Conference on Applied Statistics in Agriculture 2003 - 15th Annual Conference Proceedings

DETECTION POWER OF RANDOM, CASE-CONTROL, AND CASE-PARENT CONTROL DESIGNS FOR ASSOCIATION TESTS AND GENETIC MAPPING OF COMPLEX TRAITS

Guoping Shu

Beiyan Zeng

Oscar Smith

Recommended Citation

Shu, Guoping; Zeng, Beiyan; and Smith, Oscar (2003). "DETECTION POWER OF RANDOM, CASE-CONTROL, AND CASE-PARENT CONTROL DESIGNS FOR ASSOCIATION TESTS AND GENETIC MAPPING OF COMPLEX TRAITS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1185

Detection Power of Random, Case-Control, and Case-Parent Control Designs for Association Tests and Genetic Mapping of Complex Traits

Guoping Shu, Beiyan Zeng, and Oscar Smith

Complex Traits Genetics and Statistical Consulting, Pioneer Hi-Bred Intl, Inc., DuPont Agriculture and Nutrition, 7250 NW 62nd Ave. P.O. Box 552, Johnston, IA 50131, USA

ABSTRACT

We compared the relative detection power of random, case-control, and case-parent control (TDT) study designs by computer simulation of five parameters: Mode of inheritance (MOI), magnitude of genetic effect (γ), disease susceptibility allele frequency in the founder population (p_1), population age (t), and the genetic distance (θ) between disease susceptibility locus (D), and marker locus (M). Our results show that none of the three study designs can be claimed to be the most powerful (requiring the smallest sample size) constantly under every different genetic context (parameter combination). Our analysis indicates that both case-parent control and case-control designs have more power than the random sampling design in most genetic contexts. But the relative power between case-parent and case control depends on the specific parameter combinations. Random sampling has more power than case-parent control (although less power than case-control design) under some high genetic effect (γ) and initial allele frequency (p_1) combinations. All the three study designs show the most power under additive models of inheritance and least power under recessive mode of inheritance.

1. INTRODUCTION

Although meiotic genetic linkage analysis has been successfully used to map genes that largely control monogenic traits or diseases, this approach is less successful in the detection of genetic loci for complex and quantitative traits or diseases where the genetic effects of individual loci are relatively small and the sample size available for linkage analysis is too small to provide sufficient detection power. Association analyses or association tests, which can be done in a large population sample, have been proposed as a solution for the future of complex trait genetic mapping (Lander, 1996; Risch and Merikangas, 1996). Various study designs or sampling methods for association tests have been proposed and implemented. The three most widely used designs are: random sampling, case-control, and case-parent control. When choosing among these designs for an association study, the relative power and the minimum sample size for reaching a desired power is an important concern because of the high cost involved in sample collection and genotyping. Computer simulation under proper statistical genetic models has been employed to estimate the power for each study design (Risch and Merikangas, 1996; Teng and

Risch, 1999; Schaid, 1999; Knapp, 1999; Long and Langley, 1999; Ott, 1999; Lou and Wu, 2001; Gordon et al. 2002). The relative powers of the three study designs are difficult to derive from summarizing the results of these literatures because the results were obtained using different models and different model parameters. We are not aware of any published literature that has compared the relative powers of the three study designs under the same set of model parameters.

The goals of this study were to determine the relationship between the power of detection and sample size and to estimate the minimum samples size required to reach 80% of detection power for each of the three study designs. Computer simulation was carried out using the combination of the same set of statistical genetic parameters (mode of inheritance, initial disease allele frequency, magnitude of genetic effect).

2. METHODS

2.1 Statistical and Population Genetic Models

We assume that a complex trait is controlled by one major genetic locus and a number of minor genetic loci; the major genetic locus is biallelic (D_1 / D_2). We assume that at t generations ago, a disease susceptibility allele D_1 was introduced into a population of genotype $D_2M_1 // D_2M_2$ through a group of D_1 carrier individuals of genotype $D_1M_1 // D_2M_2$ to form a founder population, a common scenario in human immigration and in plant and animal breeding. We further assume that in the founder population (generation zero, $t = 0$), the proportion or percentage of $D_1M_1 // D_2M_2$ individuals and $D_2M_1 // D_2M_2$ individuals are P and $1 - P$ respectively and the D_1 disease susceptibility allele is in cis- position with the M_1 marker allele and normal allele D_2 is in cis- position with M_2 marker allele on a chromosome.

Table 1. Haplotype Configuration and Haplotype/Allele Frequencies At Generation ($t = 0$) with Susceptibility Allele Carrier Frequency P.

Disease Locus	Marker Locus		
	M_1 (q_1)	M_2 (q_2)	Total
Disease D_1 (p_1)	$h_{11.0} = \frac{1}{2}P$	$h_{12.0} = 0$	$\frac{1}{2}P$
Normal D_2 (p_2)	$h_{21.0} = \frac{1}{2}(1 - P)$	$h_{22.0} = \frac{1}{2}[P + (1 - P)]$	$1 - \frac{1}{2}P$
Total	$\frac{1}{2}$	$\frac{1}{2}$	1.0

