Published in final edited form as:

Genet Epidemiol. 2005 November ; 29(3): 204–214. doi:10.1002/gepi.20086.

Detection of Genotyping Errors and Pseudo-SNPs Via Deviations From Hardy-Weinberg Equilibrium

Suzanne M. Leal*

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

Abstract

Genotype error can greatly reduce the power of a genetic study. For family data, genotype error can be assessed by examining marker data for non-Mendelian inconsistencies, closely linked markers for double recombination events, and consistency of duplicate genotypes. For case-control data, duplicate samples are genotyped, and controls are tested for deviations from Hardy-Weinberg equilibrium (HWE). Duplicate samples can provide accurate estimates of genotyping error rates, unless systematic genotyping errors have occurred. Although genotyping errors can cause deviations from HWE, these deviations are usually small, and the power to detect them is low except for high rates of genotyping error and/or large sample sizes. An additional problem is that even when deviations from HWE are detected for marker loci, without additional experimentation it is not possible to unequivocally implicate genotyping error as the cause. The power and sample sizes necessary to detect deviations from HWE for single-nucleotide polymorphism (SNP) data are examined for a variety of genotyping error and pseudo-SNP models. For the majority of genotyping models examined, the power is poor to detect deviations from HWE. For example, for 1,000 controls, if an allele with a frequency of 0.1 fails to amplify for 28% of the heterozygous genotypes producing a sample error rate of 0.05, the power is 0.51 to detect a deviation from HWE at an alpha level of 0.05. On the other hand, the detection of deviations from HWE for pseudo-SNPs (paralogous and ectopic sequence variants) for the majority of models examined produces a power of >0.8 for sample sizes as small as 50 individuals.

Keywords

case-control studies; population-based studies; type I and II errors

INTRODUCTION

In the past few years, there has been a renaissance in the use of association studies to map complex traits. Association studies are carried out using either family- or population-based data. For both of these study designs, genotyping errors can be detrimental. For family-based data (i.e., trio data), genotyping errors can increase both type I and II errors [Gordon et al., 1999b, 2001]. For population-based studies (i.e., case-control), genotyping errors can

*Correspondence to: Dr. Suzanne M. Leal, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Alkek Building N1619.01, Houston, TX 77030. sleal@bcm.tmc.edu.

The supplemental materials described in this article can be found at <http://www.interscience.wiley.com/jpages/0741-0395/suppmat>

increase type II errors and thereby decrease power [Boss, 1954; Gordon and Ott, 2001; Gordon et al., 2002, 2003]. Additionally, genotyping errors can bias linkage disequilibrium (LD) measurements [Akey et al., 2001]. It is important to be able to assess genotyping error rates for individual single-nucleotide polymorphism (SNP) markers and sets of markers, so that low-performance markers can be removed from the analysis and genotyping platforms can be optimized. For trio data, genotyping errors can sometimes be detected by observation of Mendelian inconsistencies; however, the ability to detect Mendelian inconsistencies decreases with increasing marker heterozygosity [Gordon et al., 1999a; Geller and Ziegler, 2002]. For population- and family-based data, duplicate individuals can be genotyped to determine the rate of genotyping errors. However, if a systematic genotyping error has occurred, it will not be detected, and genotyping error rates will be underestimated. Individuals from population-based studies, or controls from case-control studies, can be analyzed to determine whether their genotype data are in Hardy-Weinberg equilibrium (HWE) to give clues as to whether or not genotyping errors at particular marker loci may have occurred [Tiret and Cambien, 1995; Xu et al., 2002; Hosking et al., 2004]. It should be noted that deviations from HWE are not necessarily due to genotyping error and may be due to chance, other genetic factors such as a protective allele, population admixture/stratification, inbreeding, or deletions [Cockerham, 1967, 1969; Crow and Kimura, 1970; Hartl and Clark, 1989; Deng et al., 2001; Weir et al., 2004]. Although testing for deviations of HWE is commonly used [Xu et al., 2002] to detect potential genotyping errors, the power of this technique has not been evaluated. This article examines how the rate and model of genotyping error affect deviations from HWE. What is the power to detect deviations for a given sample size? What sample sizes are needed to detect these deviations with a power of 0.8 and an α level of 0.05? In addition, the power to detect pseudo-SNPs (i.e., ectopic sequence variants (ESVs) and paralogous sequence variants (PSVs) [Estivill et al., 2002; Cheung et al., 2003] by testing for deviations from HWE is examined.

METHODS

Each SNP has two alleles denoted by A_1 and A_2 , where the A_1 has a frequency of $p = 0.1, 0.3, 0.5, 0.7, \text{ or } 0.9$, and the A_2 allele has a frequency of $1-p = q$. For each SNP marker, the proportion (P_{11}, P_{12} , and P_{22}) for each genotype G_{11}, G_{12} , and G_{22} is calculated under HWE. The proportion of G_{11}, G_{12} , and G_{22} genotypes equals $p^2, 2pq$, and q^2 , respectively. For each of the models described below, errors are introduced at a rate of $e = 0.05, 0.1, \dots, 0.95, 1.0$. The proportions of G_{11}, G_{12} , and G_{22} genotypes after errors have been introduced are denoted as P'_{11}, P'_{12} , and P'_{22} . The allele frequency for A_1 after errors have been introduced is

$$p' = \frac{2P'_{11} + P'_{12}}{2}. \quad (1)$$

The disequilibrium coefficient D [Weir, 1996], which equals 0 under complete HWE after errors have been introduced, is

$$D' = P'_{11} - p'^2. \quad (2)$$

For each error model, the sample size necessary to detect a deviation from HWE for the χ^2 test is determined by

$$N = v \frac{p'^2(1-p')^2}{D'^2}. \quad (3)$$

For all examples given, the sample size is calculated for ν power $(1-\beta) = 0.8$ and an α level of 0.05. For this situation, the noncentrality parameter ν is set equal to 7.849 [Haynam et al., 1970], for which there is a 0.80 probability that the χ^2 critical value will exceed 3.84 (the value necessary to reject the null hypothesis at an α level of 0.05) when the hypothesis of HWE is false, and a 0.05 probability that χ^2 will exceed 3.84 when the hypothesis of HWE is true. Additionally, for each error model for various allele frequencies p , error rates e , and sample sizes $n = 1,000, 500, 250, 100$, and 50, the noncentrality parameter ν for the χ^2 statistic is determined by

$$\nu = n \frac{D'^2}{p'^2(1-p')^2}. \quad (4)$$

Based on the value of ν and an α level of 0.05, the R statistical package [R Development Core Team, 2004] was used to estimate the power for each model.

