: common allele undertransmitted; case-control: common allele enriched in controls) or risk (TDT: common allele overtransmitted; case control: common allele enriched in cases) categories. $P$ values were computed using Fisher's exact test.

[a] $P = 6.0 \times 10^{-6}$.
[b] $P = 6.1 \times 10^{-3}$.
[c] $P = 3.6 \times 10^{-6}$.

estimates from the literature. The results are summarized in table 5.

### Discussion

The apparent overtransmission of common alleles observed in the literature could have several explanations, each accounting for a fraction of the total trend. First, the associations could be real and the distorted allele could be the disease-causing mutation, lending support to the common disease–common allele hypothesis. From a population genetics standpoint, however, this seems unlikely, as there will be selection against any allele that reduces reproductive fitness by even a small amount. To have a disease-causing allele reach a frequency of a few percent in the population would require that the allele also confer some selective advantage. For a disease-causing allele to become the most common allele in a population would be highly unusual. For this to happen as a rule, rather than as the exception, is unreasonable. If the diseases in the collected studies were uniformly diseases of late-onset, it might be easier to imagine that historically there was less selective pressure against them. However, the collected studies cover a wide range of diseases, with 55% of them occurring primarily in childhood, making this scenario unlikely to explain many of the compiled associations. Alternatively, the distorted marker allele may be in LD with the disease-predisposing allele. In this case, the probability that a mutation occurs
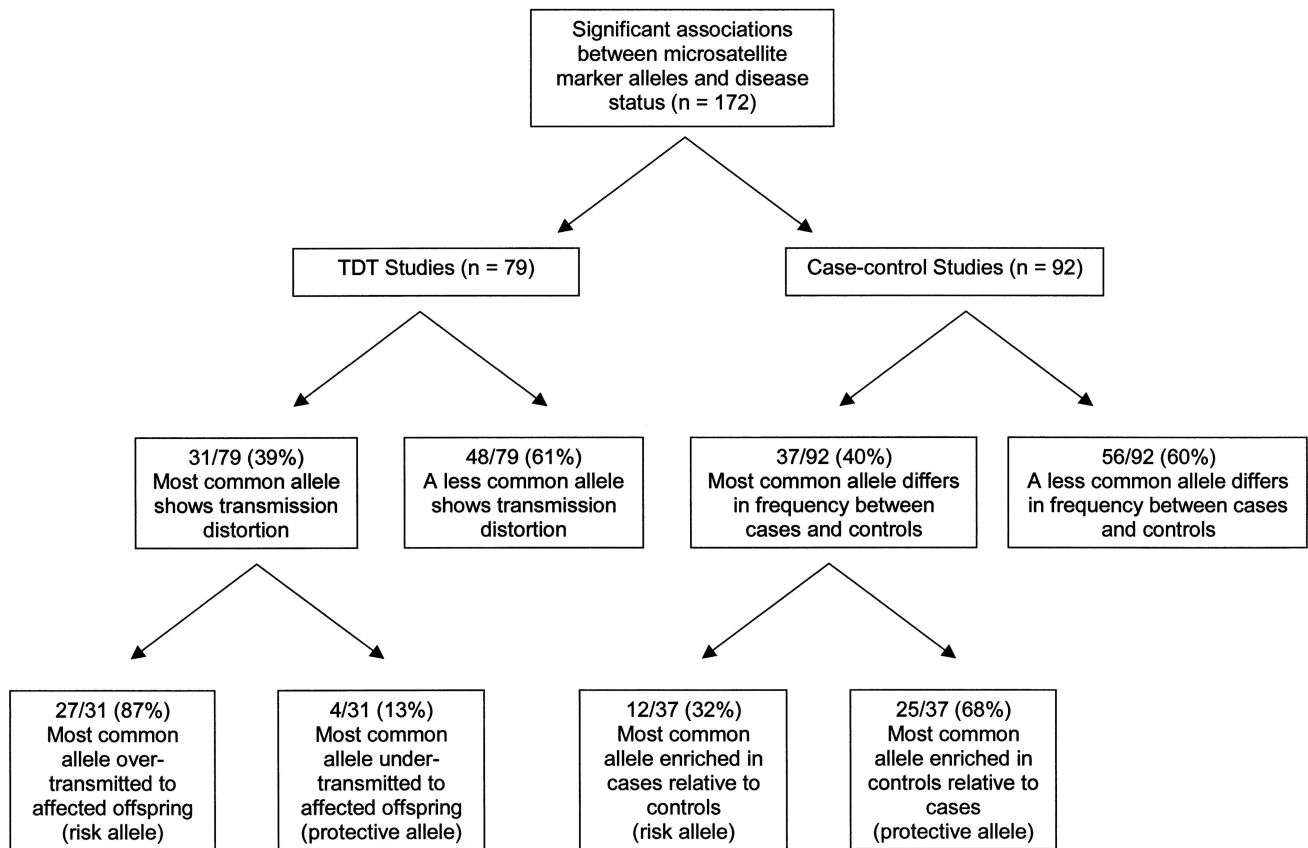
**Figure 2**    Breakdown of compiled microsatellite association studies by study type (TDT or case-control), allele implicated (most common or not), and nature of allele (risk or protective).