The haplotype configuration and haplotype/allele frequencies at generation zero are shown in Table 1. In the table, q_1 and q_2 are the allele frequencies for M_1 and M_2 respectively and the p_1 and p_2 are the allele frequency for D_1 and D_2 respectively and $h_{11.0}$, $h_{12.0}$, $h_{21.0}$, $h_{22.0}$

are the population frequencies for haplotype D_1M_1, D_1M_2, D_2M_1, D_2M_2 at generation zero respectively. We define the disease phenotype penetrance of genotype D_1D_1, D_1D_2, and D_2D_2 as $f_{11} = P(Affected \mid D_1D_1)$, $f_{12} = P(Affected \mid D_1D_2)$, and $f_{22} = P(Affected \mid D_2D_2)$ respectively. We assume the population we sample from is a random mating population with Hardy-Weinberg equilibrium for the disease susceptibility locus, and the population prevalence of the disease is thus defined as

$$K = p_1^2 f_{11} + 2p_1p_2f_{12} + p_2^2 f_{22} \qquad (1)$$

To facilitate computer simulation we express the penetrance of three disease genotypes as a function (called penetrance function) of the major effect gene penetrance parameter $\gamma(>1)$ and minor effect gene penetrance (or background penetrance) parameter $\omega(\neq 0)$, the penetrance functions for three modes of inheritance (MOIs) are listed in Table 2.

Table 2. Penetrance Functions for Three Different Modes of Inheritance (MOI)

MOI	Penetrance Function		
	$f_{22}(D_2D_2)$	$f_{12}(D_1D_2)$	$f_{11}(D_1D_1)$
Additive	ω	$\omega\gamma$	$\omega 2\gamma$
Recessive	ω	ω	$\omega\gamma$
Dominant	ω	$\omega\gamma$	$\omega\gamma$

$\omega \neq 0$, $\gamma > 1$

2.2 Computer Simulation Parameters

We assume that each chromosome is covered with selection-neutral biallelic single nucleotide polymorphic markers, such as, SNPs. The marker density is approximately 2.0 centiMorgan (cM) and the disease susceptibility locus is 1.0 cM apart from a flanking SNP marker (M_1 / M_2). In our simulation we treat Morgan map distance of 1.0 cM as the equivalent of a recombination fraction of $\theta = 0.01$.

We assume that the input data is collected from the population of generation 10 (t=10) by one of the three sampling methods: random, case-control, and case-parent control. The haplotype frequency at current generation ($t = 10$) is computed by

$$h_{ijt} = (1-\theta)^t (h_{ij0} - p_{i0}q_{j0}) + p_{i0}q_{j0} \qquad (2)$$

Here we assume Hardy-Weinberg equilibrium for the disease susceptibility locus, that is, the disease susceptibility allele frequency remains the same after t generations of random mating. Since the haplotype frequencies change over generations, the linkage disequilibrium between disease susceptibility and marker allele, measured by linkage disequilibrium coefficient D of Lewontin (1988) also degenerate over generations:

$$D = D_t = h_{ijt} - p_i q_j \tag{3}$$

2.3 Estimation of Statistical Power and Sample Size

The statistical power is defined as $Power = 1 - \beta$, where β is the probability of Type II error or the probability of incorrectly accepting the null hypothesis H_0, thus the power is the probability of correctly accepting the alternative hypothesis H_A. The formula for sample size estimation can be derived from the formula of power calculation, which differs for different study designs (see next section for detail). We use type I error rate $\alpha = 5\times10^{-8}$ in our computer simulation to reduce possible false positives which might be a concern when applying the simulation result to guiding a study design for whole genome scanning.

We can estimate the minimum sample size for reaching any level of power for statistical association between two loci of any map distance ($0 \le \theta \le 0.5$) for a sample collected from generation t under any combination of the three genetic parameters: γ, p_1, and modes of inheritance (MOI). Due to space limitation, we only report the computer simulation result of total 48 combinations of the three genetic parameters (3x4x4) for 80% detection power at $\theta = 0.01$ and $t = 10$: (1) three modes of inheritance (MOI): additive, recessive, and dominant, (2) four levels of major gene effect (γ): 1.5, 3, 5, 7, and (3) four initial disease susceptibility allele frequencies ($p_1 = \frac{1}{2}P$): 0.05, 0.15, 0.35, 0.5. All our simulation models were implemented in the SAS Language and all our simulations are done in SAS Version 8.2 for Window (SAS Institute, 2002). The following three sections give the details of three different study designs.

2.3.1 Random Sampling Design For our computer simulation, we assume a random sample of individuals was collected from the population described in section 2.1. Individuals in the sample are sorted into a 2x3 two-way table based on their phenotype (normal or disease, $i = 0,1$) and their marker allele genotypes ($j = 0,1,2$). When the null hypothesis H_0 is true, that is, when there is no linkage disequilibrium between the disease susceptibility locus and the marker locus, we expect the joint probability estimated using the observed data, p_{ij}, is the same as the product of two marginal probabilities: $\pi_{ij}(M) = \pi_{i+}\pi_{+j}$, and p_{ij} has a central chi-square distribution $\chi^2_{(v,\alpha)}$ with degrees of freedom $v = (2-1)(3-1) = 2$. When the alternative hypothesis H_A is true, that is, when the marker locus is tightly linked with the disease susceptibility locus, the joint probability p_{ij} has a noncentral chi-square distribution $\chi^2_{(v,\lambda)}$ with $v = (2-1)(3-1) = 2$ and the noncentrality parameter of λ, which can be expressed as

$$\begin{aligned}\lambda &= n\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{[p_{ij} - \pi_{ij}(M)]^2}{\pi_{ij}(M)} \\ &= \frac{nD^2\{(f_1 - 2f_2 + f_3)^2 D^2 + 2p_1p_2[q_1f_1 + (1-2q_1)f_2 - q_2f_3]^2\}}{p_1^2p_2^2[1 - q_1^2f_1 - 2q_1q_2f_2 - q_2^2f_3][q_1^2f_1 + 2q_1q_2f_2 + q_2^2f_3]}\end{aligned} \tag{4}$$

See Appendix A for the derivation of this equation.

