Errors and Linkage Disequilibrium Interact Multiplicatively When Computing Sample Sizes for Genetic Case-Control Association Studies

D. Gordon, M.A. Levenstien, S.J. Finch, J. Ott

Pacific Symposium on Biocomputing 8:490-501(2003)

ERRORS AND LINKAGE DISEQUILIBRIUM INTERACT MULTIPLICATIVELY WHEN COMPUTING SAMPLE SIZES FOR GENETIC CASE-CONTROL ASSOCIATION STUDIES

D. GORDON[1], M. A. LEVENSTIEN[1], S. J. FINCH[2], AND J. OTT[1]
[1]*Laboratory of Statistical Genetics, Rockefeller University*
1230 York Avenue, New York, NY 10021-6399
[2]*Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794*

Single nucleotide polymorphisms (SNP) may be used in case-control designs to test for association between a SNP marker and a disease. Such designs may assume that the genotype data are reported without error. Our goal is quantifying the effects that errors have on sample size for case-control studies with haplotypes formed by a disease locus and a SNP marker locus in the presence of linkage disequilibrium (LD). We consider the effects of a recently published error model on 2×3 chi-square analysis. We study the joint relation of LD and errors with sample size for three specific genetic disease models and two settings each of marker allele frequencies (total of 6 studies). Minimal sample size necessary for fixed asymptotic power is estimated as a 4th degree polynomial in the variables *S* (error) and *D'* (LD measure) via a backward step-wise regression.

We find that increased error rates lower power. In all studies, we observe that LD and errors interact in a non-linear fashion. In particular, regression analyses shows that several higher order interaction terms have coefficients significantly different from 0 in each study, with fraction of variance explained greater than 0.9999. Finally, the increase in sample size necessary to maintain constant asymptotic power and level of significance as a function of *S* is smallest when *D'* = 1 (perfect LD). The increase grows monotonically as *D'* decreases to 0.5 for all studies.

1 Introduction

Single nucleotide polymorphisms (SNPs) may be used in case-control designs to test for genetic association between marker and disease. Such designs usually assume that genotype data are reported without error. In statistical genetics, errors in genotyping or phenotyping (incorrectly assigning a case to be a control, or vice versa) can significantly affect linkage and genetic association studies. A number of authors have studied such effects[1-10]. Sobel et al. [11] summarize results to date. Major findings are that errors lead to inflation in genetic map distances, an increase in type I error or a decrease in power for statistical methods designed for gene localization, and biased estimates of parameters such as the recombination fraction among loci and the amount of linkage disequilibrium (LD) between two loci.

For case-control studies of genetic association, researchers[12,13] have found that, for a particular error model (not presented here), errors lead to a loss in power to detect association between a disease and a locus. However, to our knowledge, there has been no quantitative assessment of the relation between errors and LD in genetic case-control association studies for multiple disease models, although other authors[6,14-17] have developed methods that allow for errors in genetic linkage and/or association analyses.

The purpose of this work is therefore a quantitative assessment, in terms of increased sample size, of error rates in genetic case-control association studies. The data we consider is haplotype data for cases and controls from a SNP marker locus that is in LD with a disease locus. The SNP marker is observed, and the disease locus is unobserved. The test statistic considered is the standard χ^2 on 2 × 3 tables. We compute asymptotic power analytically by means of a non-centrality parameter. Errors affect the power of such statistics by deviating genotype frequencies in cases and controls away from their true values. Furthermore, determining sample size for fixed power level is equivalent to determining power for a fixed sample size, and it is this first question that we study in this work.

For three particular genetic disease models and two different settings of SNP marker allele frequencies (a total of 6 studies), we compute genotype frequencies for cases and controls in the presence of errors, and compute the sample size necessary to maintain constant asymptotic power and level of significance for different values of the error model parameters. Finally, we perform model fitting by regressing the minimal sample size necessary to maintain constant power on a 4th degree polynomial in the variables *S* (error parameter) and D' (LD parameter).

