[Frontiers in Bioscience 8, s68-80, January 1, 2003]

DESIGNING AN OPTIMUM GENETIC ASSOCIATION STUDY USING DENSE SNP MARKERS AND FAMILY-BASED SAMPLE

Chi. C. Gu [1] and D.C. Rao [1-2]

[1] Division of Biostatistics, and [2] Departments of Genetics and Psychiatry, Washington University School of Medicine, St. Louis, MO 63110

TABLE OF CONTENTS

1. ABSTRACT

Genetic association analysis using thousands of single nucleotide polymorphism (SNP) markers has become a promising alternative to genome-wide linkage scan. Analysis based on linage-disequilibrium (LD) is more efficient because meiotic information of past generations is utilized. However, in addition to the physical distance between the disease locus and a marker locus, numerous other factors such as admixture, genetic drift, and multiple mutations can affect the observed value of LD. The effect of these factors in a genomic LD association study must be carefully analyzed to obtain an efficient study design. In the following review, we consider studies using family-based data and carefully study the effects of some of these important design factors, including the sample size, frequency of SNP markers, and marker density. For example, we conclude that (1) for reasonably frequent SNP markers, a moderately large sample of 500 families is appropriate for a moderately stringent significance level (α = 0.00009); (2) to maintain a power of 80%, maximal difference in allele frequencies between the disease gene and a SNP marker varies between 0.1 (under additive model) and 0.5 (multiplicative); (3) a map density of *10c*M is appropriate only under idea scenario (moderately large sample size, equal trait/marker allele frequencies, maximum LD strength etc.). Results shown here should have practical implications to designing efficient LD association studies using dense SNP markers.

2. INTRODUCTION

Genetic studies of complex diseases have progressed toward a more global view as more and more high quality genetic markers become available. The movement to global analysis is also due to the large number of causative factors (genetic and non-genetic) involved in the etiology. Researchers now-a-days take more systematic approaches and perform genome-wide linkage scans using hundreds of genetic markers and hope that the global search will result in important leads. Certainly, methods are evolving for finding complex disease genes (e.g., see Rao and Province, 2001) (1).

While genome-wide linkage scan remains the workhorse of many ongoing studies of complex diseases, its utility is limited to detecting modest to large genetic effects. A promising alternative design involves linkage disequilibrium (LD) association studies using thousands of single nucleotide polymorphism (SNP) markers across the whole genome. This approach is potentially more powerful because it utilizes more information across the whole genome and from meioses in past generations. Under ideal circumstances, such a design may require substantially less sample size than linkage (2).

The choice of the LD approach was motivated because LD is a much finer measure of the physical distance between the disease susceptibility locus and a polymorphic marker. The reality is, however, not as simple as it first appears. Numerous publications of genetic association findings point to conflicting stories that rarely lead to discovery of actual disease genes. This reality reflects a lack of attention to some important factors in a LD study design. It is a well-known fact that association studies can lead to spurious results if underlying population stratification is not taken into account. This problem can be dealt with by constructing family-based controls and using

transmission/disequilibrium tests (TDT). However, everything does not stop with the application of TDT. Numerous other design factors can affect the efficiency of a LD association study.

In this article, we take a close look at the effect of several such factors in the context of LD association studies using densely placed SNP markers. Given the fact that many ongoing genetic studies of complex diseases have accumulated large samples of pedigree data that were originally collected for genomic linkage scans, we anticipate that many genetic association studies in the future will utilize powerful LD methods on such existing resources. We therefore concentrate here our analysis on association methods using family-based samples only.

3. LINKAGE DISEQUILIBRIUM MAPPING

The primary advantage of linkage disequilibrium (LD) analysis is its ability to utilize the recombination information in past generations to achieve finer mapping resolution (3). Linkage disequilibrium is a measure that reflects the fact that alleles at tightly linked loci on an ancient haplotype will result in higher than expected frequency in later generations. Consider a QTL locus A with alleles *A* and *a*, and a SNP marker locus with alleles *M* and m. Let P_{AM} denote the observed frequency of haplotype *AM*, and p_A = f(*A*) and p_M = f(*M*) the corresponding allele frequencies. Then, $D = P_{AM} - p_A * p_M$ gives a basic measure of the linkage disequilibrium. The rationale for using LD in fine mapping originates from the relationship $D_t = D_0(1-\theta)^t$, where the index of *D* indicates the initial (*0*) and the *t*-th generation, and θ the recombination fraction between the two loci. It is obvious from this simple relationship (assuming constant population size, etc.) that LD decreases as the marker moves away from the disease locus, which provides the basis for testing the existence of a disease locus by measuring the strength of LD along the genome. The LD also decays rapidly over the generations. This should lead to a refined mapping resolution for moderately old disease mutations.

The LD method has been effectively used in the final stage to localize mutations for Mendelian disorders (4-6). For complex diseases, successful stories are yet to come except for the few examples in Mendelian subsets of common diseases such as ApoE4 for Alzheimer's and BRCA1 & 2 for breast cancer. The challenge of complex diseases comes from the multiple factors involved in the etiology and the unspecified past population history rendering the LD relationship less tractable. More realistic models were introduced to account for past population history such as admixture, genetic drift, multiple mutations, and natural selection (7-11). An excellent review was given by Jorde (6) on some of the progress in this front. We will concentrate in the following sections on how controllable factors in an association study, in particular the characteristics of SNP markers, can be determined to result in an optimum study.

