Genotyping Error Detection Through Tightly Linked Markers

Guohua Zou, Deyun Pan and Hongyu Zhao[1]

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut 06520-8034

Manuscript received May 17, 2002
Accepted for publication March 19, 2003

ABSTRACT

The identification of genotyping errors is an important issue in mapping complex disease genes. Although it is common practice to genotype multiple markers in a candidate region in genetic studies, the potential benefit of jointly analyzing multiple markers to detect genotyping errors has not been investigated. In this article, we discuss genotyping error detections for a set of tightly linked markers in nuclear families, and the objective is to identify families likely to have genotyping errors at one or more markers. We make use of the fact that recombination is a very unlikely event among these markers. We first show that, with family trios, no extra information can be gained by jointly analyzing markers if no phase information is available, and error detection rates are usually low if Mendelian consistency is used as the only standard for checking errors. However, for nuclear families with more than one child, error detection rates can be greatly increased with the consideration of more markers. Error detection rates also increase with the number of children in each family. Because families displaying Mendelian consistency may still have genotyping errors, we calculate the probability that a family displaying Mendelian consistency has correct genotypes. These probabilities can help identify families that, although showing Mendelian consistency, may have genotyping errors. In addition, we examine the benefit of available haplotype frequencies in the general population on genotyping error detections. We show that both error detection rates and the probability that an observed family displaying Mendelian consistency has correct genotypes can be greatly increased when such additional information is available.

THE problem of genotyping errors has received much attention in human genetics because of its importance in the analysis and interpretation of genetic data from linkage and association studies. Terwilliger *et al.* (1990), Buetow (1991), Shields *et al.* (1991), Goldstein *et al.* (1997), Gordon *et al.* (1999), and Akey *et al.* (2001) investigated the effects of genotyping errors on various aspects of genetic data analysis. Lincoln and Lander (1992), Ott (1993), Ehm *et al.* (1996), Stringham and Boehnke (1996), Ehm and Wagner (1998), O'Connell and Weeks (1998), Douglas *et al.* (2000, 2002), and Sobel *et al.* (2002) proposed various methods to detect genotyping errors. Broman and Weber (1998), Göring and Terwilliger (2000a,b,c,d), Gordon and Ott (2001), Gordon *et al.* (2001), and Sobel *et al.* (2002) developed statistical methods for incorporating genotyping errors in the analysis of genotype data. Although it is common practice to genotype multiple tightly linked markers in a candidate region, the use of joint information from these markers to detect genotyping errors has not been investigated in the literature. In this article, we discuss this issue for nuclear families when both parents are available. We call these closely spaced markers "multiple tightly linked markers" to emphasize their close proximity on a chromosome and the fact that recombination is an unlikely event among these markers. In our analysis, the objective is to identify families, not individual markers, that likely have genotyping errors with the hope that these families will be followed up for error checking.

Mendelian consistency is the most common criterion for identifying genotyping errors. Families that fail the Mendelian-consistency check should be flagged out for error checking. In the case of single markers, Gordon *et al.* (1999, 2000) calculated the probabilities that the erroneous trio genotype and quartet and quintet genotypes can be detected on the basis of the Mendelian-consistency criterion. They found that the error detection rates are very low. Douglas *et al.* (2002) calculated the error detection rates in nuclear families by assuming that there is exactly one genotyping error per family. However, these studies examined only error detections for a single marker, and it is worthwhile to study the benefit of considering two or more tightly linked genetic markers. In the absence of phase information and when the genotyping error rate is not very high, we show that there is little to gain from considering multiple tightly linked markers for family trios if Mendelian consistency is the error checking criterion. However, when there is more than one child, error detection rates can be greatly increased by adding markers and including additional children in each family. Instead of calculating error detection rates, we also examine the problem of estimating the probability that a family displaying Mendelian

[1] *Corresponding author:* Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College St., New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

consistency has correct genotypes. These probabilities may be more relevant than error detection rates as these probability calculations allow the researchers to prioritize families if they are willing to confirm genotypes through replicate genotypings. We further consider the benefit of having information on population haplotype frequencies. We find that both error detection rates and the probability that a family displaying Mendelian consistency has correct genotypes may be greatly increased if such additional information is available.

METHODS

In this section, we discuss our methods for deriving analytical results of error detection rates for two tightly linked markers and the probability that a family displaying Mendelian consistency has correct genotypes for family trios with one or two markers. We then outline our simulation procedures for nuclear families with multiple children, multiple markers, and multiple alleles.

Error detection rates for family trios with two markers: Consider family trios where each individual is typed at two biallelic markers. The two markers, denoted by $\mathscr{A}$ and $\mathscr{B}$, have alleles A_1/A_2 and B_1/B_2, respectively. For simplicity, in the following discussion we denote A_1 and A_2 by 1 and 2, respectively, and similarly denote B_1 and B_2 by 1 and 2, respectively.

We use 2 × 2 matrices to denote two-marker diploid genotype data, where elements in each column represent the two alleles at the same marker. When phase information is known, elements in the same row represent the alleles on the same chromosome, and the two rows are exchangeable. For example, matrices

$$\begin{pmatrix}1&1\\2&2\end{pmatrix} \text{ and } \begin{pmatrix}2&2\\1&1\end{pmatrix}$$

both represent an individual with one chromosome carrying (11) and one chromosome carrying (22). We make no distinction between these two matrices in our following discussion. In addition, no distinction is made between parent 1 and parent 2. For example, the trio genotypes

$$\left(\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}1&2\\2&2\end{pmatrix}\begin{pmatrix}1&1\\1&2\end{pmatrix}\right)$$

and

$$\left(\begin{pmatrix}1&2\\2&2\end{pmatrix}\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}1&1\\1&2\end{pmatrix}\right)$$

are regarded as equivalent. For a family with genotype M, we define the conjugate of M, denoted by $\underline{M}$, as the genotype with each 1 in M replaced by 2 and each 2 in M replaced by 1 (Gordon *et al.* 2000). For each individual, 10 distinct matrices correspond to 10 possible haplotype pairs. For family trios, a total of 136 possible trio genotypes show Mendelian consistency, and this set is denoted by S.

We now define Mendelian consistency for two or more markers. If phase is known in both parents, let $(H_1^{\mathrm{P}}, H_2^{\mathrm{P}})$ denote the two haplotypes in the father, and $(H_1^{\mathrm{M}}, H_2^{\mathrm{M}})$ denote the two haplotypes in the mother. For tightly linked markers, we say the trio is Mendelian consistent if the child has one of the following genotypes, $(H_1^{\mathrm{P}}, H_1^{\mathrm{M}})$, $(H_1^{\mathrm{P}}, H_2^{\mathrm{M}})$, $(H_2^{\mathrm{P}}, H_1^{\mathrm{M}})$, or $(H_2^{\mathrm{P}}, H_2^{\mathrm{M}})$; *i.e.*, each parent passes one of the two whole haplotypes intact to the offspring. We expect this to be the case, in general, for tightly linked markers as recombinations are unlikely among them.