2 Materials and Methods

2.1 Notation

The following notation is used through the remainder of this work:

Count parameters:
N_A = number of cases
N_U = number of controls

Frequency parameters:
p_1 = allele frequency of SNP marker *1* allele
p_2 = allele frequency of SNP marker *2* allele = 1- p_1
p_d = allele frequency of disease locus *d* allele
p_+ = allele frequency of disease wild-type allele = 1- p_d
p_{Aij}= frequency of SNP marker genotype *ij* in the case population ($ij \in \{11, 12, 22\}$)
p_{Uij}= frequency of SNP marker genotype *ij* in the control population ($ij \in \{11, 12, 22\}$)
Disequilibrium parameters:

D= disequilibrium (non-standardized as defined in Hartl and Clark[18]) [Note: max (-$p_1 p_+$, -$p_2 p_d$) ≤ D ≤ min ($p_1 p_d$, $p_2 p_+$)]
$D_{\max}$ = min ($p_1 p_d$, $p_2 p_+$) (we assume in this work that disequilibrium is positive)
D' = proportion of total disequilibrium (or standardized disequilibrium [19])
= $D/ D_{\max}$

Penetrances:
$f_0 = \Pr(\text{affected} \mid ++ \text{ at disease locus})$
$f_1 = \Pr(\text{affected} \mid +d \text{ at disease locus})$
$f_2 = \Pr(\text{affected} \mid dd \text{ at disease locus})$

Conditional probabilities:
p_{A11} = Pr(*11* genotype at SNP locus | affected)
p_{A22} = Pr(*22* genotype at SNP locus | affected)
p_{U11} = Pr(*11* genotype at SNP locus | unaffected)
p_{U22} = Pr(*22* genotype at SNP locus | unaffected)

Prevalence and other parameters:
$\phi = \text{disease prevalence} = (1-p_d)^2 f_0 + 2(p_d)(1-p_d)f_1 + p_d{}^2 f_2$
(Note: We assume Hardy-Weinberg equilibrium (HWE) at the disease locus; no such assumption is made for the marker locus)

h_{ij} = haplotype frequency of i allele at disease locus (i = + or d) and j allele at marker locus (j = *1* or *2*) (see Methods)

Error model parameters:
ε_1 = Pr(true heterozygote incorrectly coded as a homozygote),
ε_2 = Pr(true heterozygote has one allele misread),
ε_3 = Pr(jointly misreading both alleles of a genotype),
ε_4 = Pr(falsely adding an allele to a true homozygote),
ε_5 = Pr(pre-gel error).
Sobel et al. [11] describe these parameters more completely. It should be noted that, for a di-allelic locus, the parameter $\varepsilon_2 = 0$, since it is not possible for one heterozygote to be incorrectly read as another heterozygote for a di-allelic locus.

When considering the χ^2 statistic on 2 × 3 tables, the sample size determination for fixed asymptotic power and significance level is completely determined by the non-centrality parameter λ, which is a function of the genotype frequencies in the case and control populations and the ratio of cases to controls. In section 2.2, we demonstrate how to compute genotype frequencies in each population as a function of the genetic model parameters (penetrance values, disease allele frequency), an LD parameter and the SNP marker allele frequency. In section 2.3, we present an error model and compute precisely how genotype frequencies determined in section 2.2 are altered for general settings of the error model parameters

2.2 Computation of genotype frequencies

We assume that we know the following six parameter values: the penetrance values f_0, f_1, f_2, the SNP marker allele frequency p_1, the disease allele frequency p_d, and the standardized disequilibrium *D'*. Using the definition of conditional probability, we calculate all such values Pr(*ab* at SNP marker locus | affection status)[20,21]. For example, we have the following case genotype frequency expressions:

$$p_{A11} = \Pr(11 \mid \text{affected}) = [1/(\phi)]\{(h_{+1})^2 f_0 + 2(h_{+1})(h_{d1})f_1 + (h_{d1})^2 f_2\},$$

$$p_{A12} = \Pr(12 \mid \text{affected}) = [2/(\phi)]\{(h_{+1})(h_{+2})f_0 + (h_{+1}h_{d2} + h_{d1}h_{+2})f_1 + (h_{d1})(h_{d2})f_2\},$$

$$p_{A22} = \Pr(22 \mid \text{affected}) = [1/\phi]\{(h_{+2})^2 f_0 + 2(h_{+2})(h_{d2})f_1 + (h_{d2})^2 f_2\}.$$

To compute the corresponding genotype frequencies for controls, replace ϕ by 1-ϕ and each f_i by $1 - f_i$ in each expression. The haplotype frequencies are functions of the parameters p_1, p_2, p_+, p_d, and *D'*. Using the notation defined above, we have:

$$h_{+1} = p_+ p_1 + D' D_{\max},$$

$$h_{+2} = p_+ p_2 - D' D_{\max},$$

$$h_{d1} = p_d p_1 - D' D_{\max},$$

$$h_{d2} = p_d p_2 + D' D_{\max}.$$

To obtain the genotype frequency expressions as functions of LD, substitute the haplotype relations above in the genotype frequency expressions.