The statistical test for the existence of linkage disequilibrium is thus a goodness of fit chi-square test between the observed joint probability p_{ij} and the expected joint probability when the null hypothesis is true, $\pi_{ij}(M)$, in a 2 x 3 contingency table (Agresti, 1990).

The power is computed from difference in accumulated probability of the central and the noncentral chi-square distributions:

$$Power = 1 - \beta = \Pr\{\chi^2_{v,\lambda} \geq \chi^2_{v,\alpha}\} \qquad (5)$$

where v is the degree of freedom and for a 2x3 table, $v = (2-1)(3-1) = 2$.

2.3.2 Case-Control Design For the case-control design, we sample from the same population as described in Section 2.1. The population can be viewed as comprising two subpopulations; n_1 individuals were sampled from the subpopulation of affected individuals (A sample or case sample) and n_2 individuals were sampled from the subpopulation of unaffected individuals (U sample or control sample). For convenience, we assume $n_1 = kn_2$, where $k = n_1 / n_2$. In our simulation study, we assume the case sample and the control sample have equal size, and $k = 1$ and $n_1 = n_2$ although any other k value can be used in our models. The sample size for power comparison is defined as $n = n_1 + n_1 = 2n_1 = 2n_2$. The n individuals are sorted into a 2 x 3 two-way table based on their phenotype category (control or case, $i = 0,1$) and their marker genotypes ($j = 0,1,2$). When the null hypothesis H_0 is true, that is, when there is no linkage disequilibrium between the disease susceptibility locus and the marker locus, we expect the marker genotype frequencies at the case subpopulation and at the control subpopulation are equal. Thus we have the joint probability estimated using the observed data in each cell of the 2 x 3 table, p'_{ij} equal to the product of two marginal probability: $\pi_{ij} = p'_{i+}p'_{+j}$. Asymptotically, the Pearson's chi-square statistic follows a central chi-square distribution $\chi^2_{(v,\alpha)}$ (Agresti, 1990). When the alternative hypothesis H_A is true, that is, when the marker locus is tightly linked with the disease susceptibility locus, we expect the difference in marker genotype frequencies between the case and the control samples, follows a noncentral chi-square distribution $\chi^2_{v,\lambda}$ with and the noncentrality parameter

$$\lambda = kn_1^2 \left[\frac{(p'_{10} - p'_{00})^2}{n_1 p'_{10} + kn_1 p'_{00}} + \frac{(p'_{11} - p'_{01})^2}{n_1 p'_{11} + kn_1 p'_{01}} + \frac{(p'_{12} - p'_{02})^2}{n_1 p'_{12} + kn_1 p'_{02}} \right] \qquad (6)$$

where $k = n_2 / n_1$, n_1 is the size of case sample, and $n_2 = kn_1$, is the size of the control sample. Equation (6) gives the relationship between sample size and power, and we compute asymptotic power using equation (5) and the degrees of freedom $v = (3-1) = 2$. See Appendix B for more detail.

2.3.3 Case-Parental Control Design There are a number of approaches to detecting statistical association using pedigree relationship or family data (Ott, 1989; Spielman et al., 1993; Ott, 1999). A widely used approach is the TDT test, or Transmission Disequilibrium Test (Spielman et al, 1993), which use case child and parent triplet or trio data. Our power and sample size estimate is for the TDT test. We assume N number of triplets or trios (one case child and its two parents) is sampled from a random mating population, the same population used for random sampling and case-control sampling. Statistically, the TDT test is McNemar test (McNemar, 1947; Agresti, 1990; Weir, 1996; Ott, 1999). There are several different approaches to computing power and sample size for TDT test (Camp, 1997; Schaid, 1999, Knapp, 1999; Ott, 1999). The method we use here is a generalization of the method proposed by Schaid (1999) (See Appendix C for detail).

We compute statistical power using the relation between sample size and power which is expressed as

$$\sqrt{Nh}\,|\pi * -0.5| = (0.5)Z_{\alpha} + Z_{\beta}\sqrt{\pi^{*}(1-\pi^{*})} \tag{7}$$

where h is the expected number of heterozygous parents per case child, N is the number of affected case children (also the number of child-parent triplets) in the sample, $\pi *$ is the probability of transmission of an D_1 disease susceptibility allele to an affected child, Z_{α} and Z_{β} are Z values of a standard normal distribution at Type I error level of α and Type II error level of β. Equation (7) gives the relationship between N and Z_{β}, from which we can obtain sample size ($3N$) and the power ($1-\beta$). Here we use $3N$ instead of N because TDT test requires genotyping both parents of a case child, thus $3N$ is comparable to the n in random sampling and case-control. See Appendix C for details about equation (7).

3. Result

Table 3 shows the minimum sample size requirement to reach 80% power and the relative power of case-control and case-parent (TDT) designs over a random sampling design under different parameter combinations.