However, phase information is usually unknown. In this case, genotypes

$$\begin{pmatrix}1&1\\2&2\end{pmatrix} \text{ and } \begin{pmatrix}1&2\\2&1\end{pmatrix}$$

are not distinguishable. It is important to keep this in mind when we calculate the error detection rate and the probability that a family trio displaying Mendelian consistency has correct genotypes. For phase-unknown data, there may be multiple haplotype sets in the parents that are consistent with the observed genotypes across the set of markers. In this case, we say that the trio is Mendelian consistent if one of the haplotype sets is Mendelian consistent in the sense defined above for phase-known data.

For genotyping errors, we assume that errors are introduced independently. At marker $\mathscr{A}$, the genotyping error rate from true allele 1 to erroneous allele 2 is e_1 and from true allele 2 to erroneous allele 1 is e_2. At marker $\mathscr{B}$, the genotyping error rates from 1 to 2 and from 2 to 1 are ε_1 and ε_2, respectively. This general error model includes the stochastic error model ($e_1 = e_2 = \varepsilon_1 = \varepsilon_2$) and the directed error model ($e_2 = \varepsilon_2 = 0$) as special cases (Akey *et al.* 2001).

For each trio genotype M, 0–12 errors may be introduced for two markers in a family trio. We say a family has undetected errors if the trio is Mendelian consistent. The probability that the errors are not detected via a Mendelian-consistency check, when there is at least one error, is

$$\beta = \sum_{i=1}^{12} P(\text{undetected errors}|i \text{ errors in trio})P(i \text{ errors in trio}|i \geq 1), \tag{1}$$

where the first conditional probability can be calculated as

$$\begin{aligned}
&P(\text{undetected errors}|i \text{ errors in trio})\\
&\quad = \sum_{M \in S} P(\text{undetected errors}|i \text{ errors in } M)P(M|i \text{ errors in trio})\\
&\quad = \sum_{M \in S} P(\text{undetected errors}|i \text{ errors in } M)P(i \text{ errors in } M)\\
&\qquad \cdot P(M)/P(i \text{ errors in trio}),
\end{aligned}$$

where S is the set of all family trio genotypes. Thus, Equation 1 is simplified to

$$\beta = \sum_{i=1}^{12} \sum_{M \in S} P(\text{undetected errors}|i \text{ errors in } M) \cdot P(i \text{ errors in } M)P(M)/[1 - P(\text{no error in trio})]. \quad (2)$$

The calculations of the probabilities that i errors exist in genotype M and no error exists in trio in Equation 2 are discussed in APPENDIX A. Note that for the stochastic error model, the probabilities that any trio genotype has i errors are the same. In this case, Equation 2 reduces to that in GORDON *et al.* (2000).

In our calculations, we first calculate P(undetected errors|i errors in M) for $1 \leq i \leq 6$, and for the cases of $7 \leq i \leq 12$, we can obtain the probabilities through the conjugate genotype $\underline{M}$ and the following lemma. Results similar to Lemma 1(i) were derived by GORDON *et al.* (1999, 2000). The values of the probability $P(M)$ and the conditional probability P(undetected errors|i errors in M) ($1 \leq i \leq 6$) are available from the authors upon request.

LEMMA 1. (i) *For any trio genotype M and for any i,* $0 \leq i \leq 12$, *we have* (a) P(*undetected errors*|*i errors in M*) = P(*undetected errors*|*i errors in M*) *and* (b) P(*undetected errors*|*i errors in M*) = P(*undetected errors*|12 − *i errors in M*).

(ii) *If* $P(M) = f(p_{11}, p_{12}, p_{21}, p_{22})$, *where f is a function of* p_{11}, p_{12}, p_{21}, *and* p_{22}; *and* p_{ij} *is the frequency of haplotype ij, where* $i, j = 1, 2$; *then*

$$P(\underline{M}) = f(p_{22}, p_{21}, p_{12}, p_{11}).$$

Error detection rates for nuclear families with more than one child and more than two markers: For the general case of multiple children, multiple markers, and multiple alleles, we conduct simulation studies to obtain error detection rates as follows:

1. We generate the genotypes of the parents according to a set of haplotype frequencies $p_{i_1 \ldots i_k}$ ($i_1 = 1, \ldots, I_1; \ldots; i_k = 1, \ldots, I_k$), where k is the number of tightly linked markers, and I_j is the number of alleles at marker j. On the basis of parental haplotypes, we simulate haplotype pairs in the children by randomly assigning one of the two haplotypes in each parent to each child. Then we introduce errors independently into the alleles of parents and children according to a given error model. On the basis of the resulting genotypes for the parents in the nuclear family, we obtain all haplotype pairs that are consistent with the genotypes of the parents.
2. We number the children in the family by the number of homozygous sites; *e.g.*, after numbering, child 1 in the family has the largest number of markers with two identical copies of an allele. For the first child, we consider all possible haplotype pairs that are consistent with this child's genotype. For each haplotype pair in the consistent haplotype pair set, we use the procedure described in APPENDIX B to (a) identify whether this pair is consistent with the parents' possible haplotype pairs and (b) if yes, determine possible haplotype pairs for other children based on this pair. If none of the haplotype pairs for the first child is consistent with the parents' haplotype pairs, then we say we have detected genotyping errors. Otherwise, we collect all the possible haplotype pairs for other children based on the first child and call this set C_1.
3. Consider child 2 in the family. If no haplotype pairs consistent with this child's genotype belong to C_1, then genotyping errors are detected. Otherwise, discard the haplotype pairs that are not consistent with the second child's genotype and call the remaining set C_2.
4. Repeat steps 2 and 3 until the nth child (assuming this family has n children) is checked and we end up with a set C_n. If C_n is empty, the errors are detected. Otherwise, the whole family is consistent with Mendelian inheritance.

To estimate error detection rates, we base our results on 100,000 simulations for single markers, 10,000 simulations when the number of markers is two or three, and 5000 simulations when there are four markers. Different numbers of simulations are used because the true error detection rates vary according to the number of markers, with lower detection rates for smaller numbers of markers. Therefore, a larger number of simulations are necessary when the number of markers is smaller.

Probability that a family trio displaying Mendelian consistency has correct genotypes: In addition to calculating error detection rates, another quantity that is of relevance is the probability that an observed trio displaying Mendelian consistency has correct genotypes. We first discuss the single-marker case. There are a total of nine trio genotypes with one marker, which is denoted by S_0. We use similar notation on trio genotypes as in the two-marker case. For example, the following trio genotypes are considered equivalent:

$$\left(\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\right) \quad \text{and} \quad \left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\right).$$

With similar genotyping error models, we can derive the probability that i errors are introduced in the trio in the one-marker case. Note that "an observed trio has correct genotypes" is not equivalent to "there is no genotyping error in the trio." For example, for an individual with genotype 12 at one marker, there may be two errors with 1→2 and 2→1, but the observed genotype is true.

For an observed genotype M that is Mendelian consistent, the probability that it is the true genotype is given by

$$P(T = M|O = M) = \frac{P(O = M|T = M)P(T = M)}{P(O = M)}, \quad (3)$$

where $P(T = M)$ is the probability that the true trio genotype is M, and $P(O = M)$ is the probability that the observed trio genotype is M, which is

$$P(O = M) = \sum_{M' \in S_0} P(O = M|T = M')P(T = M'), \quad (4)$$

where the set of S_0 was defined above. An example for calculating the conditional probability $P(T = M|O = M)$ is provided in APPENDIX C.