4. FAMILY-BASED ASSOCIATION STUDY

There are basically two types of designs of genetic association studies: population based (case-control) or family-based samples. The choice between the two approaches is determined more by the constraints we are faced with than power considerations per se, although differences in power can be substantial and has been reported (12;13). A case-control design is generally more powerful, but not without limitations. It may suffer from problems such as undetected population stratification. This arises when there are more than two subpopulations in the sample with different allele frequencies and disease prevalence. Positive association in this case may be merely a reflection of the subdivision and may have nothing to do with linkage disequilibrium of interest to the disease mapping (14).

Family-based designs on the other hand, have several advantages: (1) They construct "virtually matched" controls by utilizing alleles that were not transmitted to the affected offspring or alleles transmitted to the unaffected offspring; (2) They utilize data (genetic and epidemiological) already collected in many existing/ongoing linkage studies and can cross check with the linkage analysis results invoking two stage approaches; (3) Recently developed methods that carry out combined linkage and association analysis provide additional ability to resolve false-positive results; (4) When methods of haplotype analysis are used, multi-generation pedigree data provide better protection against poorly estimated haplotype frequencies.

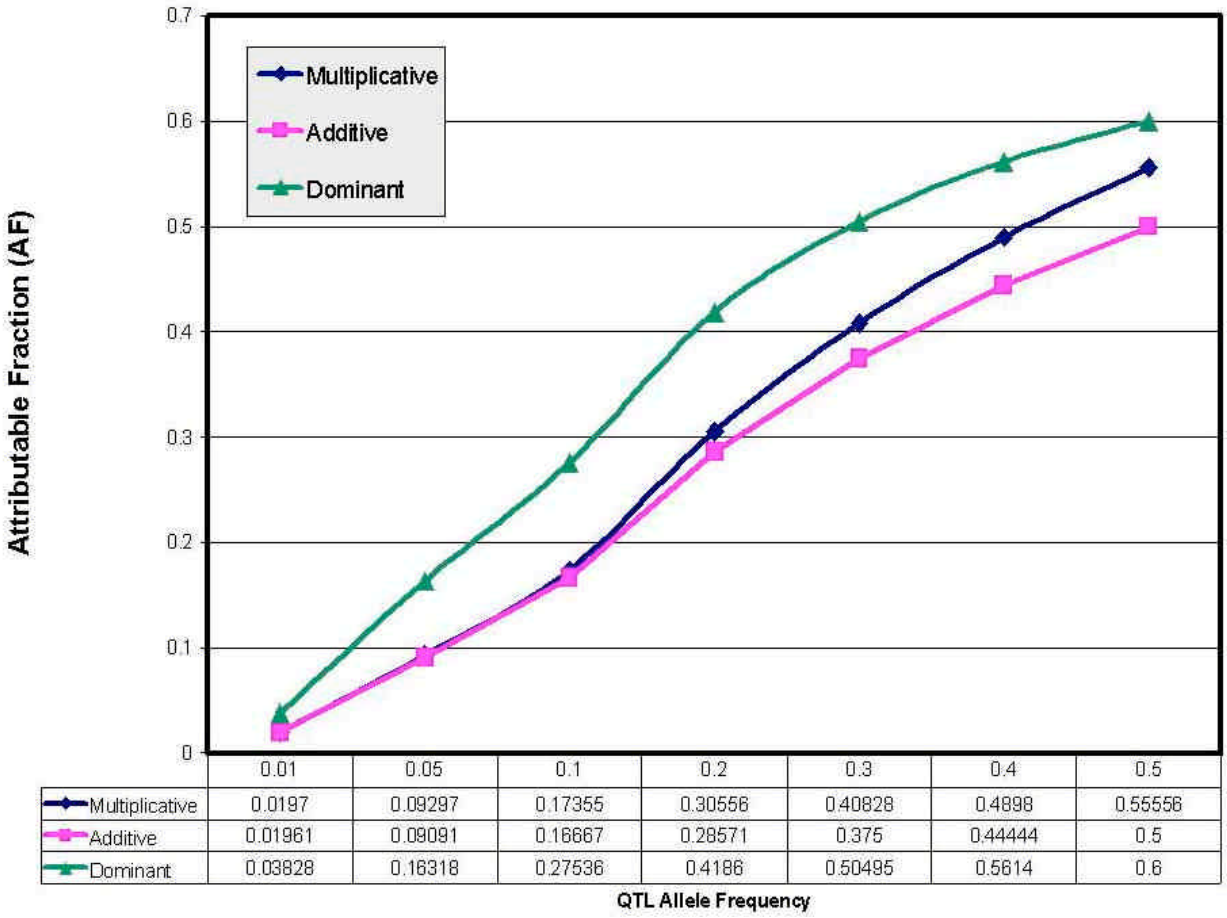
5. STUDY DESIGN ISSUES

As pointed out by many people time and again, for genetic studies of complex diseases the study design can be a determining factor for its success (15;16). Many extensions and modifications have been carefully studied since the early introduction of family controls by Falk and Rubinstein (17). Zhao gave a nice review of the mathematics behind the many variations of TDT tests currently in use (18). Most of the analysis performed in the following sections will be focused on a hypothetical LD study design with a reasonably large sample (<750) of nuclear families. A TDT extension of Knapp's method by Chen and Deng and its companion computer program will be used to calculate the power under various models (19). We will characterize properties of SNP markers that are essential for an optimum LD association study design. We want to look at the issues that influence the power of such a study.