2.3 Error model

Recently, Sobel, Papp, and Lange[11] proposed a model to describe how errors affect genotypes, in terms of the probabilities Pr(observed genotype is *ab* | true genotype is *cd*) (where $\{ab, cd\} \in \{11, 12, 22\}$). We call these probabilities *error penetrances.*

While their model generalizes to a marker locus with any number of alleles, we present in table 1 the error penetrances for a di-allelic locus.

Table 1 – Error penetrances for a SNP marker locus using the Sobel-Papp-Lange error model

	True Genotype		
Observed Genotype	11	12	22
11	$1-(\varepsilon_3+\varepsilon_4+\varepsilon_5)$	$(\varepsilon_1+\varepsilon_5)/2$	$\varepsilon_3+\varepsilon_5/2$
12	$\varepsilon_4+\varepsilon_5/2$	$1-(\varepsilon_1+\varepsilon_5)$	$\varepsilon_4+\varepsilon_5/2$
22	$\varepsilon_3+\varepsilon_5/2$	$(\varepsilon_1+\varepsilon_5)/2$	$1-(\varepsilon_3+\varepsilon_4+\varepsilon_5)$

Using table 1, we compute the observed genotypes for either cases or controls when errors are present. If table 1 is thought of as a 3 × 3 matrix *M*, we can compute the vector of observed case genotype frequencies in the presence of errors, $A = (p_{A11}, p_{A12}, p_{A22})^T$, (here, *T* is the transpose operator) by performing the matrix multiplication $M \times A$. For example,

$$p^*_{A11} = [1-(\varepsilon_3+\varepsilon_4+\varepsilon_5)]p_{A11} + [(\varepsilon_1+\varepsilon_5)/2]p_{A12} + [\varepsilon_3+\varepsilon_5/2]p_{A22}.$$

Note that the observed genotype frequencies are a function of both the error rates and the LD parameter. While the Sobel-Papp-Lange error model assumes 5 parameters, in order for us to present 3-dimensional plots of the interaction between LD and errors, we must reduce it to a single parameter. Therefore, we use fixed multiples of the settings: $\varepsilon_1 = 0.0125, \varepsilon_2 = 0, \varepsilon_3 = 0.005, \varepsilon_4 = 0.01, \varepsilon_5 = 0.0025$ from 0 up to 6 (increments of 0.5) from this point forward. Sobel et al. give these settings as the default settings for their error model parameters when considering a di-allelic locus [11]. The notation *S* represents the sum $k\sum_{i=1}^{5}\varepsilon_i$, where *k* =0.0, 0.5, 1,0, …, 6.0.

2.4 Non-centrality parameter

Using the notation above and a general result proved by Mitra[22], Gordon et al.[23] found that that the non-centrality parameter λ for the test of genotype frequency differences among cases and controls is given by:

$$\lambda = N_A N_U\left[\frac{(p^*_{A11}-p^*_{U11})^2}{N_A p^*_{A11}+N_U p^*_{U11}} + \frac{(p^*_{A12}-p^*_{U12})^2}{N_A p^*_{A12}+N_U p^*_{U12}} + \frac{(p^*_{A22}-p^*_{U22})^2}{N_A p^*_{A22}+N_U p^*_{U22}}\right]. \quad (1)$$

This formula provides us with the sample size for a fixed value of the non-centrality parameter. Assuming a fixed power and significance level, the non-centrality is known. It is then possible to solve equation (1) for sample sizes. We compute this solution for all genetic models presented in the next section.

2.5 Genetic models

Here we present values for the parameters in section 2.2. Each set of genetic model parameters (penetrances + disease allele frequency) comes from a genetic disease model in which the disease prevalence is 0.03 and the disease allele frequency is 0.2. In all studies, the non-centrality parameter is set to 15.4408, which corresponds to a fixed asymptotic power of 0.95 at the 0.05 level of significance for a χ^2 distribution with 2 degrees of freedom. Also, the LD parameter D' is varied between 0.5 and 1.0 in increments of 0.05. Finally, the SNP marker 1-allele frequency p_1 is set at both 0.2 and 0.5 in all studies. The genetic model parameter values are:

(Dominant model)
$f_0 = 0.004, f_1 = 0.07, f_2 = 0.07, p_d = 0.2$
(Additive model)
$f_0 = 0.014, f_1 = 0.028, f_2 = 0.042, p_d = 0.2$
(Multiplicative model)
$f_0 = 0.011, f_1 = 0.028, f_2 = 0.071, p_d = 0.2$

2.6 Regression analysis

As a further means of describing the quantitative relationship among sample size, LD, and errors, we perform a backward step-wise regression analysis. For each setting of error parameter S and the LD parameter D', the value of the dependent variable is the sample size necessary for asymptotic power 0.95 at level of significance 0.05. The general form of the fitted regression equation (i.e., the upper model) is:

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=0}^{4} \sum_{\substack{j=0 \\ i+j\leq 4}}^{4} \hat{\beta}_{i,j} S^i D'^j ,$$

where $\hat{Y}$ is the fitted sample size (in terms of case individuals) corresponding to a given setting of S and D', and the terms $\hat{\beta}_{i,j}$ are the parameters of the regression (*regression coefficients*) that minimize the sum of squares of differences between the fitted values for settings of S and D' (using equation 1) and the observed values for the same settings. The regression coefficients are determined using the S-PLUS 6.0 software (see Electronic Database Information).

3 Results

We have three main results. Our first is that, for the genetic models considered in section 2.5, there is multiplicative interaction between the error parameter S and the standardized LD D'. This interaction is documented graphically in figures 1 and 2 and quantitatively in our regression analysis results (Table 2).

Table 2 – Regression coefficients for all genetic model studies and SNP allele frequency settings

	Genetic Model[a]/SNP allele frequency					
Exponent pair (*i,j*) for term $S^iD'^j$	Dom/0.5	Dom/0.2	Mult/0.5	Mult/0.2	Add/0.5	Add/0.2
(0,0)(intercept)	6476	1837	17826	4906	46617	12518
(1,0)	7889	2753	21147	8932	54787	21280
(0,1)	-25030	-7134	-68367	-18727	-179104	-48223
(2,0)	5030	0	17624	1931	48206	7670
(0,2)	39822	11466	108940	30051	285705	77505
(1,1)	-23081	-8256	-61853	-25499	-160530	-61320
(0,3)	-29568	-8605	-81041	-22564	-212776	-58192
(3,0)	6946	3022	3924	6915	9685	12864
(2,1)	-10598	0	-33696	-3566	-93658	-17031
(1,2)	24739	8797	66397	26550	172977	64664
(0,4)	8449	2482	23195	6520	60972	16805
(2,2)	6365	0	16655	2312	47213	10279
(3,1)	-7629	-2834	0	-7629	0	-12526
(1,3)	-9269	-3209	-24737	-9625	-64782	-23796

[a](*Dom = Dominant, Mult = Multiplicative, Add = Additive*)

Figures 1 and 2 present the minimal sample size necessary to maintain constant asymptotic power of 0.95 at the 0.05 significance level for our dominant model with SNP 1-allele frequency of 0.5 and our additive model with SNP 1-allele frequency of 0.2, respectively. The sample size, as indicated above, is a function of S and D'.

We comment that in table 2, the non-zero coefficients, when tested (using the *t*-test) for being non-zero, are all significant at the 0.001 level (data not shown). The observations that several interaction terms in table 2 are significantly non-zero and that the fraction of variance (multiple R^2 value) for each regression is at least 0.9999 (data not shown) indicate that, for these error models, sample size is well fit by a high degree polynomial in the variables S and D', and hence there is significant interaction between these two variables in explaining the sample size increase.

Our second result is that the general trend of sample size increase as a function of the two variables S and D' is robust to genetic model specification for the models we consider here. This result may be observed quantitatively by noting that, for each monomial term in table 2, the sign of the regression coefficient for the non-zero coefficients is the same across genetic models and SNP allele frequency specifications, and may be observed graphically by studying figures 1 and 2. We comment the shape of the surfaces in figures 1 and 2 is identical to the shape of the surfaces for those figures determined by all other genetic model and SNP allele frequency specifications (data not shown).

The third result is that, for all values of S, sample size increase as a function of S is smallest when $D' = 1$, and is largest when $D' = 0.5$ (table 2; figures 1 and 2). This result suggests that high levels of LD, in addition to increasing power for genetic case-control studies, may have the additional benefit of mitigating the effects of errors in data in the sense of requiring the smallest possible increase in sample size for a given error setting.