In general, in addition to calculating $P(T = M|O = M)$ for a given genotype, we can also calculate the *overall* probability that a Mendelian-consistent family trio has correct genotypes by summing over all possible genotypes:

$$\begin{aligned} P(\text{true}|\text{a Mendelian-consistent trio}) &= \frac{\sum_{M \in S_0} P(T = M, O = M)}{\sum_{M \in S_0} P(O = M)} \\ &= \frac{\sum_{M \in S_0} P(T = M|O = M)P(O = M)}{\sum_{M \in S_0} P(O = M)}. \end{aligned} \quad (5)$$

It is readily seen that Equation 5 can also be expressed as

$$P(\text{true}|\text{a Mendelian-consistent trio}) = \frac{\sum_{M \in S_0} P(O = M|T = M)P(T = M)}{\sum_{M \in S_0} P(O = M)}.$$

In deriving the probabilities, we use the following lemma for the one-marker case.

LEMMA 2. (i) *Let* $P(O = M_0|T = M) = u(e_1, e_2)$. *Then* (a) $P(O = \underline{M}_0|T = M) = u(1 - e_1, 1 - e_2)$ *and* (b) $P(O = M_0|T = \underline{M}) = u(1 - e_2, 1 - e_1)$.
(ii) *Let* $P(T = M_0|O = M_0) = v(p, e_1, e_2)$. *Then* $P(T = \underline{M}_0|O = \underline{M}_0) = v(q, e_2, e_1)$, *where* p *is the frequency of allele* 1 *and* $q = 1 - p$.

For the case of two markers, 125 distinct trio genotypes display Mendelian consistency in the absence of phase information. The general results (3) and (4) still hold. In the calculation of terms in (3) and (4), using the same genotyping error model discussed before, we have

LEMMA 3. (i) *Let* $P(O = M_0|T = M) = g(e_1, e_2, \varepsilon_1, \varepsilon_2)$. *Then* (a) $P(O = \underline{M}_0|T = M) = g(1 - e_1, 1 - e_2, 1 - \varepsilon_1, 1 - \varepsilon_2)$ *and* (b) $P(O = M_0|T = \underline{M}) = g(1 - e_2, 1 - e_1, 1 - \varepsilon_2, 1 - \varepsilon_1)$.
(ii) *Let* $P(T = M_0|O = M_0) = h(p_{11}, p_{12}, p_{21}, p_{22}, e_1, e_2, \varepsilon_1, \varepsilon_2)$. *Then* $P(T = \underline{M}_0|O = \underline{M}_0) = h(p_{22}, p_{21}, p_{12}, p_{11}, e_2, e_1, \varepsilon_2, \varepsilon_1)$.

RESULTS

Error detection rates for family trios: Let the frequencies of alleles 1 and 2 at marker $\mathscr{A}$ be p_{1+} and p_{2+} ($= 1 - p_{1+}$), respectively. Similarly, we denote the marker allele frequencies at marker $\mathscr{B}$ by p_{+1} and p_{+2}, respectively. The haplotype frequencies are denoted by p_{11}, p_{12}, p_{21}, and p_{22}, respectively. For different sets of haplotype frequencies, we summarize the results in Table 1 when the error rates are assumed to be the same. The results are qualitatively similar when the error rates differ (data not shown). In addition, we considered the following three cases in more detail.

1. Linkage equilibrium: In this case, $p_{11} = p_{1+}p_{+1}$, $p_{12} = p_{1+}p_{+2}$, $p_{21} = p_{2+}p_{+1}$, and $p_{22} = p_{2+}p_{+2}$.
2. *Perfect* linkage disequilibrium (equal allele frequencies): In this case, only two of the four possible haplotypes are present in the population. Without loss of generality, we assume that $p_{11} = p_{1+} = p_{+1}$, $p_{12} = 0$, $p_{21} = 0$, and $p_{22} = p_{2+} = p_{+2}$.
3. *Complete* linkage disequilibrium (unequal allele frequencies): In this case, three of the four haplotypes are present in the population. Without loss of generality, we assume haplotypes 11, 12, and 21 are present.

Table 1 and the results for the above three special cases (data not shown) indicate that the error detection rates are generally low when the error rates are low if Mendelian consistency is the criterion for error checking. When the error rates are high (>20%), the error detection rates based on two markers can be significantly higher than those based on single markers, even higher than those for the case of quartet considered by GORDON *et al.* (2000). However, such high error rates are not common in practice. Compared to the results for single markers (GORDON *et al.* 1999), considering two markers only slightly increases the error detection

TABLE 1

Error detection rates for trios with two markers when e_1, e_2, ε_1, and ε_2 are all equal

True error rate	$p_{11} = 0.2$, $p_{12} = 0.3$, $p_{21} = 0.3$, $p_{22} = 0.2$	$p_{11} = 0.4$, $p_{12} = 0.4$, $p_{21} = 0.1$, $p_{22} = 0.1$	$p_{11} = 0.4$, $p_{12} = 0.2$, $p_{21} = 0.3$, $p_{22} = 0.1$
0.0010	0.2507	0.2656	0.2590
0.0050	0.2534	0.2681	0.2616
0.0100	0.2569	0.2713	0.2649
0.0200	0.2639	0.2777	0.2716
0.0500	0.2852	0.2973	0.2920
0.1000	0.3213	0.3306	0.3265
0.2000	0.3893	0.3939	0.3919
0.3000	0.4419	0.4436	0.4428

TABLE 2

Error detection rates for nuclear families with n children and k markers when the markers are in linkage equilibrium

n	$k = 1$ $(p = p_1^*)$	$k = 2$ $(p = p_2^*)$	$k = 3$ $(p = p_3^*)$	$k = 4$ $(p = p_4^*)$
1	0.3059	0.3064	0.3209	0.3217
2	0.4325	0.5076	0.5669	0.5898
3	0.5042	0.5640	0.6381	0.7111
4	0.5510	0.6248	0.6789	0.7379
5	0.5808	0.6254	0.7414	0.7937
6	0.5942	0.7014	0.7546	0.8137

The true error rate is 0.01 and the allele frequencies are unequal.
$p_1^* = (0.9, 0.1)$
$p_2^* = (0.81, 0.09, 0.09, 0.01)$
$p_3^* = (0.729, 0.081, 0.081, 0.009, 0.081, 0.009, 0.009, 0.001)$
$p_4^* = (0.6561, 0.0729, 0.0729, 0.0081, 0.0729, 0.0081, 0.0081, 0.0009, 0.0729, 0.0081, 0.0081, 0.0009, 0.0081, 0.0009, 0.0009, 0.0001)$.

rate when the error rates are low (≤5%). This is not unexpected as we can show (APPENDIX D) that if trios are Mendelian consistent for each individual marker, then the trio genotype is Mendelian consistent across all the markers even with the use of multiple tightly linked markers. Therefore, error checking through Mendelian consistency offers little more information for family trios in the absence of additional information, *e.g.*, phase and/or population genotypes.