For error models 1–4, genotyping errors are introduced into the data set. For model 5, ESVs are generated. For models 1–5, P'_{11} , P'_{12} , p' , D' , N , and ν are derived for each error model. First, for each error model, P'_{11} and P'_{12} are formulated in terms of p and e . These equations are then substituted into formula 1 to derive the allele frequency after errors have been introduced. Additionally, for each error model, by substituting the formulae for P'_{11} and p' in Equation 2 and p' and D' in Equations 3 and 4, the disequilibrium coefficient, sample size, and noncentrality parameter are formulated in terms of p and e .

ERROR MODELS

Model 1 (errors introduced randomly) The probability that a particular allele is changed to the other allele ($A_1 \rightarrow A_2$ or $A_2 \rightarrow A_1$) is $p \times e$ for the A_1 allele, and $q \times e$ for the A_2 allele. For model 1, $P'_{11} = p^2 + pe(1 - 2p)$, $P'_{12} = 2pq - 4pe(1 - p) + e$, $p' = p + \frac{1}{2}e(1 - 2p)$,

$$D' = -\frac{1}{4}e^2(1 - 2p)^2,$$

$$N = v \frac{(e(2p - 1) - 2p)^2(2 - 2p + e(2p - 1))^2}{e^4(1 - 2p)^4},$$

and

$$v = n \frac{(e - 2p)^4}{(-2e(1 - 2p)^2 + e^2(1 - 2p)^2 + 4p(p - 1))^2}.$$

The sample error rate (SER) = e .

Model 2 (errors introduced to only one of the two alleles) Errors are introduced to the A_1 allele of the G_{12} genotype and modify it to the G_{22} at a frequency of $2pq \times e$, and likewise the G_{11} genotype is altered to the G_{12} genotype at a frequency of $p^2 \times e$. For model 2,

$$\begin{aligned} P'_{11} &= p^2(1 - e), \quad P'_{12} = 2pq(1 - e) + p^2e, \\ p' &= p - pe(1 - \frac{1}{2}p), \\ D' &= -1 / 4ep^2(e(p - 2)^2 + 4(p - 1)), \\ N &= v \frac{(2 + e(p - 2))^2(2 + 2p(e - 1) - ep^2)^2}{e^2(e(p - 2)^2 + 4(p - 1))^2 p^2}, \end{aligned}$$

and

$$v = n \frac{e^2 p^2 (e(p - 2)^2 + 4(p - 1))^2}{(4(p - 1) + e^2 p(p - 2)^2 + 2e(2 - 5p + 2p^2))^2}.$$

The overall SER = $(p^2 + 2pq) \times e$.

Model 3 (genotyping errors convert heterozygous genotypes to homozygous genotypes)

It is random to which allele an error is introduced, and $G_{12} \rightarrow G_{22}$ and $G_{12} \rightarrow G_{11}$. For model 3,

$$\begin{aligned} P'_{11} &= p^2 + pe(1 - p), \quad P'_{12} = 2pq(1 - e), \quad p' = p, \\ D' &= ep(1 - p), \quad N = \frac{v}{e^2}, \quad \text{and } v = ne^2. \end{aligned}$$

For model 3b, the error is always introduced to the A_1 allele, and only genotype $G_{12} \rightarrow G_{22}$. For model 3b,

$$\begin{aligned} P'_{11} &= p^2, \quad P'_{12} = 2pq(1 - e), \quad p' = p - pe(1 - p), \\ D' &= -ep^2(p - 1)(2 + e(p - 1)), \\ N &= v \frac{(1 + e(p - 1))^2(1 + ep)^2}{e^2(2 + e(p - 1))^2 p^2}, \end{aligned}$$

and

$$v = n \frac{e^2 p^2 (2 + e(p-1))^2}{(1 + e^2 p(p-1) + e(2p-1))^2}.$$

For both models, the frequency of introduction of errors and the $SER = 2pq \times e$.

Model 4 (genotyping errors convert homozygous genotypes to heterozygous genotypes)

The frequency at which an error is introduced to the G_{11} genotype is $p^2 \times e$, and the frequency at which an error is introduced to the G_{22} genotype is $q^2 \times e$. For model 4,

$$P'_{11} = p^2(1 - e), P'_{12} = 2pq - 2pe(1 - p) + e,$$

$$p' = p + \frac{1}{2}e(1 - 2p),$$

$$D' = -\frac{1}{4}e(e(1 - 2p)^2 - 4p(p - 1)),$$

$$N = v \frac{(e(2p-1) - 2p)^2 (2 - 2p + e(2p-1))^2}{e^2 (e(1 - 2p)^2 - 4(p-1)p)^2},$$

and

$$v = n \frac{e^2 (e(2p-1)^2 - 4p(p-1))^2}{(-2e(2p-1)^2 + e^2 (2p-1)^2 + 4p(1-p))^2}.$$

The SER is $(q^2 + p^2) \times e$.

For model 4b, errors are only introduced to the G_{11} genotype, and $P'_{11} = p^2(1 - e)$,

$$P'_{12} = 2pq + p^2e, p' = p - \frac{1}{2}p^2e, D' = ep^2(p - \frac{1}{4}ep^2 - 1),$$

$$N = v \frac{(ep - 2)^2 (2 - 2p + ep^2)^2}{e^2 p^2 (4 - 4p + ep^2)^2},$$

and

$$v = n \frac{e^2 p^2 (4 - 4p + ep^2)^2}{(2p(2 + e) + e^2 p^3 - 4ep^2 - 4)^2}.$$

The SER for model 4b is $p^2 \times e$.