on any given allele may be proportional to the frequency of that allele, resulting in association between the frequency of an allele and over- or undertransmitted status. Finally, genotyping error rates may be substantial enough to be significantly contributing to the false-positive rate of the TDT and, consequently, strengthening the association between allele frequency and over- or undertransmission. Table 4 and figures 1 and 2 can help distinguish between the latter two possibilities.

As figure 1 shows, the frequency distribution of overtransmitted microsatellite alleles among controls in TDT studies is significantly different than the frequency distribution of all microsatellite alleles in the Marshfield Clinic Human Diversity Panel ($P < 1.0 \times 10^{-11}$). If the tendency of the most common microsatellite allele to appear associated with increased risk of disease were due entirely to LD between the marker allele and the disease-predisposing allele, we might expect the two frequency distributions in figure 1 to be more similar. The marked difference between the two distributions implies that the TDT finds common alleles to be disproportionately overtransmitted.

We also found significant differences between the

TDT and case-control studies, as depicted in table 4 and figure 2. Presumably, there is no difference in biology between associations that could be detected using the TDT and those that could be detected using case-control methods. The processes of mutation, recombination, natural selection, and genetic drift that influence patterns of LD should affect the TDT and case-control methods equally. Thus, the expectation is that the ratio of over- to undertransmission of the most common allele at a marker would be similar for both tests. However, as table 4 shows, this is not what is observed. Given that the most common allele at a microsatellite marker has been associated with disease status, the TDT finds it to be a risk factor 87% of the time, whereas case-control finds it to be a risk factor 32% of the time. The trend for SNPs is similar, with the most common allele being implicated as a risk factor 59% of the time with the TDT and 31% of the time with case-control. The difference between the TDT and case-control results is highly significant ($P = 3.6 \times 10^{-6}$, Fisher's exact test). It is not clear what effect differences in power to detect rare versus common alleles may have in a comparison between TDT and case-control results; further explo-