Error detection rates for families with more than one child: When additional family members are available, joint consideration of two tightly linked markers offers more information than single markers. For example, consider a family with both parents and two children. Let the two-marker quartet genotype be

$$\left(\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\1&2\end{pmatrix}\begin{pmatrix}1&1\\1&1\end{pmatrix}\right),$$

where the first two matrices denote the parents' genotypes and the last two matrices denote the children's genotypes. It is easy to see that, although the two one-marker quartet genotypes

$$\left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\right) \quad \text{and} \quad \left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\right)$$

show Mendelian consistency, the two-marker quartet genotype does not. On the basis of this, it can be expected that if we consider two-marker quartet genotypes, the error detection rate will be increased, as evidenced from our simulation results shown in Tables 2 and 3, where n and k denote the numbers of children and markers, respectively, and I_j denotes the number of alleles at marker j $(j = 1, \ldots, k)$.

Our results show that when more than one child is in a nuclear family, the error detection rates can be greatly increased by adding additional markers. Furthermore, the error detection rates increase with the number of children. The rate of increase is the greatest from one child to two children, and there is usually not much difference between having five or six children.

TABLE 3

Error detection rates for nuclear families with n children and k markers when the markers are in linkage equilibrium

n	$k = 1$ $(p = p_1^*)$	$k = 2$ $(p = p_2^*)$	$k = 3$ $(p = p_3^*)$	$k = 4$ $(p = p_4^*)$
1	0.2485	0.2580	0.2600	0.2968
2	0.3358	0.5022	0.6551	0.7003
3	0.3778	0.5393	0.7499	0.8211
4	0.3839	0.6234	0.7714	0.8842
5	0.4006	0.6254	0.8152	0.9038
6	0.4092	0.6472	0.8197	0.9119

The true error rate is 0.01 and the allele frequencies are equal.
$p_1^* = (0.5, 0.5)$
$p_2^* = (0.25, 0.25, 0.25, 0.25)$
$p_3^* = (0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)$
$p_4^* = (0.0625, \ldots\ldots, 0.0625)$, where the length of the vector is 16.

TABLE 4

Error detection rates for nuclear families with *n* children and one marker with eight alleles of equal frequency

True error rate	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
0.0100	0.6110	0.7686	0.8504	0.8698	0.8944	0.8992
0.0500	0.7742	0.9266	0.9672	0.9855	0.9928	0.9963

In addition to considering biallelic markers, we also consider markers with multiple alleles and the results are summarized in Table 4. It can be seen that, as expected, the error detection rates are higher for the case of multiple alleles. Comparing the results of Table 4 with those of the third column of Table 3 is interesting because Table 4 is for a marker with eight alleles of equal frequency and column 3 of Table 3 is for a haplotype system with eight haplotypes having equal frequencies. Although there is substantial difference between error detection rates when there is only one child and when there are two children, the difference for the case of multiple children becomes smaller when the number of children is larger.

Probability that a family trio displaying Mendelian consistency has correct genotypes: In addition to error detection rates, the probability that a family with Mendelian consistency has correct genotypes may be of more relevance as these probabilities will help the investigators to prioritize families for genotyping error checking among Mendelian-consistent families.

Figure 1 reveals the results of the probability that a trio with Mendelian consistency has correct genotypes for the case of single markers when the error rates are the same. It is seen that when the true error rates are low (≤1%), most of the Mendelian-consistent trios have correct genotypes. On the other hand, when the error rates are high (≥20%), most of the Mendelian-consistent trios have incorrect genotypes. A similar observation can be obtained for different values of e_1 and e_2 (see Figures 2 and 3). Furthermore, it can be seen from the probability curves for $p = 0.1$ and 0.5 in Figure 1 that, when $e_1 = e_2$, the probability that a trio displaying consistency has correct genotypes is not much affected by allele frequencies.

In the case of two markers, we consider the cases for $p_{11} = p_{12} = 0.4$, $p_{21} = p_{22} = 0.1$, and $e_1 = e_2 = \varepsilon_1 = \varepsilon_2 = 0.005$; $p_{11} = 0.4$, $p_{12} = 0.2$, $p_{21} = 0.3$, $p_{22} = 0.1$, and $e_1 = e_2 = \varepsilon_1 = \varepsilon_2 = 0.01$; and the three special cases considered in error detection rates. Our analytical results (data not shown) indicate that if the error rates are low (≤0.5%), the probability of an observed genotype being true is >95%, which implies that a trio genotype displaying consistency is often true. If the error rates are between 0.5 and 2%, then the trio genotypes still tend to be true. However, if the error rates are large (≥10%), then the probability is often <40%, which means that a trio genotype displaying consistency is usually not true. As in the case of one marker, the probability that an observed genotype displaying consistency is correct is only slightly affected by haplotype frequencies under the stochastic error model.

Probability that a nuclear family with more than one child displaying Mendelian consistency has correct genotypes: For the general case of multiple children and multiple markers, we conduct simulation studies to estimate the probability that a family displaying Mendelian consistency has correct genotypes. The simulation results are presented in Table 5. It can be seen from the table that when the error rate is 0.01, the probability that a family displaying Mendelian consistency has correct genotypes is high, even though multiple children and

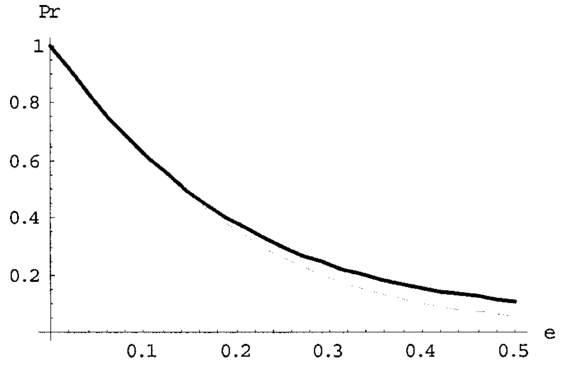

FIGURE 1.—The probability that an observed genotype with one marker is true for $p = 0.1$ and 0.5 when $e_1 = e_2 = e$ (the dotted line denotes the curve for $p = 0.1$).

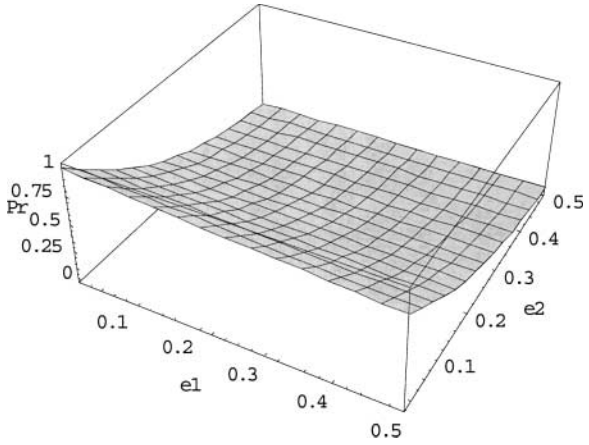

FIGURE 2.—The probability that an observed genotype with one marker is true when $p = 0.1$.