4 Summary and Discussion

In this work, we have demonstrated that it is possible to compute analytically sample size requirements for genetic case-control studies in the presence of errors. In sections 2.2-2.5, we have described how these computations are done for the test of genotypic association using the 2 × 3 contingency table. Further, we have shown that, for our genetic model, error model, and LD parameter settings, sample size is accurately predicted by a polynomial of high degree in the variables S and D'. From the viewpoint of marker selection, we have documented that high levels of LD have the smallest cost, in terms of increased sample size, for a given setting of error parameters. This result should be reassuring to researchers who are planning association studies and who are concerned about errors in their data.

This work generalizes to an analytic method for sample-size calculations in the presence of errors when the observed data are haplotypes or multi-locus genotypes. One only needs to specify multi-locus error models. Perhaps the simplest and most reasonable model is one in which errors in individual marker loci are independent of errors in other marker loci. Also, this work is not restricted to just di-allelic loci; it can also be extended to markers loci with any number of alleles. The analytic price is that one has to specify multiple LD parameters and multiple allele or haplotype frequency parameters for the marker loci.

We have considered the question of interaction between errors and LD over a larger set of values for the genetic model parameters specified in section 2.2; our observation is that the interaction between S and D' is robust to genetic model specifications. That is, the shape of figures 1 and 2 is repeated for every set of genetic model parameters considered (data not shown).

An important question for this work regards parameter estimation. We are currently working on methods to determine genotyping error rates. Also, LD parameters can be estimated using inter-marker LD patterns. With traits for which the genetic model parameters are difficult to estimate, one can specify genetic model-free parameters[23] rather than the genetic model-based parameters we have specified in this work.

Software performing these calculations will be available from our website http://linkage.rockefeller.edu/pawe/ by January 2003. The program is called PAWE (Power of Association Tests With Errors).

Acknowledgments

The authors gratefully acknowledge grants K01-HG00055 and MH59492 from the the National Institutes of Health.

Electronic Database Information

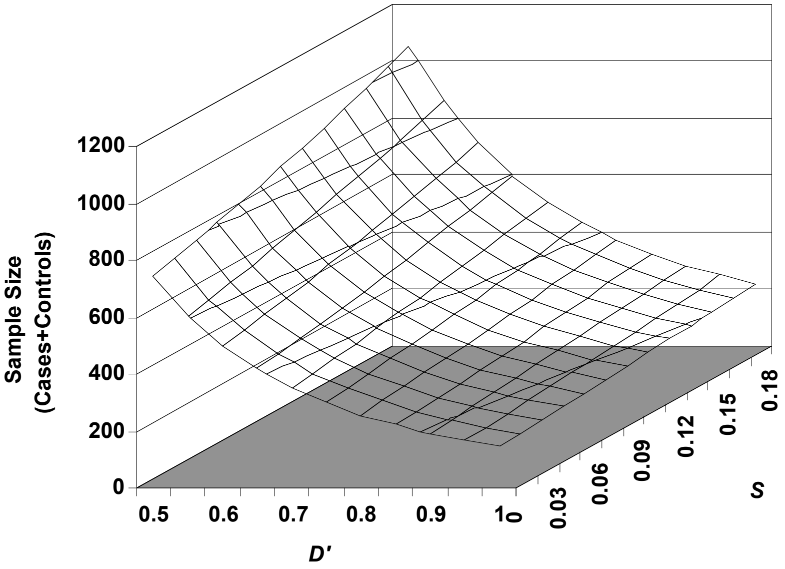
S-PLUS 6.0 Academic Site Edition Release 2. Copyright 1988-2001 Insightful Corp.

References

1. Douglas, J.A., Boehnke, M. & Lange, K. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am J Hum Genet **66**, 1287-97 (2000).
2. Shields, D.C., Collins, A., Buetow, K.H. & Morton, N.E. Error filtration, interference, and the human linkage map. Proc Natl Acad Sci **88**, 6501-5 (1991).
3. Buetow, K.H. Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet **49**, 985-94 (1991).
4. Terwilliger, J.D., Weeks, D.E. & Ott, J. 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**, A201 (1990).
5. Gordon, D., Matise, T.C., Heath, S.C. & Ott, J. Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. Genet Epidemiol **17 Suppl 1**, S587-92 (1999).
6. Gordon, D., Heath, S.C., Liu, X. & Ott, J. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet **69**, 371-80 (2001).
7. Goldstein, D.R., Zhao, H. & Speed, T.P. The effects of genotyping errors and interference on estimation of genetic distance. Hum Hered **47**, 86-100 (1997).
8. Cherny, S.S., Abecasis, G.R., Cookson, W.O., Sham, P. & Cardon, L.R. The effect of genotype and pedigree error on linkage analysis: analysis of three asthma genome scans. Genet Epidemiol **2001**, S117-22 (2001).
9. Abecasis, G.R., Cherny, S.S. & Cardon, L.R. The impact of genotyping error on family-based analysis of quantitative traits. Eur J Hum Genet **9**, 130-4 (2001).