**Table 5**

**Estimates of Microsatellite Genotyping Error Rates**

| Reference and Error Rate[a] | Notes |
|---|---|
| Brzustowicz et al. 1993: | |
| 3.2% | Global average error rate in CEPH database |
| 5.9% | Global average error rate in CEPH among "non-obvious equivalents" (i.e., blind duplicates) |
| .7%–6.8% | Range in per-marker error rates in CEPH database at four markers examined on chromosome 5 |
| Smith et al. 1995: | |
| 1%–3% | Includes only errors due to nontemplated nucleotide addition by *Taq* DNA polymerase |
| Hall et al. 1996: | |
| .55% | 35 discrepancies in duplicate genotyping of 12 markers in 268 individuals |
| Weeks et al. 2002: | |
| .65%–6.8% | Comparison between MGS and CIDR genotypes; range is due to differing treatments of non one-to-one mappings of allele sizes between centers |
| 0%–63% | MGS-CIDR discordance rate per marker that could not be resolved by linear transformation of allele sizes from one center to the other. Actual values were: 0%, 205 markers; 3.3%, 28 markers; 3.3%–17%, 42 markers; and 50%–63%, 3 markers |
| Ghosh et al. 1997: | |
| 2.6% | Discordance rate in allele calling between two technicians reading the same gel |
| Idury et al. 1997: | |
| 1.9% | Mendelian-detectable error rate in nuclear families |
| Mansfield et al. 1994: | |
| 1.5% | Mendelian-detectable error rate in nuclear families |
| Sobel et al. 2002: | |
| 4%–13% | Average error rates per chromosome for alleles called with automated Genotyper software. Note that very high error rate on one chromosome was due to a single marker with initial error rate of 29%; after removal of this problem, marker error rate was 4%–5%. |
| 0%–8% | Average error rates per chromosome for alleles called manually. After removal of problem marker: 0%–2%. |
| 0%–7% | Average error rates per chromosome for data cleaned by quality-control procedures described in reference. After removal of problem marker: 0%. |

[a] Error rates are per genotype unless otherwise specified. Note that the error rates used in the text and in tables 1–3 are per allele.

ration of this issue is warranted. However, as figure 2 shows, there appears to be no difference in ascertainment rate of the most common allele, as 39% and 40% of significant microsatellite TDT and case-control results, respectively, involve the most common allele. We believe that the tendency of the TDT to identify the most common allele as a risk factor is a direct effect of the presence of undetected genotyping errors.

We used our random error model to estimate the magnitude of genotyping error that could account for the apparent overtransmission of common microsatellite alleles found in the literature. We found that approximately one-third of the results could not be due to random undetected genotyping error as we have modeled it (table B). Another third would require error rates in the range of 5%–10% per allele, or 10%–20% per genotype. The final third would require 0.3%–3% error per allele, or 0.6%–6% per genotype.

To assess whether these error rates are realistic, we returned to the literature, compiling all available analyses of microsatellite genotyping error rates. As table 5 shows, estimates in global mean error rates range from 0.55% (Hall et al. 1996) to 5.9% (Brzustowicz et al.

1993) per genotype. However, because the TDT is computed on a per-marker basis, one must consider both the mean error rate across all markers and the variance in error rates between markers when estimating true error rates. Several of the studies listed in table 5 address this issue. Brzustowicz et al. (1993) examined four markers on chromosome 5 and found per-marker error rates of 0.7%, 2.7%, 3.4%, and 6.8%. If only the mean error rate across these markers (3.0%) were reported, the high error rate at the last marker would not be evident. Weeks et al. (2002) compared genotype data generated by two high-throughput genotyping centers, the Center for Inherited Disease Research (CIDR) (National Human Genome Research Institute, National Institutes of Health, Bethesda, MD) and the Mammalian Genotyping Service (MGS) (Marshfield Clinic, Marshfield, WI). A total of 321 markers were typed by both centers in 30 samples. Of the 321 markers, 276 showed approximately one-to-one correspondence in allele size between centers (0% discordance). Among the remaining 45 markers, there were 116 discordant pairs of alleles. Interestingly, the discordancies were not distributed evenly across the 45 markers; 50 of the 116

discordancies occurred at 3 markers, resulting in per-marker discordance rates >50% at these markers. Fourteen markers showed discordance rates of 3.3%–17%, and 28 markers showed discordance rates of 3.3% for a global mean discordance rate of 1.3%. Again, the global mean does not reflect the high variance between markers. Finally, in an analysis of 50 markers on three chromosomes, Sobel et al. (2002) observed a large variance in per-marker error rates. A global mean error rate of 5.6% included one marker with a 29% error rate.