Model 5 (ESV) Two types of ESVs are examined. For model 5, one site is monomorphic for the A_1 allele, while the second site is polymorphic, with allele frequencies $p = 0.1, 0.3, 0.5$,

0.7, and 0.9 for the A_1 allele. The genotype frequencies at the second site are in complete HWE. For model 5,

$$P'_{11} = p^2, P'_{12} = 2pq + q^2, p' = \frac{1}{2}(1 + p^2),$$

$$D' = \frac{1}{4}(p^2 - 1)^2,$$

$$N = v \frac{(1 + p^2)^2}{(p^2 - 1)^2}, \text{ and } v = n \frac{(p^2 - 1)^2}{(1 + p^2)^2}.$$

For model 5b, both locus sites are polymorphic ($p = 0.05, 0.1, 0.3, 0.5, 0.7$, and 0.9) for the same alleles and in complete HWE. The allele frequency for the A_1 allele at the first locus is denoted by p_a , and the allele frequency for the A_2 allele at the first locus is denoted by q_a . Likewise, at the second locus, the allele frequency for the B_1 allele is denoted by p_b , and the allele frequency for the B_2 allele is denoted by q_b . For model 5b,

$$P'_{11} = p_a^2 p_b^2, P'_{12} = 1 - p_a^2 p_b^2 - q_a^2 q_b^2,$$

$$p' = \frac{1}{2} + \frac{1}{2}(p_a^2 p_b^2 - q_a^2 q_b^2),$$

$$D' = p_a^2 p_b^2 - \frac{1}{4}(1 + p_a^2 p_b^2 - q_a^2 q_b^2)^2,$$

$$N = v \frac{(p_a^4 p_b^4 - 2p_a^2 p_b^2 q_a^2 q_b^2 + q_a^4 q_b^4 - 1)^2}{(p_a^4 p_b^4 + (q_a^2 q_b^2 - 1)^2 - 2p_a^2 p_b^2 (1 + q_a^2 q_b^2))}.$$

and

$$v = n \frac{(p_a^2 p_b^2 + (q_a^2 q_b^2 - 1)^2 - 2p_a^2 p_b^2 (1 + q_a^2 q_b^2))^2}{(p_a^4 p_b^4 + q_a^4 q_b^4 - 2p_a^2 p_b^2 q_a^2 q_b^2 - 1)^2}.$$

All possible combinations of allele frequencies at both sites were examined.

RESULTS

For a given power and alpha level, two parameters, D' and p' , play a role on the sample size needed to reject the null hypothesis of HWE. Likewise, the parameters D' and p' influence the power to reject the null hypothesis of HWE for a set sample size and α level. For a set sample size and fixed p' , the power to detect deviations from HWE increases, the more divergent D' is from 0. In addition, for a set p' , power, and α level, the sample size will decrease with increasing divergents of D' . How divergent D' is from 0 is dependent on to what extent the error biases P'_{11} and p' (see Formula 2). For biallelic loci, the minimum value for D' is -0.25 when all genotypes are heterozygous, and the maximum value for D'

is 0.25 when all individuals are homozygous, with 50% of the genotypes homozygous for one allele and 50% of the genotypes homozygous for the other allele. For a set D' , sample size, and α level, the power increases as p' or $(1-p')$ approach 0, and the lowest power is obtained when $p' = 0.5$.

For the majority of genotyping error models examined (models 1–4b), even with sample sizes of 1,000 control individuals (Tables I–III), the power to detect genotyping errors through deviations in HWE is low even for high error rates (i.e., 15%). In most cases, but not always, deviations from HWE, D' , and the power to detect them increase with increasing genotyping error rates. For specific error rates and sample sizes, the effect of allele frequency on the extent of departure of D' from 0 and the power to detect deviations from HWE are model-dependent. On the other hand, for ESV models examined for most situations, even small sample sizes of 50 individuals or less are sufficient to detect deviations from HWE with a power of 0.8 and α level of 0.05.

For a random error (model 1), when $p = 0.5$, the introduction of errors does not cause a deviation from HWE ($D' = 0$) for all error rates (Supplemental Fig. I). The power in this situation is 0.05, the same that would be expected by chance for an α level of 0.05, and the sample size is undefined (Table I). When $p \neq 0.5$, random errors always lead to $D' < 0$ (Supplemental Fig. I). For $p < 0.5$, $p' > p$, with the difference between p' and p increasing with decreasing p and increasing e (Supplemental Fig. II). For allele frequencies of $p = 0.1$ and 0.3 and error rates of 0.15, more than 10,000 individuals were necessary to obtain a power of 0.8 at an α level of 0.05 (Table I). For allele frequencies that were examined for the same error level, the highest power to detect a deviation from HWE was obtained when the A_1 allele $p = 0.1$. For $p = 0.5$, the power to detect deviations from HWE increased with increasing error rates (Table I; Supplemental Fig. III). For 1,000 controls for $p = 0.1$ and 0.3, it was not until $e = 0.35$ and 0.75, respectively, that the power was >0.8 to detect a deviation of HWE at an α level of 0.05 (Supplemental Fig. III).