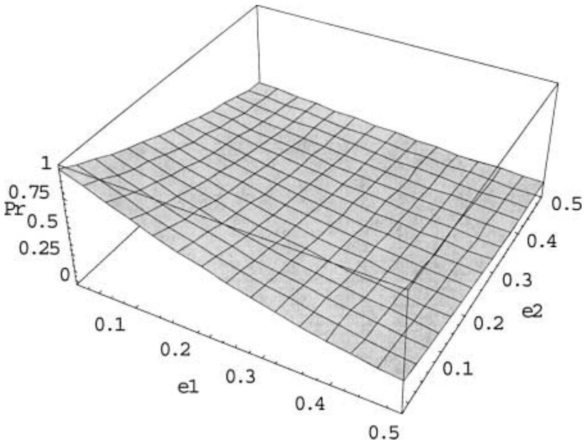

FIGURE 3.—The probability that an observed genotype with one marker is true when $p = 0.5$.

multiple markers are considered. Also, except for the case of only one child, these probabilities are very similar when the number of children differs.

Use of population haplotype information: One way to increase error detection rates is to make use of some other information. Consider the special case of perfect linkage disequilibrium with equal allele frequencies. In this scenario, there are only two possible haplotypes in the population, (11) and (22), and there are a total of 10 possible trio genotypes. Therefore, any patterns differing from these 10 will be identified as caused by genotyping errors. The conditional probabilities that a family trio with i errors ($1 \leq i \leq 6$) will be undetected, *i.e.*, fall into one of the 10 categories, are available from the authors upon request. Note that we can still use Lemma 1 to calculate the probabilities when i is between 7 and 12. When all the error rates are the same, the error detection rates are as summarized in Table 6. As expected, the error detection rates are indeed greatly increased.

TABLE 6

Error detection rates for trios when e_1, e_2, ε_1, and ε_2 are all equal in the presence of the information on perfect linkage disequilibrium (LD) with equal allele frequencies

True error rate	$p_{1+} = p_{+1} = 0.1$	$p_{1+} = p_{+1} = 0.5$
0.0010	0.9993	0.9992
0.0050	0.9964	0.9958
0.0100	0.9929	0.9918
0.0200	0.9865	0.9844
0.0500	0.9707	0.9662
0.1000	0.9541	0.9475
0.2000	0.9444	0.9388
0.3000	0.9482	0.9457

Such additional information will also affect the calculation of the probability that a family displaying Mendelian consistency has correct genotypes. Still consider the above case where two markers are in perfect linkage disequilibrium with equal allele frequency. The conditional probability that a Mendelian-consistent trio is true is presented in Table 7. It is apparent that even when the error rates are as high as 10%, the probability that an observed Mendelian-consistent trio is true is quite high.

DISCUSSION

In this article, we investigated genotyping error detections through multiple tightly linked markers in nuclear families. Our error detection rate is calculated using families, not markers, as a unit, with the objective of being able to identify families having genotyping errors. We first calculated the error detection rates for family trios with two markers using an analytical method. We showed that in the absence of phase information, genotyping errors can be detected if and only if there is Mendelian inconsistency at one or more of the markers. This means that only the information on each marker is helpful for detecting genotyping errors. Joint consideration of multiple tightly linked markers will not pro-

TABLE 5

The values of P(true|a Mendelian-consistent family) for markers in linkage equilibrium

n	$k = 1$ ($p = p_1$*)	$k = 2$ ($p = p_2$*)	$k = 3$ ($p = p_3$*)	$k = 4$ ($p = p_4$*)
1	0.9567	0.9130	0.8761	0.8382
2	0.9556	0.9170	0.8970	0.8677
3	0.9506	0.9078	0.8870	0.8737
4	0.9451	0.9056	0.8830	0.8704
5	0.9399	0.8975	0.8674	0.8679
6	0.9338	0.8928	0.8642	0.8534

The true error rate is 0.01.
p_1* = (0.9, 0.1)
p_2* = (0.81, 0.09, 0.09, 0.01)
p_3* = (0.729, 0.081, 0.081, 0.009, 0.081, 0.009, 0.009, 0.001)
p_4* = (0.6561, 0.0729, 0.0729, 0.0081, 0.0729, 0.0081, 0.0081, 0.0009, 0.0729, 0.0081, 0.0081, 0.0009, 0.0081, 0.0009, 0.0009, 0.0001).

TABLE 7

The values of P(true|a Mendelian-consistent trio) when e_1, e_2, ε_1, and ε_2 are all equal in the presence of the information on perfect LD with equal allele frequencies

True error rate	$p_{1+} = p_{+1} = 0.1$	$p_{1+} = p_{+1} = 0.5$
0.0010	1.0000	1.0000
0.0050	0.9999	0.9999
0.0100	0.9996	0.9995
0.0200	0.9983	0.9981
0.0500	0.9885	0.9876
0.1000	0.9495	0.9463
0.2000	0.7671	0.7616
0.3000	0.4622	0.4762

vide more information. Therefore, the error detection rates will not be greatly increased when the error rates are low. As a result, the error detection rates are generally low if Mendelian consistency is used as the unique criterion for checking errors. However, when more than one child is in a family, joint consideration of tightly linked markers can offer more information than single markers. In fact, the error detection rates can be greatly increased by adding tightly linked markers.

Tables 2 and 3 reveal different properties between markers with equal and unequal allele frequencies: If the number of markers k is small (≤ 2) or only one child is in the family, the error detection rates for markers with unequal allele frequencies are greater than those for markers with equal allele frequencies. However, if there are more than two markers and more than one child is in the family, the error detection rates for markers with equal allele frequencies are greater. This is also seen for the case of linkage disequilibrium (data not shown). An explanation for this phenomenon is as follows. For the case of unequal haplotype (allele) frequencies, the genotype of the first child can often be used to detect the error, but for the case of equal haplotype (allele) frequencies, the genotype of the first child is often used to determine the haplotypes of parents and often cannot be used to detect the error except for the case of $k = 1$. Thus, when $n = 1$, the error detection rates for unequal haplotype (allele) frequencies are greater. If k is not small, the errors will often be introduced into each of the genotypes of parents and children, and the errors are often easier to detect for the case of equal haplotype (allele) frequencies because when k becomes large, more and more alleles at each marker will be heterozygous for the case of unequal haplotype (allele) frequencies but the genotype at each marker is more likely to change to homozygotes for the case of equal haplotype (allele) frequencies. Let us consider an extreme case of the following two three-marker genotypes in a family with two parents and two children,

$$\left(\begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}\begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}\begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}\begin{pmatrix}1&1&1\\1&1&1\end{pmatrix}\right) \tag{6}$$

and

$$\left(\begin{pmatrix}1&1&1\\2&2&2\end{pmatrix}\begin{pmatrix}1&1&1\\2&2&2\end{pmatrix}\begin{pmatrix}1&1&1\\2&2&2\end{pmatrix}\begin{pmatrix}1&1&1\\2&2&2\end{pmatrix}\right). \tag{7}$$