6. WHAT SAMPLE SIZE IS APPROPRIATE

The first question that should be dealt with in any genetic study is how big a sample is appropriate. This is essential both for scientific reasons and for practical reasons such as budget and fieldwork. However, there is no uniform answer to this question: the same sample size may not lead to the same power because of the information content of a sample may vary (missing parental information etc.). Several recent publications gave detailed analysis of power using nuclear families and various forms of TDT tests (20-22).

To come up with practically usable sample sizes, we assume for the moment that a SNP marker is tightly



	0.01	0.05	0.1	0.2	0.3	0.4	0.5
Multiplicative	0.0197	0.09297	0.17355	0.30556	0.40828	0.4898	0.55556
Additive	0.01961	0.09091	0.16667	0.28571	0.375	0.44444	0.5
Dominant	0.03828	0.16318	0.27536	0.4186	0.50495	0.5614	0.6

Figure 1. Genetic models used for power calculation. Three types of genetic models (multiplicative, additive and dominant) are considered. This figure shows the effect sizes of the underlying disease allele in terms of attributable fraction of cases in the population (23).

linked to the trait locus and the frequencies are the same for the alleles in LD. Relaxation of this assumption will reduce power, as will other factors that may weaken the strength of linkage disequilibrium. These will be considered in the following sections. We considered three practical sample sizes for the power calculation: a moderate sized study with 250 nuclear families each with 3 offspring, a large study with 500 such families, and a really large study with a mixture of 500 families with 3 offspring and 250 with 2 offspring. We then tested a wide spectrum of underlying disease models to study the power of TDT. We let the disease allele frequency vary from very rare (0.01) to fairly common (0.5). Three types of genetic models were considered: multiplicative, additive and dominant. A measure of attributable fraction (23) is used to measure the genetic effect size at each allele frequency under the 3 genetic models. As shown in Figure 1, the models were chosen so that for a fixed value of allele frequency, the genetic effect attributable to the disease allele is closely comparable across the three models. The power under all these models and sample sizes is summarized in Table 1. We see that even using a moderately stringent threshold of $\alpha = 0.00009$, susceptible genes having common effects in most of the 500 families can be detected by TDT test under all the models tested for marker alleles that are reasonably frequent ($p >= 0.1$).

7. WHAT MARKER ALLELE FREQUENCY

We have seen above how the frequency of the disease allele affects the attributable fraction of cases in population to the gene effect, as well as the power of LD association test. Unlike linkage designs where rare disease and low disease allele frequencies lead to more power, association analysis is more efficient when disease allele is more frequent for a common disease. Analysis by Chapman and Wijsman (24) showed that markers with equifrequent alleles give most powerful LD test. For a biallelic SNP, this also points to common disease alleles.

The matter is complicated by the fact that LD is dependant on allele frequencies. In its basic form, D, the strength of LD has upper and lower bounds dependant on the allele frequencies: $D_{max} = \min[p_A*(1- p_M), (1-p_A)*p_M]$ if $D>0$; $\max[-p_A*p_M, -(1-p_A)*(1- p_M)]$, if $D<0$. Through simple mathematics, one can see why LD strength achieves its maximum when alleles in LD have similar frequencies and the magnitude of LD depends on the value of the

Table 1. Power to detect genetic association for studies of moderate to large sample sizes

	a = **0.001**			a = **0.00009**		
p	**N=250**	**N=500**	**N=500 +250**	**N=250**	**N=500**	**N=500 +250**
Multiplicative						
0.01	1.0%	3.1%	7.4%	0.1%	0.6%	1.8%
0.05	15.2%	48.9%	80.8%	4.6%	24.8%	58.5%
0.10	43.6%	88.5%	99.2%	20.7%	70.8%	96.0%
0.20	80.8%	99.6%	100.0%	58.5%	97.8%	100.0%
0.30	93.2%	100.0%	100.0%	79.7%	99.8%	100.0%
0.40	97.1%	100.0%	100.0%	89.1%	100.0%	100.0%
0.50	98.4%	100.0%	100.0%	93.2%	100.0%	100.0%
Additive						
0.01	1.0%	3.1%	7.3%	0.1%	0.6%	1.8%
0.05	14.3%	46.7%	78.4%	4.2%	23.0%	55.2%
0.10	39.5%	85.4%	98.6%	17.8%	65.4%	93.7%
0.20	71.9%	98.8%	100.0%	46.8%	94.5%	99.9%
0.30	83.3%	99.7%	100.0%	61.9%	98.4%	100.0%
0.40	86.2%	99.9%	100.0%	66.7%	99.0%	100.0%
0.50	85.0%	99.8%	100.0%	64.8%	98.8%	100.0%
Dominant						
0.01	2.2%	8.9%	27.0%	0.3%	2.1%	9.5%
0.05	41.8%	89.1%	99.6%	18.3%	70.2%	97.7%
0.10	81.5%	99.7%	100.0%	57.7%	98.2%	100.0%
0.20	97.3%	100.0%	100.0%	88.8%	100.0%	100.0%
0.30	98.5%	100.0%	100.0%	93.0%	100.0%	100.0%
0.40	97.6%	100.0%	100.0%	89.7%	100.0%	100.0%
0.50	92.2%	100.0%	100.0%	76.4%	99.8%	100.0%