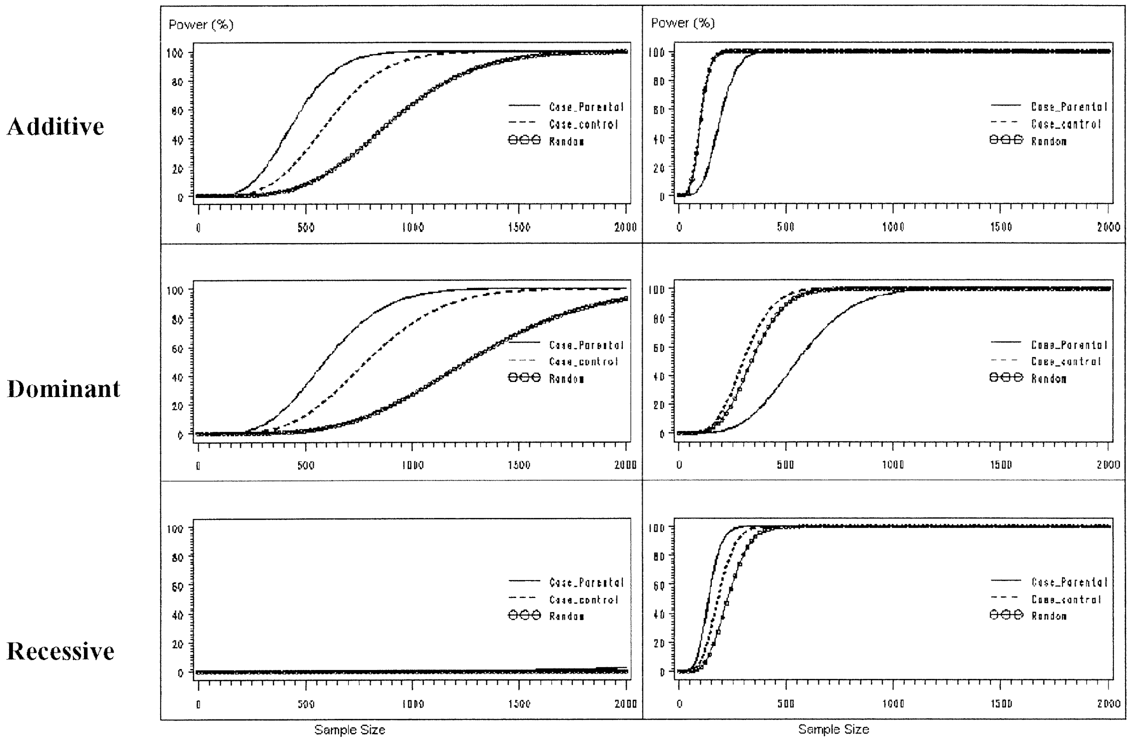
Table 3. Sample Size Necessary to Gain 80% Detction Power ($t=10, \theta=0.01$)

Gamma	p_1 (initial disease allele frequency)	Additive					Recessive					Dominant				
		Random (R)	Case-Control (CC)	Relative Power (R/CC)	Case-Parent (TDT)	Relative Power (R/TDT)	Random (R)	Case-Control (CC)	Relative Power (R/CC)	Case-Parent (TDT)	Relative Power (R/TDT)	Random (R)	Case-Control (CC)	Relative Power (R/CC)	Case-Parent (TDT)	Relative Power (R/TDT)
1.5	0.05	>100000	>100000	NA	68140	NA	>100000	>100000	NA	>100000	NA	>100000	>100000	NA	90700	NA
	0.15	28030	11785	2.38	6730	4.16	>100000	>100000	NA	>100000	NA	64475	25705	2.51	14835	4.35
	0.35	3490	1855	1.88	1140	3.06	47195	18675	2.53	13775	3.43	20395	8715	2.34	5970	3.42
	0.5	1380	885	1.56	570	2.42	11925	5135	2.32	3770	3.16	14165	6245	2.27	5625	2.52
3	0.05	26480	11175	2.37	6400	4.14	>100000	>100000	NA	>100000	NA	30625	12785	2.40	7360	4.16
	0.15	3430	1855	1.85	1180	2.91	86251	33325	2.59	25085	3.44	5245	2635	1.99	1730	3.03
	0.35	730	550	1.33	435	1.68	3385	1670	2.03	1200	2.82	1915	1165	1.64	1035	1.85
	0.5	360	325	1.11	310	1.16	955	580	1.65	415	2.30	1395	915	1.52	1155	1.21
5	0.05	7740	3695	2.09	2230	3.47	>100000	>100000	NA	>100000	NA	8705	4095	2.13	2490	3.50
	0.15	1180	795	1.48	590	2.00	22371	9215	2.43	6815	3.28	1665	1035	1.61	795	2.09
	0.35	270	260	1.04	305	0.89	985	595	1.66	425	2.32	630	505	1.25	610	1.03
	0.5	130	130	1.00	250	0.52	300	242	1.24	190	1.58	450	400	1.13	735	0.61
7	0.05	3870	2045	1.89	1300	2.98	>100000	>100000	NA	>100000	NA	4315	2235	1.93	1145	3.77
	0.15	630	495	1.27	430	1.47	10305	4495	2.29	3285	3.14	865	640	1.35	555	1.56
	0.35	130	130	1.00	265	0.49	495	350	1.41	255	1.94	312	295	1.06	495	0.63
	0.5	50	20	2.50	230	0.22	155	150	1.03	125	1.24	205	215	0.95	620	0.33

NA: no available

For additive mode of inheritance, the table shows that for a small genetic effect ($\gamma = 1.5,\ 3$), both case-control and case-parental (TDT) are more powerful than random sampling, and the TDT is more powerful than the case-control design. At a large genetic effect value ($\gamma = 5,\ 7$), the relative power of different study designs depends on the level of initial frequency of p_1 allele. At low p_1, both case-control and the TDT show more power, but at high p_1 allele frequency, the TDT has the least power and case-control has the most power in most combinations. Notes that random sampling design outperforms TDT in a number of combinations. The above result is a summary for the additive mode; it is largely true for the dominant mode. For the recessive mode, both case-control and TDT outperform random sampling designs and TDT also outperforms the case-control design (Table 3).