The first configuration is more likely to occur if the allele frequencies are 0.9 and 0.1 at each marker, and the second one is more likely to occur if the allele frequencies are 0.5 and 0.5 at each marker. If only one error is introduced into some marker for each person, say marker 1 for parent 1, marker 2 for parent 2, then when the genotypes of parents in (6) become

$$\begin{pmatrix}1&1&1\\2&1&1\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1&1&1\\1&2&1\end{pmatrix},$$

the probability that the errors can be detected is $20(1 - \varepsilon)^{10}\varepsilon^2$ (where ε is the genotyping error rate from true allele 1 to erroneous allele 2 and from true allele 2 to erroneous allele 1); and when the genotypes of parents in (7) become

$$\begin{pmatrix}1&1&1\\1&2&2\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1&1&1\\2&1&2\end{pmatrix}$$

$$\left(\text{or } \begin{pmatrix}1&1&1\\1&2&2\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1&2&1\\2&2&2\end{pmatrix}\right),$$

the probability of error detection is $22(1 - \varepsilon)^{10}\varepsilon^2$. The difference between the former and the latter is $-2(1 - \varepsilon)^{10}\varepsilon^2$. On the other hand, if the family trio is considered [*i.e.*, only one child is considered in (6) and (7)], then the corresponding difference is $2(1 - \varepsilon)^5\varepsilon - 2(1 - \varepsilon)^5\varepsilon = 0$. Note that when k is small, the possibility that the errors are introduced into each of the genotypes of parents and children is not great. For this case, the possibility that the haplotypes of parents can be determined through the first child is small.

We have also examined error detections for multiallelic markers and the error detection rates are greater for equal allele frequencies (see Table 4). This can be readily understood by noting that unlike the case of biallelic markers, for the case of multiallelic markers, the errors in the genotypes of parents have greater effect on error detections. Although haplotypes can be thought of as a multiallelic marker, the error detection rates are lower for a haplotype system than for a multiallelic marker with the same allele frequencies as the set of haplotype frequencies. However, the difference is smaller when a larger number of children are considered in a nuclear family.

The probability formula derived in this article, *e.g.*, the probability that i errors are introduced under the general error model, can be used to calculate error detection rates for other sampling types such as quartet under the general error model. For example, for the case of quartet considered by Gordon *et al.* (2000), if

$e_1 = 0.005$, $e_2 = 0.01$, and $p = 0.5$, we can obtain the error detection rate to be 0.3374.

In addition to error detection rates, we have also calculated the probability that a family displaying Mendelian consistency has correct genotypes. The calculations of such quantities are useful as they may point to certain families that, although showing Mendelian consistency, are likely to have genotyping errors. The calculations require haplotype frequencies from the population and estimated error rates. A potential application of calculating these probabilities is to conduct transmission/disequilibrium tests in the presence of genotyping errors. We showed that when the error rates are low, the overall probability that a Mendelian-consistent trio has correct genotypes is quite high, and the overall probability is not very sensitive to haplotype frequencies under the stochastic error model. We expect that the number of families showing Mendelian consistency and having correct genotypes decreases with the increase of the number of children in the family and the number of markers. Our simulation results indeed show this property (data not shown). On the other hand, our simulation results show that conditional on a family showing Mendelian consistency, the probability that this family has correct genotypes is not a monotonic function of n and k. We offer an explanation by considering two markers with equal allele frequencies: $p_1 = p_2 = 0.5$. Consider the following two-marker genotype,

$$\left(\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\right), \tag{8}$$

which is the most common family configuration. After the errors are introduced, if the parents and the first child show Mendelian consistency and they have correct genotypes, then their genotypes must be

$$\left(\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\right).$$

Now we consider the case of adding another child, that is, a quartet (see Equation 8). After the errors are introduced into the genotype of the second child, P(the resulting quartet genotype is Mendelian consistent) $= 1 - 4 \times (1 - 0.01)^3 \times 0.01 - 4 \times (1 - 0.01) \times 0.01^3 = 0.9612$, and P(the genotype of the second child is correct$|$the resulting quartet genotype is consistent) $= [(1 - 0.01)^4 + 2 \times (1 - 0.01)^2 \times 0.01^2 + 0.01^4]/0.9612 = 0.9996$. Thus, the ratio of the probabilities that the genotype is true and consistent is 1.04. This shows that conditional on Mendelian consistency, the probability of having correct genotypes becomes larger when an additional child is added.

For the case of one marker, we consider the following one-marker genotype:

$$\left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\right). \tag{9}$$

After the errors are introduced, if the parents and the first child show Mendelian consistency and they have correct genotypes, then their genotypes must be

$$\left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\right).$$

If we add one child, then after the errors are introduced into the genotype of the second child (see Equation 9), P(the resulting quartet genotype is consistent) $= 1$, and P(the genotype of the second child is true$|$the resulting quartet genotype is consistent) $= 1 - 2 \times (1 - 0.01) \times 0.01 = 0.9802$. Thus, the ratio of the probabilities that the genotype is correct and consistent is 0.9802, which means that conditional on Mendelian consistency, the probability of having correct genotypes becomes smaller when an additional child is added.

If the phase information is known, errors can be detected although each individual marker shows Mendelian consistency. For example, consider a family whose individual marker genotypes are

$$\left(\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\right) \quad \text{and} \quad \left(\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\right)$$

and their haplotypes are

$$\left(\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}1&2\\2&1\end{pmatrix}\begin{pmatrix}1&1\\2&2\end{pmatrix}\right). \tag{10}$$

Although each marker is Mendelian consistent, the joint haplotypes are not, unless we assume there is a recombination event between these two tightly linked markers. Therefore, when phase information is available, error detection rates may be improved. However, phase information may be difficult to obtain except through some molecular techniques. Instead, we have examined the benefit of perfect linkage disequilibrium information, which can be regarded as partial haplotype information in genotyping error detections. We have considered a situation where only two of the haplotypes are known to exist in a given population. In this case, utilizing this information may significantly increase the chance to detect errors through tightly linked markers and increase the confidence that a Mendelian-consistent trio has correct genotypes, and this line of research is worth pursuing.

In this article, we have considered tightly linked markers by assuming no recombination events among these markers. If we allow the occurrence of recombinations, there would be little benefit from using a Mendelian-consistency check as the only criterion for identifying families with genotyping errors. However, if reliable estimates of recombination fractions among these markers are available, we can calculate the probability for each family, incorporating recombination fraction information as well as population haplotype frequency information if it is available. Therefore, although fewer families can be detected as having genotyping errors purely on the basis of Mendelian-consistency check, we are still able to order families by the likelihoods of their genotypes and pursue those with very small likelihoods to be observed.

The authors are grateful to the two reviewers for their valuable comments and suggestions, which greatly improved the original manuscript. This work was supported in part by grant GM59507 from the National Institutes of Health.

LITERATURE CITED

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

Broman, K. W., and J. L. Weber, 1998 Estimation of pairwise relationships in the presence of genotyping errors. Am. J. Hum. Genet. **63:** 1563–1564.

Buetow, K. H., 1991 Influence of aberrant observations on high-resolution linkage analysis outcomes. Am. J. Hum. Genet. **49:** 985–994.