frequencies. Several variations of D were proposed in attempt to reduce the dependency of D on allele frequencies. Devlin and Risch (25) studied commonly used measures of LD, including Lewontin's D', Hill's D, Bengtsson and Thomson's d, and Yule's Q (26-29). They concluded that d is directly proportional to the genetic distance and is the best measure for simple LD mapping. For rare diseases and randomly sampled haplotypes, $D' \approx d$ is also more favorable than the other two measures. These various forms of LD measures share the same numerator, which is D, and uses different denominator to standardize D. All of them are frequency dependant, though some are less sensitive than others. Even the well-behaved ones can become dependent with certain past population history such as heterogeneous haplotype backgrounds (30). Recently, Malecot's measure r of LD was introduced for a better resolution of the dependency problem (31).

When the frequency of a chosen marker allele is not in sync with that of the disease allele in LD (disparity of allele frequencies), the total LD strength is compromised. This will lead to reduced power of LD association test. Since we do not know the allele frequency of the disease allele in question, we cannot determine the best marker allele frequencies for an optimum study. However, we can approach the problem from a slightly different angle. We can test the effect on power of various combinations of disease and SNP allele frequencies across a wide spectrum of hypothetical models and look for ranges of (p,q) that give acceptable statistical power over a large variety of models. In Figure 5, we plot such power surfaces under the three types of disease models over all combinations of p and q between 0.05 and 0.95 on a 0.05-grid. It is clear that the underlying genetic model influences the level of robustness of the LD association test against the disparity of allele frequencies. Such influences are more apparent in the contour plots shown in Figure 6. Area colored green in Figure 6 covers those (p, q) pairs that can achieve a power of 80% and above at a significant level of 0.0009; and in light blue those can achieve a power 60-80%. The moderately large sample of 500 families was used for the calculation. Additive models are most sensitive, for which to maintain higher than 80% of power the difference between QTL and SNP allele frequencies has to be less than 10% for even fairly common disease alleles ($p=0.2$). Maximum tolerance is about 20% when QTL allele is fairly common ($p\sim0.5$). We see that although multiplicative models is not most powerful for a given disease allele frequency (see table 1 and Figure 1), it has the highest tolerance of disparity in allele frequencies. For highly common QTL alleles ($p\sim0.5$-0.7), it can tolerate an allele frequency up to about 50% and still maintain a power above 80%.

8. VARIABLE LD STRENGTH

We have seen in the previous section that total LD strength is dependent on the allele frequencies. Adding to the complexity, observed values of LD are often less than the maximum possible LD values. Even when the marker locus is tightly linked to the disease locus, where genetic distance is essentially zero, other factors such as genetic drift and allelic heterogeneity can lead to weakened LD strength observed in a sample. Tu and Whittemore (32)

gave a detailed analysis of the effect of weakened LD strength on the power of LD analysis. Their analysis showed that lost of LD strength combined with less frequent disease allele could render a LD association study less favorable to a linkage design.

We have analyzed the power loss due to weakened LD strength caused by any known or unknown reasons. Displayed in Figure 2 is a study for fairly common alleles (p=0.2), and use our large sample of 500 families, such a design can tolerate up to (~25%) loss in LD strength without suffering substantial loss in power. Combined with knowledge learned from population genetic studies on distribution of LD among loci in interested region (33;34), results of the power analyses can provide a practical guideline of the sizes of detectable association effects.

9. FEASIBLE MARKER DENSITY

Since the past history of a population is beyond our control, much attention has been turned to searching for an optimal choice of marker density used for dense SNP association study. While the physical distance between the QTL and SNP markers are not the sole factor that influence the strength of LD, it is still considered a main factor in most of the cases. Both theoretical studies and population genetic studies have tried to address this important question (12;33-35). Most of these studies used LD tests for population-based samples. Conclusions were somewhat different: some suggested to use a very high dense map with about 500,000 SNP markers spanning the whole genome; others suggested that strong LD can be extended up to 1*cM* and 30,000 SNPs will probably be enough for a whole-genome association study.

At a practical level, studies may only afford to have several hundreds to lower thousands SNPs either distributed along the whole genome or clustered around candidate regions. To get a practical sense of how power of LD test changes as a function of the genetic distance between the marker and the QTL locus, we calculated the power of our hypothetical studies with 500 nuclear families under all the models specified in Figure 1. We allow the genetic distance between QTL and the SNP marker to vary from 1*cM* to 10*cM*, which translate to a marker map density with between-marker distance of 2 to 40*cM*.

In Figure 3, we display the result for a moderately common allele frequency of 0.2. As the distance between the marker and the trait locus increases, the power to detect LD at the marker dissipates. With 500 families, under best scenario (equal frequencies and maximum LD strength), a map density of 10*cM* apart can still achieve a decent statistical power above 80% across the models tested. If genetic heterogeneity is a problem and only part of the sample is informative (say, 250 families), a much denser map (<2cM) would be needed (data not shown).