Figure 1. Power and Sample Size By Modes of Inheritance ($t = 10, \theta = 0.01, \gamma = 5.0$)



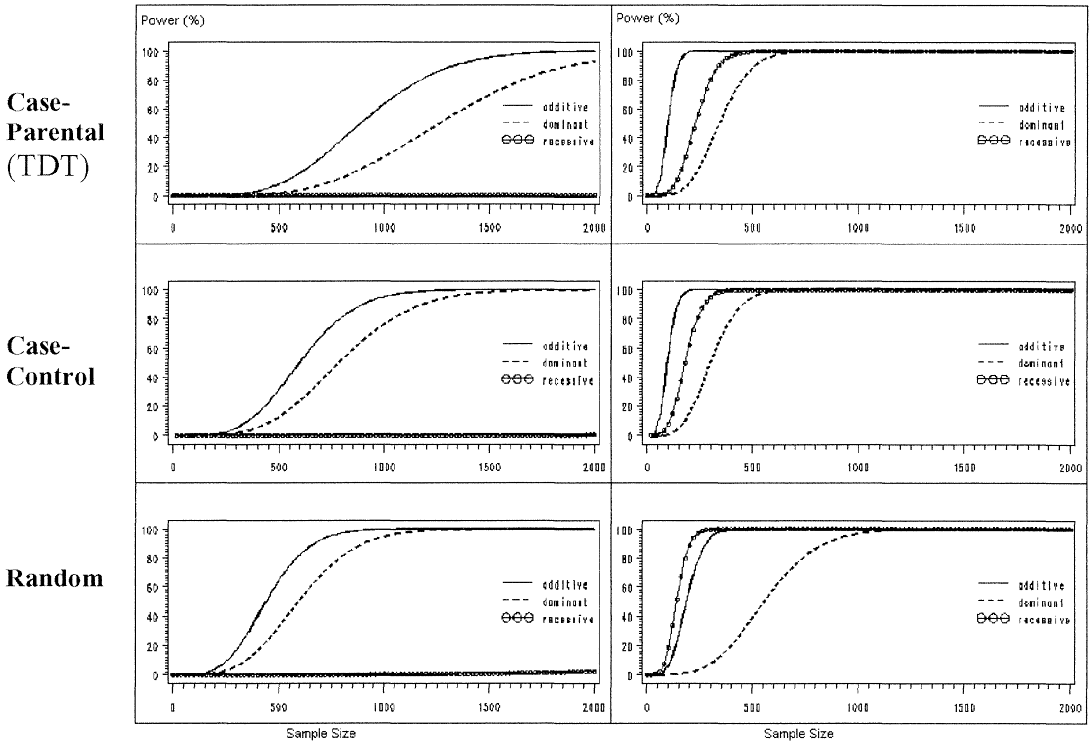
A. $p_1 = 0.15$

B. $p_1 = 0.5$

Figure 1 and Figure 2 show the relationship between the sample size and statistical power at two allele frequencies $p_1 = 0.15$ and $p_1 = 0.50$ under different study designs (Figure 1) and different modes of inheritance (Figure 2). The figures are based on the simulation results at

$\gamma = 5.0$ and $\theta = 0.01$. The left and right panel of each figure shows a difference in the rate of power increase at two levels of initial allele frequencies. Both figures show that the rate of reaching 100% power is faster at $p_1 = 0.5$ than at $p_1 = 0.15$. Figure 1 also indicates that at $p_1 = 0.15$, the sample size needed to reach high power is extremely large for recessive mode compare with the dominant mode and additive mode. However at $p_1 = 0.5$, the detection power increases rapidly for recessive mode.

Figure 2. Power and Sample Size By Study Design ($t = 10, \theta = 0.01, \gamma = 5.0$)



A. $p_1 = 0.15$ B. $p_1 = 0.5$

Figure 1 also shows that at $p_1 = 0.15$, case-parental (TDT) design shows the fastest increase in power when sample size increases whereas the random sampling shows the slowest increase. At recessive mode, all study designs require more than 2000 informative individuals to reach 80% or more power. The relative performance of three study designs are quite different at $p_1 = 0.50$.

Figure 2 shows the comparison the power and sample size among three different modes of inheritance at each study design. We can see that at $p_1 = 0.15$, the additive mode shows the fastest increase in detection power then, the dominant mode, and the recessive mode. That is, the sample size to reach the same power is the smallest for additive mode, the second for dominant mode, and the largest for recessive mode. At $p_1 = 0.50$ we see much faster increase of power for recessive mode in all three designs.