It is evident that genotyping error rates are not uniform across markers; some markers show error rates that are many times higher than those shown by others. The erratic behavior of a subset of markers could be a problem particularly when genotyping is performed by an individual laboratory, rather than a high-throughput center, as the markers included in high-throughput panels have been studied extensively. They have specifically been selected for inclusion because they produce highly replicable genotypes, and their individual PCR reactions have been optimized (CIDR, MGS). In addition, large-scale genotyping centers are highly practiced, calling millions of genotypes per year. These centers report global mean error rates on the order of 0.5% per genotype for blind duplicates run on the same gel. Since few of the studies included in our analyses outsourced their genotyping to such centers, it is not difficult to imagine that they would have mean error rates that are several times higher than those of the large centers, particularly when gel-to-gel and technician-to-technician variation are included in the error rate calculation. More importantly, there is no reason to expect that the studies included in our analyses would have variances in error rates between markers that would be any smaller than those listed in table 5. Therefore, we believe that in practice genotyping error rates are sufficiently high to bias the TDT, inflating the false-positive rate and causing common alleles to appear to be overtransmitted.

We recommend several strategies for coping with the TDT's inflated false-positive rate associated with undetected genotyping error. First, an association between a common allele and disease susceptibility or a rare allele and protection, identified thorough the TDT, should be treated with caution. More confidence can be placed in an association in the opposite direction than the predicted bias. For an association in the predicted direction, confidence can be increased by repeat genotyping, analysis of nearby markers, use of a haplotype rather than an individual marker, or use of a functional assay in the case of a SNP. Second, we recommend re-genotyping all individuals at any marker showing significant transmission distortion. For SNPs, the regenotyping should be carried out using a different technology than was originally employed. If the same genotypes are obtained through both methods, one can be more confident in the association. Third, a likelihood-based TDT method that is robust to undetected genotyping errors (Gordon et al. 2001) should be used in conjunction with the standard TDT. Fourth, a more stringent cutoff for significance could be employed, on the basis of the frequency of the allele under study and the sample size. Fifth, investigators should explore the variance in genotyping error rates across the markers in their own data sets. This is particularly important if many markers are considered, as the error rate at an individual marker will have a much smaller effect on the mean error rate in a large pool of markers than in a small one. Every effort should be made to ensure that the error rate is under 0.5% at each marker and to identify markers with higher error rates. If a significant association is identified through any type of analysis at a marker with a higher error rate, it should raise suspicion.

Our analyses underscore the need to continue to improve data quality through the pursuit of increasingly accurate genotyping technology. Highly accurate genotype data will become increasingly critical as human geneticists attempt genomewide association studies using thousands of SNPs in an effort to identify genes of increasingly smaller effect in the etiology of complex traits.

## Acknowledgments

## Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

Marshfield Clinic Human Diversity Panel, http://www.marshfieldclinic.org/research/genetics/Freq/FreqInfo.htm

## 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

Brzustowicz LM, Merette C, Xie X, Townsend L, Gilliam TC, 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

Clarke LA, Rebelo CS, Goncalves J, Boavida MG, Jordan P (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol Pathol 54: 351–353

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A (2001) High-throughput variation detection and genotyping using microarrays. Genome Res 11: 1913–1925

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 66:1287–1298

Douglas JA, Skol AD, Boehnke M (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. Am J Hum Genet 70: 487–495

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

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

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 of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. Genome Research 7:165–178

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

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

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

Goring HHH, Terwilliger JD (2000) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. Am J Hum Genet 66:1107–1118

Hall JM, LeDuc CA, Watson AR, Roter AH (1996) An approach to high-throughput genotyping. Genome Research 6:781–790

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

Idury RM, Cardon LR (1997) A simple method for automated allele binning in microsatellite markers. Genome Res 7: 1104–1109

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

Mansfield DC, Brown AF, Green DK, Carothers AD, Morris SW, Evans HJ, Wright AF (1994) Automation of genetic linkage analysis using fluorescent microsatellite markers. Genomics 24:225–233

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

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

Smith JR, Carpten JD, Brownstein MJ, Ghosh S, Magnuson VL, Gilbert DA, Trent JM, Collins FS (1995) Approach to genotyping errors caused by nontemplated nucleotide addition by *Taq* DNA polymerase. Genome Research 5:312–317

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 70:496–508

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

Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes: consistency between two high-throughput genotyping centers. Genome Research 12:430–435