More interestingly, the genetic distance between the QTL and a SNP also affects how fast the power of LD test decays. In figure 4, we plot the loss of power of LD test as a function of genetic distance between QTL and a SNP marker, under various models. For moderately common to fairly common alleles (*p=0.2 and 0.4*), across the three types of models, the power is less sensitive to increased distance when the marker is close to QTL. For less frequent alleles (*p*=0.1), the power under dominant model is more sensitive to increased distance in the close vicinity of the QTL. Under multiplicative and additive models, however, it is much more sensitive at the distal sites (loss of >3% of power per 1*cM* increased). This should have implications to placing of SNP markers when some prior knowledge is available about the underlying genetic mode of the disease, especially a candidate gene.

10. QUANTITATIVE TRAITS AND COMBINED LINKAGE AND ASSOCIATION ANALYSIS

Many complex diseases manifest quantitative phenotypes that are measured on a continuous scale. Such quantitative traits generally contain more information than mere affection status of a disease. Many linkage methods are routinely used to map such quantitative traits. Several methods also have been developed to perform LD association studies using pedigree samples. Several score tests were proposed by Allison (36) and one based on *F*-ratio was shown to perform better under a variety models. Multiple regression procedure were used by George (37) and extended by Zhu and Elston (38) with careful considerations of power. Rabinowitz (39) considered a score test utilizing information of multiple siblings by using parental information to correct for correlation and admixture. The method was extended to 1-TDT by Sun (40) and colleagues to handle cases when only one parent genotype is available.

Alternatively, score tests based on likelihood of distribution of the phenotype employ a general regression treatment (20;41;42). Of special note are such methods using a variance components framework and can perform combined linkage and LD analysis using family-based data. Such analyses can be carried out with a sibpair design as suggested by Fulker. (43;44). Both the general regression and the variance components frameworks can entertain the addition of covariates and interaction terms into the general model. Using the variance components approach to analyze linkage and LD simultaneously, however, can enhance the true positive signal on a background around the QTL. It is because that the linkage effect in the combined model will dissipate quickly as markers move away from the QTL site. When compared the results to a pure linkage model, an elevated bump around the true QTL will appear in the plot of test scores, which leads to a refined resolution of mapping (Figure 7).

We demonstrated such effect in an earlier simulation study (45). Three simulated samples were used: Study A with a sample of 120 nuclear families each with 5 offspring, Study B with 600 nuclear families each with 2 offspring and Study C with 1000 families each with 2 offspring (total number of subjects are 600, 1200, and 2000). We simulated positive LD signal for Study A and B, and no LD effect for Study C. The QTL was simulated

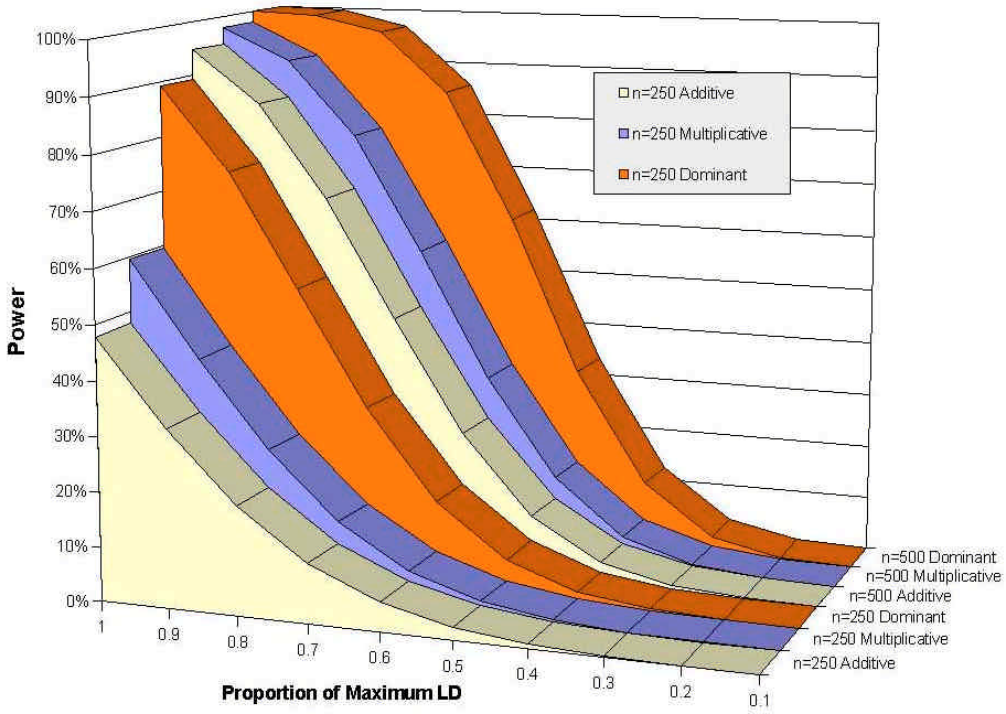


Figure 2. Tolerance of weakened LD strength in terms of TDT power loss. A reasonably large study with a sample size of 500 can tolerate a sizable reduction in LD strength (25% of D_{max}) while maintaining an acceptable power of ~70% across all three types of models.