4. Discussion

It is still being debated on which study design for association analysis is the most effective (Risch, 1997; Morton and Collins, 1998; Teng and Risch, 1998; Long and Langley, 1999). Random sampling, case-control, and case-parental control are the mostly widely used study designs. Random sampling design, although widely used in population genetic modeling and molecular evolution studies because of its ease in parameterization, has been considered the least effective study design for genetic mapping especially when population disease prevalence is low (corresponding to low disease susceptibility allele frequency and low phenotype penetrance). The case-control design is widely used in genetic epidemiology and large-scale association mapping because it is a very effective design when the disease prevalence is low. But it will produce spurious associations when the population is stratified. The case-parental control design, a type of family-based study design, is a robust design in the presence of population stratification, however collecting nuclear families or parent-child trios for TDT test is costly or, in some situation, impossible (for example, for a late onset trait or disease, the parental data are often not available). Although the power for each individual study design has been reported (Luo and Wu, 2001, random sampling; Slager and Schaid, 2001, case-control, Risch and Merikangas, 1996, Teng and Risch, 1999, Long and Langley, 1999, Schaid, 1999, Knapp, 1999, case-parental control), direct comparison of their relative detection power is not available in the literature. This study is the first attempt to comparing all three designs for their relative powers using computer simulation under different combinations of five statistical and genetic parameters.

The focus of this study is to determine which study design is the most powerful design in term of minimum sample size required under the assumption that the population is not stratified and that sampling and genotyping cost per sample unit is the same for each study design. We found from these simulation studies that none of the three designs are universally superior in terms of power for all parameter combinations. For example, TDT shows more power than case-control and random sampling in most parameter combinations, but shows the lowest performance at some high γ and high p_1 combinations (Figure 1-B2, Table 3). Random sampling shows the lowest performance at most parameter combinations but performances the best at the parameter combination of $\gamma = 7$, $p_1 = 0.5$, and dominant mode (Table 3). Our simulation results demonstrate that MOI, γ, p_1 and their combinations are all critical in determining the performance for each study design. Since the true values of these factors are seldom known in real life, the power for any design is uncertain, which presents a challenge for applying association study to genetic mapping.

In addition to three simulation parameters that were widely used in previous studies: MOI, γ, and p_1, we also introduce two new parameters in our simulation models: genetic distance between the marker locus and the disease susceptibility locus, θ, and the age of population, or the number of generations after founders, t. We can estimate the power and sample size for any combination of the two parameters, although we only reported here the results for one combination of these two parameters: $\theta = 0.01$ $t = 10$ in this work. The published power analysis results of Schaid (1999) and Lou and Wu (2001) can be reproduced using our models by setting the parameter values to: $\theta = 0$, $t = 0$. Thus, their models are the special cases of our models. The significance of this generalization or extension is that our models can be used to estimate the power and sample size for both whole genome candidate gene screening and fine mapping. For candidate gene genome screening where the marker locus and the disease susceptibility locus are assumed to reside within the same candidate gene sequence and are completely linked, we set $\theta = 0$ (Risch and Merikangas, 1996; Schaid, 1999; Knapp, 1999; Lou and Wu, 2001). For fine mapping where the two loci are tightly linked but do not overlap ($0 < \theta \leq 0.5$), we can set θ to any value to simulate fine mapping using maps of different marker density. Fine mapping has been widely used for refining the genetic map location of an unknown trait/disease locus by its association with a set of mapped SNP or SSR markers of usually unknown biological functions.

In order to extend the TDT design of Schaid (1999) from complete linkage ($\theta = 0$) to tight linkage ($\theta > 0$), we use the apparent penetrance of marker genotype $f_{ij}^{'}$ instead of marker genotype relative risk $r_1 = f_{12}/f_{11}$ and $r_2 = f_{12}/f_{11}$ used in Schaid (1999). Since the disease penetrance f_{ij} is a constant for a given genotype of disease susceptibility locus, the apparent penetrance of a marker genotype is also a constant in Schaid (1999) because $f_{ij}^{'} = f_{ij}$ when $\theta = 0$. In our model, the apparent penetrance $f_{ij}^{'}$ is a function of θ (see equation (2) and (11)). The advantage of using the apparent penetrance notation is that we can express the conditional probability $P(g_k \mid A, m_l)$ (see Appendix C and column 6, Table 4) for three different MOIs using a single formula instead of three different formula as used in Schaid (1999).

Since our main interest in this study is to compare the relative detection power of different study designs, we estimate sample size using type I error rate $\alpha = 5 \times 10^{-8}$ under the assumption that the chromosome location of a trait or disease gene locus is completely unknown and a whole genome scanning is needed. In many studies, the chromosome location of a trait or disease gene locus is known (from previous linkage analysis, for example), association test is only used for fine mapping not for genome scanning. If this is the case, the type I error rate can be set to a much larger value, and the actual sample size needed to reach 80% power should be smaller than we provide here. Therefore, the minimum sample sizes reported here should be treated as very conservative estimates.

Acknowledgments

We thank our colleagues, Antoni Rafalski and Stanley Luck for many helpful discussions. This work benefited from suggestions and comments of anonymous reviewers. We thank Dr. George Milliken and other conference organizers at the Department of Statistics, Kansas State University for their support.