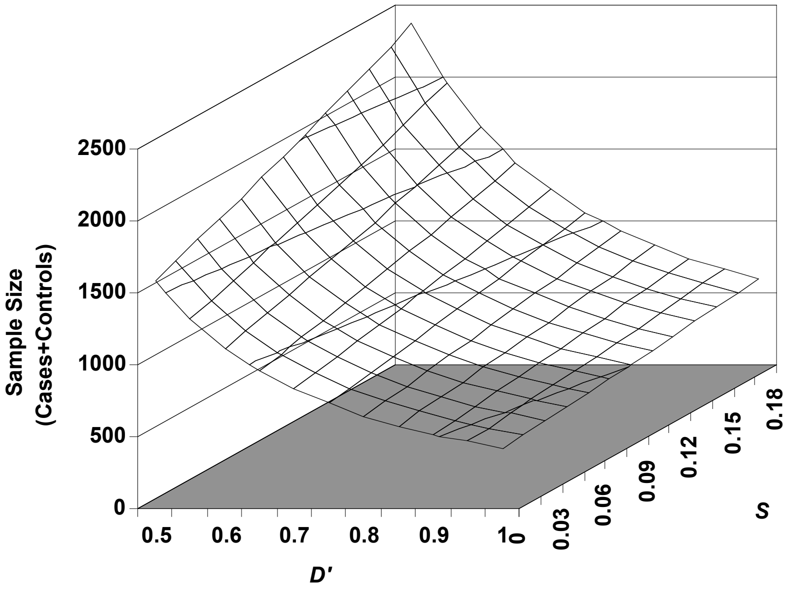
10. Akey, J.M., Zhang, K., Xiong, M., Doris, P. & Jin, L. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet **68**, 1447-56 (2001).
11. Sobel, E., Papp, J.C. & Lange, K. Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet **70**, 496-508 (2002).
12. Bross, I. Misclassification in 2 x 2 tables. Biometrics **10**, 478-486 (1954).
13. Gordon, D. & Ott, J. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput, 18-29 (2001).
14. Goring, H.H. & Terwilliger, J.D. Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. Am J Hum Genet **66**, 1107-1118 (2000).
15. Goring, H.H. & Terwilliger, J.D. Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. Am J Hum Genet **66**, 1095-106 (2000).
16. Goring, H.H. & Terwilliger, J.D. Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. Am J Hum Genet **66**, 1298-309 (2000).
17. Goring, H.H. & Terwilliger, J.D. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletnos when the mode of inheritance cannot be accurately specified. Am J Hum Genet **66**, 1310-27 (2000).
18. Hartl, D.L. & Clark, A.G. Principles of population genetics, (Sinauer Associates, Sunderland, 1989).
19. Lewontin, R.C. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49**, 49-67 (1964).
20. Risch, N. A general model for disease-marker association. Ann Hum Genet **47**, 245-52 (1983).
21. Sham, P. Statistics in Human Genetics, (J. Wiley and Sons, Inc., New York, 1998).
22. Mitra, S.K. On the limiting power function of the frequency chi-square test. Ann Math Stat **29**, 1221-1233 (1958).
23. Gordon, D., Finch, S.J., Nothnagel, M. & Ott, J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered **(in press)** (2002).

Figure 1. Minimum sample size necessary to maintain 0.95 power at 0.05 significance level for dominant genetic model, SNP 1 allele frequency = 0.5



Minimum Sample Size: 146, for *D'*=1, *S*=0
Maximum Sample Size: 1056, for *D'*=0.5, *S*=0.18

Figure 2. Minimum sample size necessary to maintain 0.95 power at 0.05 significance level for additive genetic model, SNP 1 allele frequency = 0.2



Minimum Sample Size: 416, for *D'*=1, *S*=0
Maximum Sample Size: 2376, for *D'*=0.5, *S*=0.18