For model 2, errors are introduced to the A_1 allele, causing genotypes $G_{11} \rightarrow G_{12}$ and $G_{12} \rightarrow G_{22}$. For model 2, the allele frequency plays an important role in the amount of deviation from HWE. For example, when $p = 0.9$, for error rates 0.30, D' ranges from 0.0–0.007 and then becomes negative for error rates 0.35, dropping to -0.164 when $e = 1.0$. When $p = 0.1$, the estimates of D' are only slightly greater than 0.0, regardless of the error rates. When $p = 0.5$, the highest estimate of $D' = 0.028$ occurs at $e = 0.45$ and drops below 0 for $e = 0.9$ (Supplemental Fig. IV). When errors are introduced, $p' < p$. The difference between p' and p increases with increasing value of p and error rates (Supplemental Fig. V). For all p values examined ($p = 0.1, 0.3, 0.5, 0.7$, and 0.9), the power increases with increasing error rates, but then reaches a threshold that is dependent on p' , where the power drops and then increases again except for $p = 0.1$, where the power does not increase for a second time (Supplemental Fig. VI). For model 2, a higher error rate does not necessarily lead to increased power to detect a deviation from HWE. For error rates of 0.2, for the p values examined, $p = 0.7$ leads to the highest power to detect a deviation from HWE; for example, for an error rate of 0.2, for a sample of 1,000 individuals, power = 0.8 for an α level of 0.05. For the same sample size and error rate, the power drops to 0.71 and 0.25 for $p = 0.5$ and 0.9, respectively (Table II). For $p = 0.9$, the power increases from 0.17 to 0.31 for

error rates of 0.05–0.15, but then drops to 0.25 for an error rate of 0.2. The power continues to decline, reaching 0.06 at $e = 0.35$, and then starts increasing until it reaches 1.0 at $e = 0.7$ (Table II; Supplemental Fig. VI). For model 2, when $SER = 0.05$ for 1,000 individuals, the power is < 0.2 , regardless of the allele frequency (Table III).

For model 3, errors are only introduced to the G_{12} genotype. When an error is introduced to the heterozygous genotype, each allele has an equal probability of being miscalled, resulting in either a G_{11} or G_{22} genotype. For model 3, the introduction of errors does not bias p' , $p' = p$. With increasing error rates and as p goes to 0.5, D' positively departs from 0 until it reaches its maximum of 0.25, when $e = 1.0$ (Supplemental Fig. VII). For model 3, the power to detect a deviation from HWE is only dependent on error rates and not allele frequencies. For an error rate of 0.1, the power to detect a deviation from HWE at an α level of 0.05 is 0.89, 0.61, and 0.35 for 1,000, 500, and 250 individuals, respectively (Table I; Supplemental Fig. VIII).

For model 3b, errors are only introduced to the G_{12} genotype an $G_{12} \rightarrow G_{22}$. For models 3 and 3b, for equivalent values of p and e , the SER is the same. With increasing error rate, D' positively departs from 0 and reaches its maximum at $e = 1.0$. It should be stated that the amount of deviation of D' from 0 is highly dependent on p , and for the maximum error rate of $e = 1.0$, $D' = 0.01, 0.08, 0.19, 0.25$, and 0.15 for $p = 0.1, 0.3, 0.5, 0.7$, and 0.9 , respectively (Supplemental Fig. IX). For model 3b, the introduction of errors leads to $p' < p$ (Supplemental Fig. X). Higher error rates lead to a greater deviation between p' and p . For a fixed error rate, the amount of deviation of p' from p is symmetrical for p and $(1-p)$, and the greatest deviation between p' and p is reached when p and $(1-p)$ are in the vicinity of 0.5. For model 3b, for a constant error and sample size, the power to detect a deviation from HWE increases with increasing p (Supplemental Fig. XI). For example, for an α level of 0.05 when $e = 0.05$ and $p = 0.1$ for 1,000 individuals, the power is 0.06, and increases to 0.78 for $p = 0.9$. When the sample size is reduced to 500 for $p = 0.9$, the power drops to 0.49 (Table II; Supplemental Fig. XI). For a power of 0.8 at an α level of 0.05, when the SER was kept constant at 0.05, the sample size necessary to detect a deviation from HWE reduced with increasing allele frequency from $N = 1,974$ for $p = 0.1$ down to 48 individuals for $p = 0.9$. For both $p = 0.1$ and $p = 0.9$ for $SER = 0.05$, errors were introduced to the G_{12} genotype at frequency $e = 0.278$ (Table III).

For model 4, errors are introduced to the G_{11} or G_{12} genotypes, and $G_{11} \rightarrow G_{12}$ and $G_{22} \rightarrow G_{12}$. For both error model 4 and model 1,

$$p' = p + \frac{1}{2}e(1 - 2p),$$

and for $p < 0.5$, $p' > p$ (Supplementary Fig. II). With increasing error rates, D' goes to its minimum value of -0.25 at $e = 1$ for all values of p . With the exception of $e = 1$ for the same error rates, $p = 0.5$ produces smaller D' values than for $p = 0.3$ and 0.1 (Supplemental Fig. XII). The power to detect a deviation from HWE for 1,000 individuals at an α level of 0.05 is poor when $e = 0.05$ for all allele frequencies; $(1-\beta) = 0.31, 0.34$, and 0.35 for $p = 0.1, 0.3$, and 0.5 , respectively. When the error rate is increased to 0.15, the power is > 0.8 for all allele

frequencies examined, even when the sample size is reduced to 500 individuals (Table I; Supplemental Fig. XIII).

For model 4b, errors are introduced to the G_{11} and $G_{11} \rightarrow G_{12}$. For all allele frequencies, the introduction of errors causes a deviation from HWE with $D' < 0$, with the greatest deviations occurring when errors are introduced to the more frequent allele (e.g., $p = 0.7$ or 0.9) (Supplemental Fig. XIV). The introduction of genotyping error always causes $p' < p$, with increasing deviations with increasing values of p and e (Supplemental Fig. XV). The power at an α level of 0.05 to detect deviations from HWE increases with increasing allele frequency; for $p = 0.1$ and $e = 0.1$, the power to detect a deviation from HWE for 1,000 controls is 0.06, and increases to 0.72 when $p = 0.9$ (Table II; Supplemental Fig. XVI).

For model 5 (ESV), i.e., pseudo-SNPs, in almost all situations the power to detect deviations from HWE is high. In the case where one site is monomorphic and the other site is polymorphic, with an allele frequency of 0.1, 0.3, or 0.5 for the same allele at the monomorphic site, the power is >0.9 to detect a deviation from HWE for a sample size of 50 individuals. When $p = 0.7$ and 0.9 at the polymorphic site, the power is reduced, and sample sizes of 67 and 712, respectively, are necessary to obtain a power of 0.8 for $\alpha = 0.05$ to reject the null hypothesis of HWE.