Table 4. Conditional Probability of Observing Case Marker Genotype in Case-Parent Control Designs

Parental Mating Type (m_l)	$P(m_l \mid A)$	g_k	$P(g_k \mid m_l)$	Y^{TDT}	$P(g_k \mid A,\ m_l)$	$P(g_{kil} \mid A)$
$M_1M_1 \times M_1M_1$	$q_1^4 \frac{f'_{11}}{K'}$	M_1M_1	1	—	1	$q_1^4 \frac{f'_{11}}{K'}$
$M_1M_1 \times M_1M_2$	$2q_1^3q_2 \frac{f'_{11}+f'_{12}}{K'}$	M_1M_1 M_1M_2	1/2 1/2	1 0	$f'_{11}/(f'_{11}+f'_{12})$ $f'_{12}/(f'_{11}+f'_{12})$	$2q_1^3q_2 \frac{f'_{11}}{K'}$ $2q_1^2q_2^2 \frac{f'_{12}}{K'}$
$M_1M_1 \times M_2M_2$	$2q_1^2q_2^2 \frac{f'_{12}}{K'}$	M_1M_2	1	—	1	$2q_1^2q_2^2 \frac{f'_{12}}{K'}$
$M_1M_2 \times M_1M_2$	$q_1^2q_2^2 \frac{f'_{11}+2f'_{12}+f'_{22}}{K'}$	M_1M_1 M_1M_2 M_2M_2	1/4 1/2 1/4	1,1 1,0 0,0	$f'_{11}/(f'_{11}+2f'_{12}+f'_{22})$ $2f'_{12}/(f'_{11}+2f'_{12}+f'_{22})$ $f'_{22}/(f'_{11}+2f'_{12}+f'_{22})$	$q_1^2q_2^2 \frac{f'_{11}}{K'}$ $q_1^2q_2^2 \frac{2f'_{12}}{K'}$ $q_1^2q_2^2 \frac{f'_{22}}{K'}$
$M_1M_2 \times M_2M_2$	$2q_1q_2^3 \frac{f'_{12}+f'_{22}}{K'}$	M_1M_2 M_2M_2	1/2 1/2	1 0	$f'_{12}/(f'_{12}+f'_{22})$ $f'_{22}/(f'_{12}+f'_{22})$	$2q_1q_2^3 \frac{f'_{12}}{K'}$ $2q_1q_2^3 \frac{f'_{22}}{K'}$
$M_2M_2 \times M_2M_2$	$q_2^4 \frac{f'_{22}}{K'}$	M_2M_2	1	—	1	$q_2^4 \frac{f'_{22}}{K'}$

Appendix A

The joint probability of the phenotype and marker genotype of an individual in a 2 x 3 two-way table, p_{ij}, described in Section 2.3.1 is estimated by

$$\begin{aligned}
p_{10} &= \Pr(A, M_2M_2) = f_{11}h_{12}^2 + f_{12}(2h_{12}h_{22}) + f_{22}h_{22}^2 \\
p_{11} &= \Pr(A, M_1M_2) = f_{11}(2h_{11}h_{12}) + f_{12}[(2h_{11}h_{22}) + (2h_{12}h_{21})] + f_{22}(2h_{21}h_{22}) \\
p_{12} &= \Pr(A, M_1M_1) = f_{11}h_{11}^2 + f_{12}(2h_{11}h_{21}) + f_{22}h_{21}^2 \\
p_{00} &= \Pr(U, M_2M_2) = (1-f_{11})h_{12}^2 + (1-f_{12})(2h_{12}h_{22}) + (1-f_{22})h_{22}^2 \\
p_{01} &= \Pr(U, M_1M_2) = (1-f_{11})(2h_{11}h_{12}) + (1-f_{12})[(2h_{11}h_{22}) + (2h_{12}h_{21})] + (1-f_{22})(2h_{21}h_{22}) \\
p_{02} &= \Pr(U, M_1M_1) = (1-f_{11}h_{11}^2) + (1-f_{12})(2h_{11}h_{21}) + (1-f_{22})h_{21}^2
\end{aligned} \tag{8}$$

From equation (3), we have

$$\begin{aligned}
h_{11} &= h_{D_1M_1} = p_1q_1 + D, \quad h_{12} = h_{D_1M_2} = p_1q_2 - D \\
h_{21} &= h_{D_2M_1} = p_2q_1 - D, \quad h_{22} = h_{D_2M_2} = p_2q_2 + D
\end{aligned} \tag{9}$$

Replace h_{ij} in (8) and using (9), we obtain (4). See Lou and Wu (2001) for a different approach of obtaining (4).

Appendix B

For computing noncentrality parameter λ, we obtain the conditional probability of marker genotype given a phenotype in case-control study design of 2.3.2 from the joint probability given in (8) of Appendix A,

$$\begin{aligned}
p_{10}^{'} &= \Pr(M_2M_2 \mid A) = \frac{p_{10}}{K}, \; p_{11}^{'} = \Pr(M_1M_2 \mid A) = \frac{p_{11}}{K}, \; p_{12}^{'} = \Pr(M_1M_1 \mid A) = \frac{p_{12}}{K} \\
p_{00}^{'} &= \Pr(M_2M_2 \mid U) = \frac{p_{00}}{1-K}, \; p_{01}^{'} = \Pr(M_1M_2 \mid U) = \frac{p_{01}}{1-K}, \; p_{02}^{'} = \Pr(M_1M_1 \mid U) = \frac{p_{02}}{1-K}
\end{aligned} \tag{10}$$

By incorporating (10) into Mitra (1958), we have equation (6) (Mitra, 1958; Agresti, 1990; Gordon et. al., 2002).

Appendix C

There are four steps to obtain the value of equation (7).