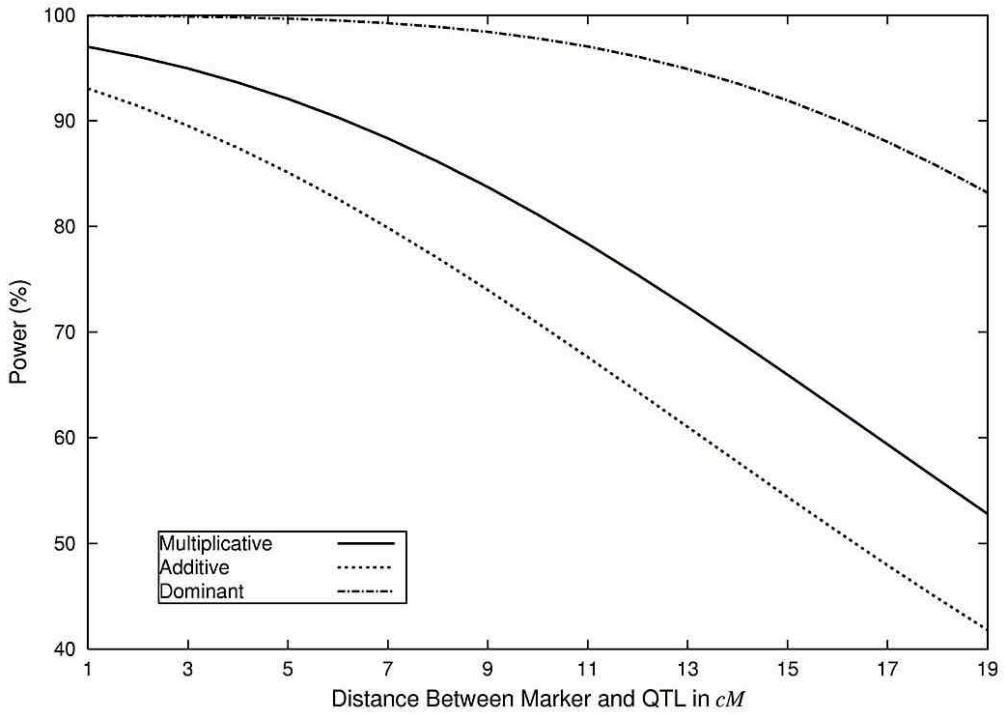


Figure 3. Effect of SNP markers density on power of TDT. Power displayed here were calculated based on a sample size of 500 and a disease allele frequency of p=0.2, and by assuming the marker allele in LD has the same frequency. The map density is reflected on the x-axis, where a distance value translate to a map density twice of that value, e.g., for a map density of 10cM apart the maximum distance between the QTL and closest marker will be 5cM. Power under dominant models is generally better than the other models, although for high frequent alleles power under multiplicative can be as high (data not shown)