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

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

Ehm, M. G., M. Kimmel and R. W. Cottingham, Jr., 1996 Error detection for pedigree data, using likelihood methods. Am. J. Hum. Genet. **58:** 225–234.

Ehm, M. G., and M. Wagner, 1998 A test statistic to detect errors in sib-pair relationships. Am. J. Hum. Genet. **62:** 181–188.

Goldstein, D. R., H. Zhao and T. P. Speed, 1997 The effects of genotyping errors and interference on estimation of genetic distance. Hum. Hered. **47:** 86–100.

Gordon, D., and J. Ott, 2001 Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac. Symp. Biocomput. **6:** 18–29.

Gordon, D., S. C. Heath and J. Ott, 1999 True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Hum. Hered. **49:** 65–70.

Gordon, D., S. M. Leal, S. C. Heath and J. Ott, 2000 An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. Pac. Symp. Biocomput. **5:** 663–674.

Gordon, D., S. C. Heath, X. Liu and J. Ott, 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.

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

Göring, H. H. H., and J. D. Terwilliger, 2000b Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. Am. J. Hum. Genet. **66:** 1107–1118.

Göring, H. H. H., and J. D. Terwilliger, 2000c Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. Am. J. Hum. Genet. **66:** 1298–1309.

Göring, H. H. H., and J. D. Terwilliger, 2000d 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.

Lincoln, S. E., and E. S. Lander, 1992 Systematic detection of errors in genetic linkage data. Genomics **14:** 604–610.

O'Connell, J. R., and D. E. Weeks, 1998 PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am. J. Hum. Genet. **63:** 259–266.

Ott, J., 1993 Detecting marker inconsistencies in human gene mapping. Hum. Hered. **43:** 25–30.

Shields, D. C., A. Collins, K. H. Buetow and N. E. Morton, 1991 Error filtration, interference, and the human linkage map. Proc. Natl. Acad. Sci. USA **88:** 6501–6505.

Sobel, E., J. Papp and K. Lange, 2002 Detection of genotyping errors. Am. J. Hum. Genet. **70:** 496–508.

Stringham, H. M., and M. Boehnke, 1996 Identifying marker typing incompatibilities in linkage analysis. Am. J. Hum. Genet. **59:** 946–950.

Terwilliger, J. D., D. E. Weeks and J. Ott, 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. **47** (Suppl.): A201.

Communicating editor: Y.-X. Fu

APPENDIX A: THE CALCULATION OF $P(i$ ERRORS IN $M)$

For family M, let ξ_{Mi} be the number of errors at marker i, where $i = 1, 2$. Then we have

$$P(i \text{ errors in } M) = P(\xi_{M1} + \xi_{M2} = i) = \sum_{k=0}^{i} P(\xi_{M1} = k, \xi_{M2} = i - k). \tag{A1}$$

Let ξ_{Mij} denote the number of allele j's errors at marker i in M, $i, j = 1, 2$. For example, ξ_{M12} is the number of allele 2's errors, *i.e.*, from true allele 2 to erroneous allele 1, at marker 1 in M. Then $\xi_{M1j} \sim \text{Binomial}(N_{1j}, e_j)$, and $\xi_{M2j} \sim \text{Binomial}(N_{2j}, \varepsilon_j)$, $j = 1, 2$, where N_{ij} is the number of allele j at marker i in M $(i, j = 1, 2)$. Note that $\xi_{Mij} = 0$ when $N_{ij} = 0$. Then,

$$\begin{aligned} P(\xi_{M1} = k) = P(\xi_{M11} + \xi_{M12} = k) &= \sum_{\substack{j=0 \\ j \le N_{11},\, k-j \le N_{12}}}^{k} P(\xi_{M11} = j, \xi_{M12} = k - j) \\ &= \sum_{j=0}^{k} \binom{N_{11}}{j} e_1^j (1 - e_1)^{N_{11}-j} \cdot \binom{N_{12}}{k-j} e_2^{k-j} (1 - e_2)^{N_{12}-(k-j)}. \end{aligned} \tag{A2}$$

(Here we define $\binom{N}{n} = 0$ if $N < n$). Similarly,

$$P(\xi_{M2} = i - k) = \sum_{\ell=0}^{i-k} \binom{N_{21}}{\ell} \varepsilon_1^{\ell} (1 - \varepsilon_1)^{N_{21}-\ell} \cdot \binom{N_{22}}{i-k-\ell} \varepsilon_2^{i-k-\ell} (1 - \varepsilon_2)^{N_{22}-(i-k-\ell)}. \tag{A3}$$

Substituting (A2) and (A3) in (A1), we obtain

$$P(i \text{ errors in } M) = \sum_{k=0}^{i} \left[\sum_{j=0}^{k} \binom{N_{11}}{j} e_1^j (1 - e_1)^{N_{11}-j} \cdot \binom{N_{12}}{k-j} e_2^{k-j} (1 - e_2)^{N_{12}-(k-j)} \right] \cdot \left[\sum_{\ell=0}^{i-k} \binom{N_{21}}{\ell} \varepsilon_1^{\ell} (1 - \varepsilon_1)^{N_{21}-\ell} \cdot \binom{N_{22}}{i-k-\ell} \varepsilon_2^{i-k-\ell} (1 - \varepsilon_2)^{N_{22}-(i-k-\ell)} \right]. \tag{A4}$$

In particular,

$$P(\text{no error in } M) = (1 - e_1)^{N_{11}} (1 - e_2)^{N_{12}} (1 - \varepsilon_1)^{N_{21}} (1 - \varepsilon_2)^{N_{22}}.$$

Thus,

$$\begin{aligned} P(\text{no error in trio}) &= \sum_{M \in S} P(\text{no error in trio} | M) \cdot P(M) \\ &= \sum_{M \in S} (1 - e_1)^{N_{11}} (1 - e_2)^{N_{12}} (1 - \varepsilon_1)^{N_{21}} (1 - \varepsilon_2)^{N_{22}} \cdot P(M). \end{aligned}$$

Noting that

$$\sum_{j=0}^{k} \binom{N_{11}}{j} \binom{N_{12}}{k-j} = \binom{N_{11}+N_{12}}{k} = \binom{6}{k}$$

and

$$\sum_{\ell=0}^{i-k} \binom{N_{21}}{\ell} \binom{N_{22}}{i-k-\ell} = \binom{N_{21}+N_{22}}{i-k} = \binom{6}{i-k},$$

we see that for the stochastic error model, (A4) reduces to

$$P(i \text{ errors in } M) = \binom{12}{i} e_1^i (1 - e_1)^{12-i}.$$

For the directed error model, we have

$$P(i \text{ errors in } M) = \sum_{k=0}^{i} \binom{N_{11}}{k} e_1^k (1 - e_1)^{N_{11}-k} \cdot \binom{N_{21}}{i-k} \varepsilon_1^{i-k} (1 - \varepsilon_1)^{N_{21}-(i-k)}.$$

APPENDIX B

In the following, we describe how to determine whether a haplotype pair consistent with child 1 is also consistent with both parents.