(1) Compute the conditional probability of observing case child given the marker genotype $g_k = M_iM_j$ (which is also called apparent penetrance of a marker genotype):

$$\begin{aligned}
f_{ii}^{'} &= P(A \mid g_k = M_iM_i) = \frac{f_{11}h_{1i}^2 + f_{12}(2h_{1i}h_{2i}) + f_{22}h_{2i}^2}{q_i^2} \\
f_{ij}^{'} &= P(A \mid g_k = M_iM_j) = \frac{f_{11}(2h_{1i}h_{1j}) + f_{12}(2h_{1i}h_{2j} + 2h_{1j}h_{2i}) + f_{22}(2h_{2i}h_{2j})}{2q_iq_j}
\end{aligned} \tag{11}$$

The conditional probability given by Schaid (1999) is a special case of this general formula at $\theta = 0$ where $f'_{ij} = f_{ij}$.

(2) Compute the conditional probability of the kth marker genotype (g_k) for the case child given the lth parental mating type (m_l), $P(g_k \mid A,\ m_l)$:

$$P(g_k \mid A,\ m_l) = \frac{P(A \mid g_k)P(g_k \mid m_l)}{\sum_{k=1}^{3}[P(A \mid g_k)P(g_k \mid m_l)]} \qquad (12)$$

where $P(A \mid g_k)$ is the apparent penetrance f'_{ij} defined in expression (11) and $P(g_k \mid m_l)$ is the probability of observing marker genotype g_k in a child given parental mating type m_l (column 4, Table 4).

There are three marker genotypes for all offspring cases (column 3 Table 4) and 6 possible parental mating types (column 1, Table 4). However only the conditional probability $P(g_k \mid A,\ m_l)$ for three informative mating types (type 2, 4, 5, column 1, Table 4) need to be computed and the three noninformative types (type 1, 3, 6) are always 1 (column 6, Table 4).

(3) Compute probability of observing marker genotype g_k in case child given mating type m_l:

$$P(g_{k|l} \mid A) = P(m_l \mid A)P(g_k \mid A,\ m_l) \qquad (13)$$

which is the product of column 2 and column 6 in Table 4. The result is reported in column 7 of Table 4. Equation (13) applies to any mode of inheritance (MOI).

(4) Specify y^{TDT}, Compute h and π^*: here, y^{TDT} is the number of M_1 marker alleles transmitted or not transmitted from a heterozygous parent to the case child, which has values either 1 or 0 and is listed in column 5 of Table 4. The h is the expected number of heterozygous parents per case child and is computed using

$$h = \sum_{l=2,4,5} w_l \sum_k P(g_{k|l} \mid A) \qquad (14)$$

where w_l is the weight, which has values of 2 when both parents are heterozygous and 1 for all other mating types. The π^* in equation (7), the probability of transmission of and D_1 disease susceptibility allele from n_k heterozygote parents to an affected child, is computed from

$$\pi^* = [\sum_{l=2,4,5} \sum_k y^{TDT} P(g_{k|l} \mid A)] / Nh \qquad (15).$$

Reference

Agresti A. 1990. Categorical Data Analysis. Johns Wiley and Sons, New York.

Camp N. J. 1997. Genome wide Transmission/Disequilibrium testing-consideration of the genotypic relative risks at disease susceptibility loci. Am. J. Hum. Genet. 61:1424-1430.

Chapman N. H. and Wijsman E. M. 1998. Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am. J. Hum. Genet. 63: 1872-1885.

Gordon D., Finch S. J., Nothnagel M., and Ott J. 2002. Power and sample size calculations for case-control genetic association tests when errors are present: applications to single nucleotide polymorphisms. Hum. Hered. 54: 22-33.

Knapp M. 1999. A note on power approximations for the Transmission/Disequilibrium test. Am. J. Hum. Genet. 64: 1177-1185.

Long A.D. and Langley C.H. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9: 720-731.

Lander E. S. 1996. The new genomics: global views of biology. Science 274: 536-539.

Lewontin R.C. 1988. On measures of gametic disequilibrium. Genetics 120: 849-852.

Lou Z. W. and Wu Chung-I. 2001. Modeling linkage disequilibrium between a polymorphic marker locus and a locus affecting complex dichotomous traits in natural populations. Genetics 158: 1785-1800.

McNemar Q. 1947. Note on the sampling error of difference between correlated proportions or percentages. Psychometrika 12: 153-157.

Mitra S. K. 1958. On the limiting power function of the frequency chi-square test. Annals of Mathematical Statistics 29: 1221-1233.

Morton N. E. and Collins A., 1998. Tests and estimates of allelic association in complex inheritance. Proc. Natl. Acad. Sci. USA 95: 11389-11393.

Ott J. 1989. Statistical properties of the haplotype relative risk. Genet. Epidemiol. 6: 127-130.

Ott J. 1999. Human Genetic Linkage Analysis. John Hopkins University Press.

Sasieni P. D. 1997. From genotypes to genes: double the sample size. Biometrics 53: 1253-1261.

Slager S. L. and Schaid D.J. 2001. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum. Hered. 52: 149-153.

SAS Institute Inc. 2000. The SAS System Version 8.2. Cary, NC.

Schaid D. J. 1999. Likelihoods and TDT for the Case-Parents design. Genet. Epidemiol. 16: 250-260.

Spielman R. S., McGinnis R.E., and Ewens W.J. 1993. The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM), Am. J. Hum. Genet. 52: 506-516.

Risch N. and Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273: 1516-1517.

Risch N. 1997. Genetic analysis of complex disease. Science 275: 1329-1330.

Teng J. and Risch N. 1999. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. Genome Research 9: 234-242.

Weir B. S. 1996. Genetic Data Analysis II. Methods for Discrete Population Genetic Data. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.