For model 5, $p' < p$ and $D' < 0$ (Table IV). When both sites are polymorphic (model 5b), the power is also excellent (>0.8) at an α level of 0.05 to detect deviations from HWE in 50 control individuals, as long as both sites do not have a low frequency for the same allele ($p < 0.1$). When these conditions are not met, much larger sample sizes are necessary to obtain a power of 0.8, e.g., if the frequency at both sites for the A_1 allele is 0.05, a sample size of 752 individuals is necessary to reject the null hypothesis of HWE (Table V).

DISCUSSION

Although it is simple to carry out χ^2 or likelihood ratio tests [Weir, 1996] to evaluate data sets of controls or unrelated individuals for deviations from HWE, this should not be the only method used to evaluate whether or not a genotyping error is present. Even though testing marker genotypes for deviations from HWE can give clues to a potential genotyping error in many instances, low power will result in non-detection of deviations from HWE due to genotyping errors. If an α level of 0.05 is being used as the criterion to reject HWE, then data sets of controls with $> 5\%$ of the markers rejecting HWE could indicate nontrivial rates of genotyping error [Hosking et al., 2004]. On the other hand, higher rates of rejecting HWE could also indicate a population with stratification/admixture or other population-based causes of violation of HWE. Although deviations from HWE can be due to genotyping error, there are many other plausible explanations. Therefore, other methods must be used to determine if genotyping errors have occurred. These methods include genotyping duplicate individuals for the same marker, genotyping the same individuals and marker loci using different genotyping platforms, or genotyping the marker loci in family data. In pedigree data it is possible to evaluate genotyping errors by detecting Mendelian inconsistencies. For trio data, the detection of errors through the observation of Mendelian inconsistencies is low. Only 0.25–0.30 of errors can be detected through the observation of Mendelian

inconsistencies, if it is assumed that the type of genotyping error which occurs is random [Gordon et al., 1999a]. A more realistic model is genotyping error due to poor amplification of one of the alleles, which leads to heterozygous genotypes being read as homozygous genotypes. For this error model, detection rates as high as 0.61 can be observed for trio data [Geller and Ziegler, 2002]. When pedigree data are genotyped with two or more offspring, the detection rates increase over those for which only trios are examined [Gordon et al., 2000; Douglas et al., 2002]. For nuclear families with a random genotyping error, error detection rates for SNP markers ranged from 0.13–0.75 [Douglas et al., 2002]. In addition, for family genotype data, it is possible to examine tightly linked markers for double recombination events which could be indicators of genotyping error [Lincoln and Lander, 1992; Brzustowicz et al., 1993; Ehm et al., 1996; Stringham and Boehnke, 1996; Douglas et al., 2000].

Genotyping error rates that are platform- and/or site-specific can be estimated by comparing generated genotypes to publicly available genotypes, such as those from the HapMap project [Matsuzaki et al., 2004]. The HapMap project has made publicly available phase I and II quality-control (QC) data. The phase II QC data consist of 1,500 validated SNPs that were chosen from the HapMap project and subsequently regenotyped on the 30 Centre d'Etude du Polymorphisme Humaine (CEPH) trios, using a number of technologies including fluorescent polarization [Hsu et al., 2001], BeadArray (Illumina) [Oliphant et al., 2002], Molecular Inversion Probe (Parallele) [Hardenbol et al., 2003], and Invader (Third Wave) [Lyamichev and Neri, 2003]. In addition, the 1,500 markers were regenotyped on the 30 CEPH trios using some of the same technologies (e.g., BeadArray, Illumina) at multiple sites. Genotyping error rates for a specific platform and/or site can be assessed by regenotyping a subset of phase II QC markers on the same CEPH samples which were used for the HapMap project, and comparing them to these publicly available genotypes.

In order to assess genotypes of new SNP marker loci for errors so that poorly performing SNPs can be removed, markers can be genotyped on a second platform, using the same samples or subset of samples. However, when comparing genotypes for the same individuals between two different platforms, it can be difficult to assess for which platform a genotyping error has occurred when an inconsistency is observed. In addition, there can be systematic errors due to primer design. Using the gold standard of bidirectional sequencing can reduce these problems.

Signal intensity, background noise and/or clustering properties all play a role in the ability to correctly assign genotypes and the types of errors which occur. Although all the genotyping error models that were examined could potentially occur, the model that is most likely to occur is one where a heterozygous genotype is scored as a homozygous genotype (model 3b). This type of genotyping error will occur when there is poor amplification of one of the alleles, and heterozygous genotypes are read as homozygous. For this model, for markers where the allele with a low allele frequency (e.g., 0.1) have amplification problems, large sample sizes are necessary to detect deviations from HWE, even with allele dropout rates as high as 0.15 (Table II). The power to detect deviations from HWE is greatly improved when the allele with poor amplification has a frequency of 0.7–0.9. For an allele with a frequency

of 0.9 that has a 10% dropout rate, 575 controls are necessary to detect a deviation from HWE with a power of 0.8 and α level of 0.05.

Rates of genotyping error can vary from sample-to-sample and batch-to-batch due to quality of DNA, instability of genotyping reagents, and operator error. Within a study, variation in genotyping error rates may not be random if the order of sample-processing is done in a systematic fashion, e.g., controls are genotyped mainly in one batch and cases mainly in another batch. For this example, a genotyping error could lead to an increase in type I and II error.

If a sample is contaminated with DNA from one or more other samples, genotyping errors can arise. In this situation homozygous genotypes would be called as heterozygous genotypes (Model 4) and an excess of heterozygous genotypes would be observed. If it is suspected that DNA contamination has taken place, the samples should be genotyped with several microsatellite markers. The observation of three or more alleles for the microsatellite genotypes indicates that DNA contamination has occurred.