Let

$$\binom{h_1}{h_2}^0$$

denote a consistent haplotype pair of the first child, where $(\;)^0$ means phase information is known. Further, if h_F (h_M) is a haplotype of the father (the mother), let $\bar{h}_F$ ($\bar{h}_M$) denote the complementary haplotype in the sense that $\bar{h}_F$ ($\bar{h}_M$) consists of the remaining alleles of the father (the mother).

1. If h_1 is consistent with the father but not the mother, then h_2 has to be consistent with the mother unless there are genotyping errors. Thus, $\bar{h}_F$ can be determined by h_1 and the genotype of the father and $\bar{h}_M$ can be determined by h_2 and the genotype of the mother. Hence, possible haplotype pairs for the children in this family determined by such

$$\binom{h_1}{h_2}^0$$

are

$$\left\{ \binom{h_1}{h_2}^0, \binom{h_1}{\bar{h}_{2M}}^0, \binom{\bar{h}_{1F}}{h_2}^0, \binom{\bar{h}_{1F}}{\bar{h}_{2M}}^0 \right\}.$$

2. If h_1 is consistent with the mother but not the father, then h_2 has to be consistent with the father unless there are genotyping errors. Thus, possible haplotype pairs for the children in this family determined by

$$\binom{h_1}{h_2}^0$$

are

$$\left\{\binom{h_1}{h_2}^0, \binom{h_1}{\bar{h}_{2\mathrm{F}}}^0, \binom{\bar{h}_{1\mathrm{M}}}{h_2}^0, \binom{\bar{h}_{1\mathrm{M}}}{\bar{h}_{2\mathrm{F}}}^0\right\}.$$

3. If h_1 is consistent with both the father and the mother, then when h_2 is consistent with the father but not the mother, possible haplotype pairs for the children in this family determined by

$$\binom{h_1}{h_2}^0$$

are the same as those in possibility 2. When h_2 is consistent with the mother but not the father, possible haplotype pairs for the children in this family determined by

$$\binom{h_1}{h_2}^0$$

are the same as those in possibility 1. When h_2 is consistent with both the father and the mother, possible haplotype pairs for the children in this family determined by

$$\binom{h_1}{h_2}^0$$

are

$$\left\{\binom{h_1}{h_2}^0, \binom{h_1}{\bar{h}_{2\mathrm{F}}}^0, \binom{\bar{h}_{1\mathrm{M}}}{h_2}^0, \binom{\bar{h}_{1\mathrm{M}}}{\bar{h}_{2\mathrm{F}}}^0\right\}$$

and

$$\left\{\binom{h_1}{h_2}^0, \binom{h_1}{\bar{h}_{2\mathrm{M}}}^0, \binom{\bar{h}_{1\mathrm{F}}}{h_2}^0, \binom{\bar{h}_{1\mathrm{F}}}{\bar{h}_{2\mathrm{M}}}^0\right\}.$$

APPENDIX C

As an example, we calculate $P(T = M|O = M)$, where

$$M = \left(\binom{1}{1}\binom{1}{2}\binom{1}{2}\right).$$

It can be shown that

$$\begin{aligned}P(O = M|T = M) &= P\left(O = \left(\binom{1}{1}\binom{1}{2}\binom{1}{2}\right)\middle| T = \left(\binom{1}{1}\binom{1}{2}\binom{1}{2}\right)\right)\\ &= (1 - e_1)^4(1 - e_2)^2 + 4e_1(1 - e_1)^3\, e_2(1 - e_2) + 3e_1^2(1 - e_1)^2 e_2^2.\end{aligned}$$

Similarly, we can get the other conditional probability $P(O = M|T = M')(M' \neq M)$. Substituting these formulas into (4) and using the values of $P(T = M)$, we obtain

$$\begin{aligned}P(O = M) = 8e_1^2(1 - e_1)^4 &\\ \cdot p^4 + [&2e_1(1 - e_1)^4(1 - e_2) + 6e_1^2(1 - e_1)^3\, e_2 + (1 - e_1)^4(1 - e_2)^2\\ &+ 4e_1(1 - e_1)^3 e_2(1 - e_2) + 3e_1^2(1 - e_1)^2 e_2^2]\\ \cdot 2p^3q + [&(1 - e_1)^3 e_2(1 - e_2)^2 + 2e_1(1 - e_1)^2 e_2^2(1 - e_2)\\ &+ e_1^2(1 - e_1)e_2^3 + e_1(1 - e_1)^3 e_2(1 - e_2) + e_1^2(1 - e_1)^2 e_2^2\\ &+ (1 - e_1)^3 e_2(1 - e_2)^2 + 2e_1(1 - e_1)^2 e_2^2(1 - e_2) + e_1^2(1 - e_1)e_2^3\\ &+ (1 - e_1)^2 e_2^2(1 - e_2)^2 + e_1(1 - e_1)e_2^3(1 - e_2)]\\ \cdot 4p^2q^2 + [&6(1 - e_1)e_2^3(1 - e_2)^2 + 2e_1e_2^4\,(1 - e_2) + 3(1 - e_1)^2 e_2^2(1 - e_2)^2\end{aligned}$$

$$+ 4e_1(1-e_1)e_2^3(1-e_2) + e_1^2e_2^4]$$

$$\cdot 2pq^3 + 8e_2^4(1-e_2)^2 \cdot q^4,$$

where p is the population frequency of allele 1, and $q = 1 - p$. Thus, we have

$$P(T = M|O = M) = [(1-e_1)^4(1-e_2)^2 + 4e_1(1-e_1)^3e_2(1-e_2) + 3e_1^2(1-e_1)^2e_2^2] \cdot 2p^3q/P(O = M).$$

APPENDIX D

Theorem. If there is no phase information and each marker in a set of tightly linked markers is Mendelian consistent, the trio is Mendelian consistent across these markers.

Proof. Let $\mathcal{P}$, $\mathcal{M}$, and $\mathcal{C}$ denote the genotypes for the two parents and the child across a set of k markers, where

$$\mathcal{P} = \begin{pmatrix} a_{11} & \cdots & a_{k1} \\ a_{12} & \cdots & a_{k2} \end{pmatrix}$$

$$\mathcal{M} = \begin{pmatrix} b_{11} & \cdots & b_{k1} \\ b_{12} & \cdots & b_{k2} \end{pmatrix}$$

$$\mathcal{C} = \begin{pmatrix} c_{11} & \cdots & c_{k1} \\ c_{12} & \cdots & c_{k2} \end{pmatrix}$$

Let c_i^{p} be the allele consistent with one of the two alleles a_{i1} and a_{i2} in the father and c_i^{m} be the allele consistent with one of the two alleles b_{i1} and b_{i2} in the mother. Then we would infer that one of the haplotypes in the father is $(c_1^{\mathrm{p}}c_2^{\mathrm{p}} \ldots c_k^{\mathrm{p}})$ and one of the haplotypes in the mother is $(c_1^{\mathrm{m}}c_2^{\mathrm{m}} \ldots c_k^{\mathrm{m}})$. It is easy to see that such inference would imply Mendelian consistency for this family trio without recombinations among these markers.