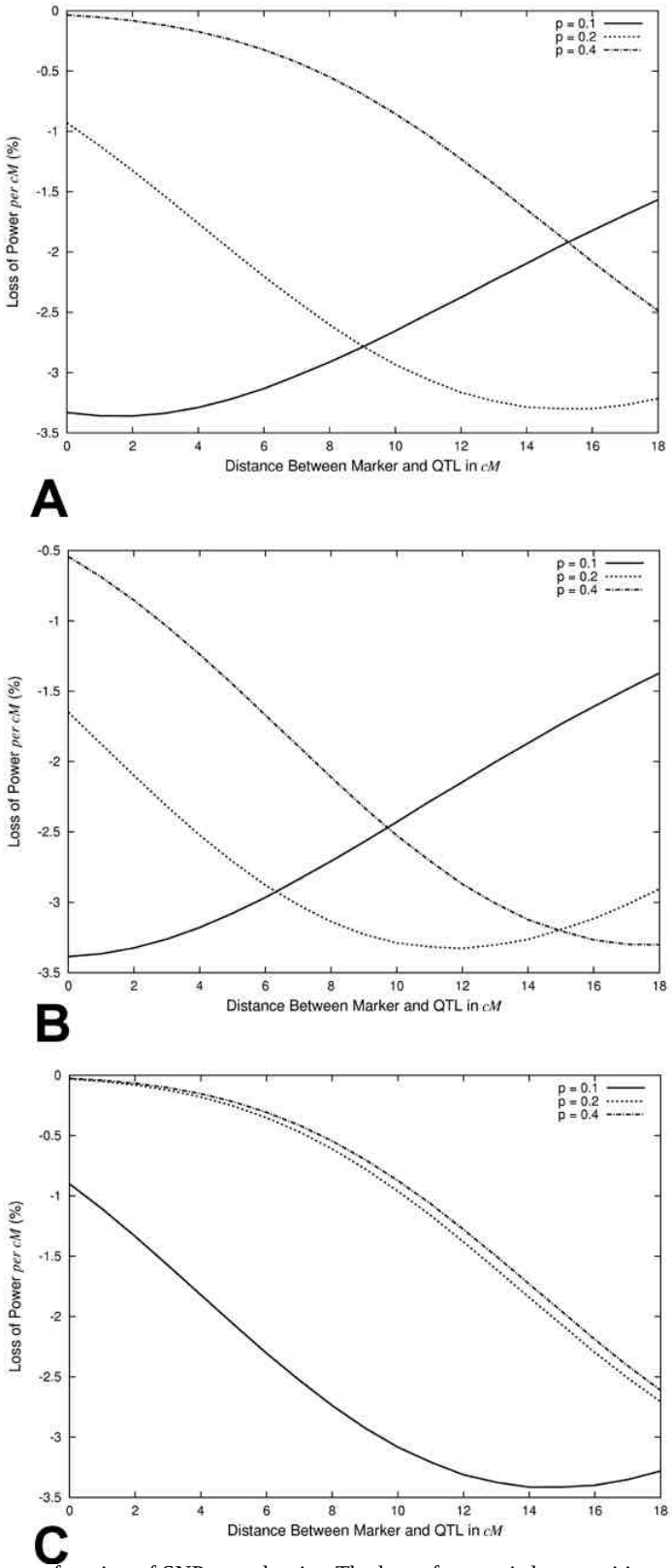


Figure 4. Rate of power loss as a function of SNP map density. The loss of power is less sensitive to increased distance when the marker is close to QTL for moderately common to fairly common alleles (*p>=0.2*). For less frequent alleles (*p*<=0.1), it can be more sensitive either in the close vicinity of the QTL or at distal sites, depending on the underlying model. Power under three types of models are plotted separately as shown.

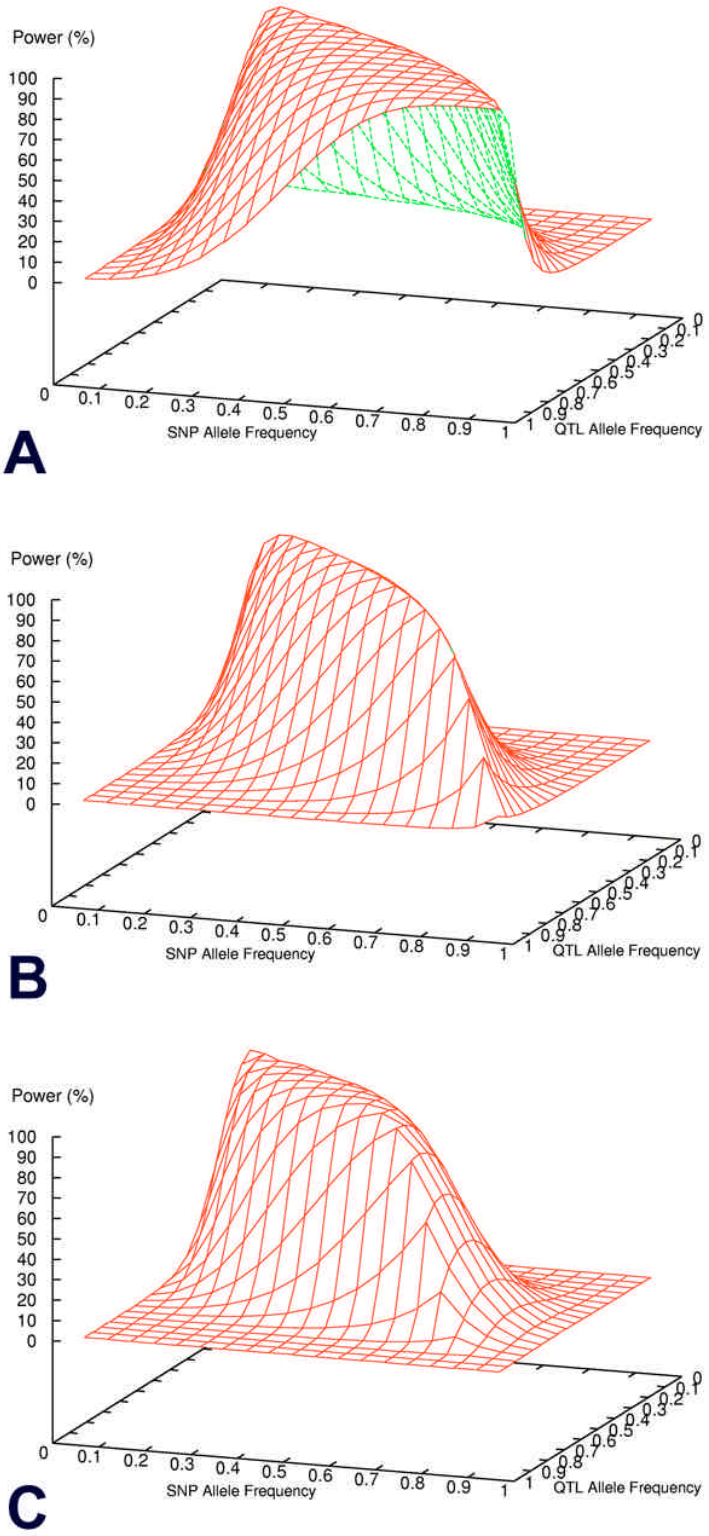


Figure 5. Power surface of TDT as a function of disparity between trait- and marker-allele frequencies. Power is calculated and plotted over combinations of trait- and marker-allele frequencies, ranging from 0.05 to 0.95 on a 0.05-grid, based on the moderately large sample of 500 families. Varying shapes of the power surface in (a) to (c) indicate that the robustness of the LD association test in respect to the disparity of allele frequencies is largely model dependent.