Pseudo-SNPs can occur when a primer amplifies two different sites in the genome, which can occur due to duplications within the genome. Duplications and pseudogenes within the genome are not rare [Bailey and Eichler, 2003]. The frequency of pseudo-SNPs is not only dependent on the frequency of duplications within the genome, but also on primers designed for SNP marker loci. Although using bioinformatics tools such as BLAST [Altschul et al., 1997] is greatly beneficial in identifying primers that could potentially amplify more than one region of the genome, this method will not totally alleviate the problem. The majority of pseudo-SNPs are PSVs where two sites are monomorphic for a different base. For an ESV, one or more sites are polymorphic. If an ESV produces a SNP with more than two alleles, the SNP would be removed, and the primers would not be used for future genotyping, since it is unusual to utilize SNPs which are not diallelic. The occurrence of ESV pseudo-SNPs is most probably rare.

The power to detect deviations from HWE due to ESVs in most situations is high, even with small sample sizes (e.g., 50 controls). The ESV models examined had either one or two polymorphic sites. For PSVs where both sites are monomorphic but for different alleles, all individuals are heterozygous if no genotyping errors are present, which leads to $D' = -0.25$. In this situation, when carrying out a χ^2 test to detect a deviation from HWE, 8 individuals will provide 80% power at an α level of 0.05. A red flag should be raised if an SNP presents with few or no homozygous genotypes observed.

Under HWE, if the frequency of one allele is more than twice that of the other, the proportion of G_{12} genotypes is intermediate between the frequency of the G_{11} and G_{22} genotypes. For a population in equilibrium, the relationship between heterozygous and homozygous genotypes is $4(P_{11} \times P_{22}) = P_{12}^2$. When there is an excess of heterozygous genotypes, D' will be negative, while for an excess of homozygous genotypes, D' will be positive. For the models that were examined, the introduction of genotyping errors led to an excess of heterozygous genotypes for models 1 (except when $p = 0.5$, where $D' = 0$), 4, and 4b, while models 3 and 3b led to an excess of homozygous genotypes. Model 2 led to an

excess of homozygous or heterozygous genotypes, depending on the allele frequency and error rates (Supplemental Fig. IV). All pseudo-SNP models led to negative D' values. The observation of a statistically significant deviation from HWE, with an excess of heterozygous genotypes, would suggest that a pseudo-SNP is present, there is DNA contamination, or a genotyping error has occurred at this locus. Another explanation is that the SNP has a heterozygous advantage or is in LD with a variant that has a heterozygous advantage or occurred by chance. For those SNP markers where HWE has been rejected and there is an excess of homozygous genotypes, these observations could have occurred due to chance, population history (e.g., stratification/admixture, inbreeding), or genotyping error. It is of interest to note that genotyping errors due to allele dropout (model 3b) led to an excess of homozygous genotypes. Multiple markers in LD which are not HWE would suggest that a genetic factor was involved in the violation of HWE, and not genotyping error. For those SNPs for which there is a significant deviation from HWE, the call rate and cluster quality score should be examined. Low passing rates for either of these two measures are suggestive that deviations from HWE are due to genotyping error. Additionally, for those SNP loci for which a significant deviation from HWE is due to an excess of heterozygous genotypes, BLAST scores for the primers for these marker loci can be examined in order to elucidate if these are potentially pseudo-SNPs. Those SNP loci whose primers have high similarity scores to another sequence may be indicative that the SNP is actually a pseudo-SNP.

In case data, tests for deviations from HWE should not be used to try to detect genotyping errors. Cases are ascertained due to their disease status, and therefore disease susceptibility genotypes/alleles should be present at high rates in the case sample. Thus, deviations from HWE are greatest at trait susceptibility loci, but can also be detected for benign polymorphisms that are in LD with the susceptibility locus [Feder et al., 1996; Nielsen et al., 1999]. HW disequilibrium decreases with decreasing LD between the marker and susceptibility loci. Additionally, de novo mutations and deletions in cases can lead to deviations from HWE [Dhami et al., 2005]. Therefore, detection of HW disequilibrium at SNP loci in the case sample could be an indicator that a disease/susceptibility gene is within the region.

Tests for HWE can be used to test for the potential presence of genotyping errors in any non-ascertained sample. For the majority of models tested, the power to detect genotyping errors was low even for large samples, unless error rates were high. Therefore, the inability to detect a deviation from HWE, especially for small sample sizes, does not rule out that a nontrivial genotyping error may be present. Only pseudo-SNP models had adequate power, even with small sample sizes, to reveal deviations from HWE for most situations. Even when deviations from HWE are detected for samples of SNP loci, this observation may not be due to genotyping error. Therefore, additional experimentation is necessary to determine rates of genotyping errors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The author thanks John Belmont, Andy DeWan, and Bertram Müller-Myhsok for useful comments and suggestions, and Kai Yan for help in constructing the tables and supplemental figures for this article. This work was supported by the National Institutes of Health grants R01-DC03594 and HG02755.

Grant sponsor: National Institutes of Health; Grant numbers: R01-DC03594, HG02755.

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. [PubMed: 11359212]
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402. [PubMed: 9254694]
- Bailey JA, Eichler EE. 2003 Genome-wide detection and analysis of recent segmental duplications within mammalian organisms. *Cold Spring Harbor Symp Quant Biol* 68:115–124. [PubMed: 15338609]
- Boss I. 1954 Misclassification in 2×2 tables. *Biometrics* 10:478.
- Brzustowicz LM, Mérette 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. [PubMed: 8213837]
- Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW. 2003 Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol* 4:25.
- Cockerham CC. 1967 Group inbreeding and coancestry. *Genetics* 56:789–792.
- Cockerham CC. 1969 Variance of gene frequencies. *Evolution* 23:72–78. [PubMed: 28562963]
- Crow J, Kimura M. 1970 An introduction to population genetics theory. New York: Harper and Row.
- Deng HW, Chen WM, Recker RR. 2001 Population admixture: detection by Hardy-Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. *Genetics* 157:885–897. [PubMed: 11157005]
- Dhami P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP, Woodward KJ, Andrews RM, Langford C, Vetrie D. 2005 Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet* 76:750–762. [PubMed: 15756638]
- Douglas JA, Boehnke M, Lange K. 2000 A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet* 66:1287–1297. [PubMed: 10739757]
- 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. [PubMed: 11791214]
- Ehm MG, Kimmel M, Cottingham RW, Jr. 1996 Error detection for genetic data, using likelihood methods. *Am J Hum Genet* 58:225–234. [PubMed: 8554060]
- Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC. 2002 Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet* 11:1987–1995. [PubMed: 12165560]
- Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R, Jr, Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS, Laver P, Lee UK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Stames SN, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR, Wolff KK. 1996 A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13: 399–408. [PubMed: 8696333]
- Geller F, Ziegler A. 2002 Detection rates for genotyping errors in SNPs using the trio design. *Hum Hered* 54:111–117. [PubMed: 12626844]

- Gordon D, Ott J. 2001 Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac Symp Biocomput* 18–29. [PubMed: 11262939]
- Gordon D, Heath SC, Ott J. 1999a True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Hered* 49:65–70. [PubMed: 10077724]
- Gordon D, Matisse TC, Heath SC, Ott J. 1999b Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. *Genet Epidemiol* 17: 587–592.
- Gordon D, Leal SM, Heath SC, Ott J. 2000 An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. *Pac Symp Biocomput* 663–674. [PubMed: 10902214]
- 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. [PubMed: 11443542]
- Gordon D, Finch SJ, Nothnagel M, Ott J. 2002 Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered* 54:22–33. [PubMed: 12446984]
- Gordon D, Levenstien MA, Finch SJ, Ott J. 2003 Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. *Pac Symp Biocomput* 490–501. [PubMed: 12603052]
- Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, Davis RW. 2003 Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* 21:673–678. [PubMed: 12730666]
- Hartl DL, Clark AG. 1989 *Principles of population genetics*, 2nd edn. Sunderland, MA: Sinauer Associates, Inc.
- Haynam GE, Govindarajulu Z, Leone FC. 1970 Tables of the cumulative non-central chi-square distribution In: Harter HL, Owen DB, editors. *Selected tables in mathematical statistics*, volume 1 Providence, RI: American Mathematical Society p 1–78.
- Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF. 2004 Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet* 12:395–399. [PubMed: 14872201]
- Hsu TM, Chen X, Duan S, Miller RD, Kwok PY. 2001 Universal SNP genotyping assay with fluorescence polarization detection. *Biotechniques* 31:560–568. [PubMed: 11570500]
- Lincoln SE, Lander ES. 1992 Systematic detection of errors in genetic linkage data. *Genomics* 14:604–610. [PubMed: 1427888]
- Lyamichiev V, Neri B. 2003 Invader assay for SNP genotyping. *Methods Mol Biol* 212:229–240. [PubMed: 12491914]
- Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, Yang G, Webster T, Cawley S, Walsh P, Jones KW, Mei R. 2004 Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 1:146–151.
- Nielsen DA, Ehm MG, Weir BS. 1999 Detecting marker-disease associations by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* 63:1531–1540.
- Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. 2002 Bead Array technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques [Suppl]* 56-8: 60–61.
- R Development Core Team. 2004 *R: a language and environment for statistical computing*. Version 1.9. Vienna, Austria: R Foundation for Statistical Computing.
- Stringham HM, Boehnke M. 1996 Identifying marker typing incompatibilities in linkage analysis. *Am J Hum Genet* 59: 946–950. [PubMed: 8808612]
- Tiret L, Cambien F. 1995 Departure from Hardy-Weinberg equilibrium should be systematically tested studies of association between genetic markers and disease. *Circulation* 92:3364–3365. [PubMed: 7586328]
- Weir BS. 1996 *Genetic data analysis II*. Sunderland, MA: Sinauer.
- Weir BS, Hill WG, Cardon LR. 2004 Allelic association patterns for a dense SNP map. *Genet Epidemiol* 27:442–450. [PubMed: 15543640]

Xu J, Turner A, Little J, Bleecker ER, Meyers DA. 2002 Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? *Hum Genet* 6:573–574.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I.

Displays for Models 1, 3, and 4^a

Allele frequency (p)	Error rate (e)	Model 1			Model 3			Model 4		
		N	1- β	SER	N	1- β	SER	N	1- β	SER
0.1	0.05	>10,000	0.052	0.050	3,140	0.353	0.009	3,645	0.311	0.041
	0.10	>10,000	0.071	0.100	785	0.885	0.018	1,013	0.795	0.082
	0.15	>10,000	0.135	0.150	349	0.997	0.027	485	0.981	0.123
	0.20	4,175	0.278	0.200	196	1.0	0.036	287	1.0	0.164
0.3	0.05	>10,000	0.050	0.050	3,140	0.353	0.021	3,196	0.347	0.029
	0.10	>10,000	0.050	0.100	785	0.885	0.042	812	0.875	0.058
	0.15	>10,000	0.052	0.150	349	0.997	0.063	366	0.996	0.087
	0.20	>10,000	0.056	0.200	196	1.0	0.084	208	1.0	0.116
0.5	0.05	UD	0.050	0.050	3,140	0.353	0.025	3,140	0.353	0.025
	0.10	UD	0.050	0.100	785	0.885	0.050	785	0.885	0.050
	0.15	UD	0.050	0.150	349	0.997	0.075	349	0.997	0.075
	0.20	UD	0.050	0.200	196	1.0	0.100	196	1.0	0.100

^aSample size (N) necessary to detect deviation from HWE with power of 0.8 at α level of 0.05; power (1- β) for 1,000 individuals to detect deviation from HWE at α level of 0.05; and sample error rate (SER). Note that for allele frequencies $p = 0.7$ and 0.9 , results are equivalent to those obtained for $p = 0.3$ and 0.1 , respectively. UD, sample size undefined.

TABLE II.

Displays for Models 2, 3b, and 4b^a

Allele frequency (p)	Error rate (e)	Model 2			Model 3b			Model 4b		
		N	1-β	SER	N	1-β	SER	N	1-β	SER
0.1	0.05	>10,000	0.053	0.0095	>10,000	0.062	0.0090	>10,000	0.053	0.0005
	0.1	>10,000	0.061	0.0190	>10,000	0.101	0.0180	>10,000	0.062	0.0010
	0.15	>10,000	0.075	0.0285	7,731	0.172	0.0270	>10,000	0.077	0.0015
0.3	0.2	>10,000	0.094	0.0380	4,144	0.280	0.0360	>10,000	0.098	0.0020
	0.05	>10,000	0.075	0.0255	8,667	0.158	0.0210	>10,000	0.077	0.0045
	0.1	9,752	0.146	0.0510	2,148	0.481	0.0420	8,516	0.160	0.0090
0.5	0.15	4,594	0.257	0.0765	944	0.822	0.0630	3,739	0.305	0.0135
	0.2	2,745	0.394	0.1020	524	0.972	0.0840	2,078	0.494	0.0180
	0.05	>10,000	0.116	0.0375	3,216	0.346	0.0250	>10,000	0.125	0.0125
0.7	0.1	3,941	0.292	0.0750	822	0.871	0.0500	3,059	0.360	0.0250
	0.15	1,969	0.515	0.1125	372	0.996	0.0750	1,341	0.677	0.0375
	0.2	1,249	0.708	0.1500	213	1.0	0.1000	743	0.901	0.0500
0.9	0.05	8,033	0.167	0.0455	1,690	0.577	0.0210	6,435	0.197	0.0245
	0.1	2,516	0.423	0.0910	445	0.988	0.0420	1,611	0.598	0.0490
	0.15	1,404	0.657	0.1365	207	1.0	0.0630	715	0.912	0.0735
0.9	0.2	998	0.801	0.1820	122	1.0	0.0840	400	0.993	0.0980
	0.05	7,919	0.169	0.0495	1,053	0.780	0.0090	4,416	0.266	0.0405
	0.1	3,975	0.290	0.0990	285	0.999	0.0180	1,206	0.723	0.0810
0.9	0.15	3,689	0.308	0.1485	137	1.0	0.0270	569	0.960	0.1215
	0.2	4,870	0.246	0.1980	83	1.0	0.0360	333	0.998	0.1620

^aSample size (N) necessary to detect deviation from HWE with power of 0.8 at α level of 0.05; power (1-β) for 1,000 individuals to detect deviation from HWE at α level of 0.05; and sample error rate (SER).

TABLE III.

Displays for sample error rate of SER = 0.05 for Models 2, 3, 3b, 4, and 4b^a

Allele frequency (p)	Model 2				Model 3				Model 3b				Model 4				Model 4b			
	e	N	1- β	D'	e	N	1- β	D'	e	N	1- β	D'	e	N	1- β	D'	e	N	1- β	D'
0.1	0.263	15,851	0.108	0.002	0.278	102	1.0	0.025	0.278	1,974	0.514	0.004	0.061	2,017	0.423	-0.006	NA ^d			
0.3	0.098	10,124	0.142	0.006	0.119	554	0.964	0.025	0.119	1,510	0.626	0.014	0.086	1,095	0.766	-0.018	0.556	246	1.0	-0.036
0.5	0.067	8,215	0.165	0.008	0.100	785	0.885	0.025	0.100	822	0.871	0.024	0.100	790	0.885	-0.025	0.200	744	0.901	-0.026
0.7	0.055	6,803	0.189	0.007	NA ^b				0.119	320	0.999	0.034	NA ^b				0.102	1,548	0.615	-0.016
0.9	0.051	7,817	0.171	0.003	NA ^c				0.278	48	1.0	0.044	NA ^c				0.062	2,968	0.370	-0.006

^aError rate (e) at which errors were introduced. Sample size (N) necessary to detect deviation from HWE with power of 0.8 at α level of 0.05; power (1- β) for 1,000 individuals to detect deviation from HWE at α level of 0.05; and disequilibrium coefficient D' . For model 1, SER = e; therefore, results for model 1 can be found in Table I

^bEquivalent to allele frequency 0.3.

^cEquivalent to allele frequency 0.1.

^dHigh enough error rate cannot be introduced to obtain sample error rate of 0.05.

TABLE IV.Displays for Model 5^a

Allele frequency (p)	p'	D'	N	1- β
0.1	0.505	-0.245	8	1.0
0.3	0.545	-0.207	11	1.0
0.5	0.625	-0.141	22	0.989
0.7	0.745	-0.065	67	0.677
0.9	0.905	-0.009	712	0.115

^aEstimates of allele frequency (p') for specific allele frequencies at second polymorphic site; disequilibrium coefficient (D'); sample size (N) necessary to detect deviation from HWE with power of 0.8 at α level of 0.05; and power (1- β) for 50 individuals to detect deviation from HWE at α level of 0.05.

TABLE V.

Displays for Models 5b^a

	Frequency site A (p_A)	Frequency site B (p_B)	p'	D'	N	$1-\beta$
0.05	0.05		0.093	-0.009	752	0.112
	0.1		0.135	-0.018	326	0.195
	0.3		0.279	-0.078	53	0.779
	0.5		0.387	-0.150	20	0.994
	0.7		0.460	-0.210	11	1.0
0.1	0.9		0.497	-0.244	8	1.0
	0.1		0.172	-0.030	183	0.310
	0.3		0.302	-0.090	43	0.858
	0.5		0.400	-0.158	18	0.996
	0.7		0.466	-0.212	11	1.0
0.3	0.9		0.500	-0.242	8	1.0
	0.3		0.384	-0.139	23	0.986
	0.5		0.450	-0.180	15	0.999
	0.7		0.500	-0.206	12	0.999
	0.9		0.534	-0.212	11	1.0
0.5	0.5		0.500	-0.187	14	1.0
	0.7		0.550	-0.180	15	0.999
	0.9		0.600	-0.158	18	0.996

^aEstimates of allele frequency (p') for specific allele frequencies at first and second polymorphic sites; disequilibrium coefficient (D'); sample size (N) necessary to detect deviation from HWE with power of 0.8 at α level of 0.05; and power ($1-\beta$) for 50 individuals to detect deviation from HWE at α level of 0.05.