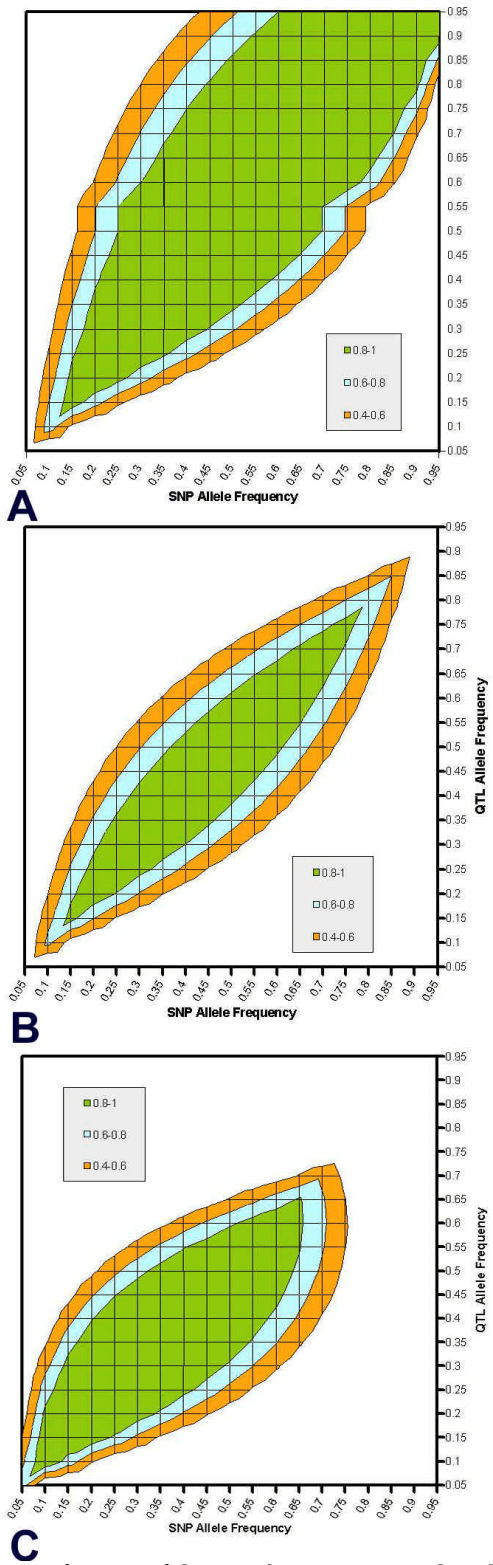


Figure 6. Contour plot of power surface as function of disparity between trait- and marker allele frequencies. Contour plots of the power surface displayed in the previous figure. The model dependence of the robustness of the LD test becomes more apparent. Multiplicative models are most tolerant to possible disparity between trait- and marker- allele frequencies (p and q). For a given significant level of 0.0009, areas colored green represent (p, q) pairs with a power of 80% and in light blue a power of 60-80%.

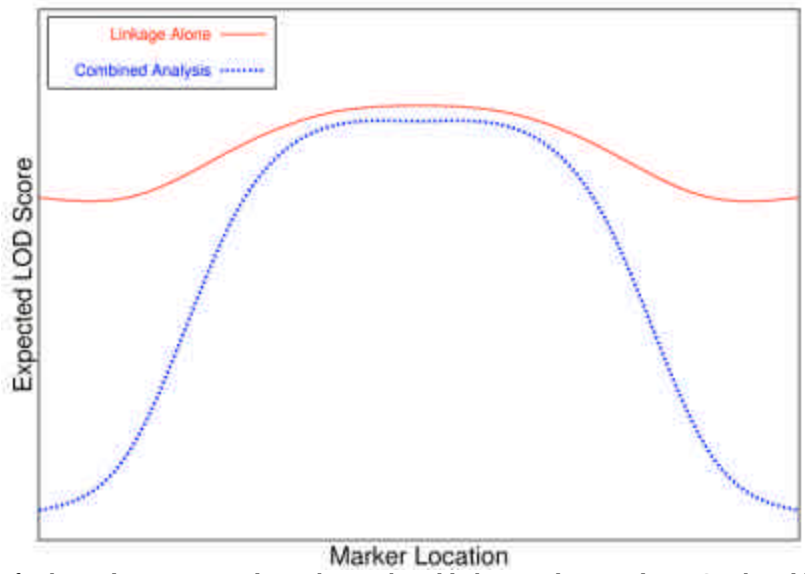


Figure 7. Illustration of enhanced mapping resolution by combined linkage and LD analysis. Combined linkage and association analysis produces test LOD scores that decays rapidly as the marker moves away from the trait QTL and gives enhanced or comparable test score near the trait QTL compared to linkage analysis alone. This may lead to improved resolution for fine mapping of disease genes.

Table 2. Summary of a simulation study of the utility to reduce false positives by combined linkage and LD analysis

	Linkage analysis only (p-value)	**Combined analysis of Linkage & LD (p-value)**	
D=0.08	7.09 (0.008)	A: N=120 K=5	59.56 (< 0.1e-6)
	4.84 (0.029)	B: N=600 K=2	94.14 (< 0.1e-6)
D=0	6.89 (0.009)	C: N=1000 K=2	3.86 (0.2770)

using an additive model with a heritability of 20% and a disease allele frequency of 0.125. An equifrequent marker with an allele frequency of 0.125 is simulated for $D = 0.08$ for sample A and B, and for $D = 0$ for sample C. Analysis was carried out using Fulker's method and a modified version of the SEGPATH program which performs likelihood tests on general pedigrees using a wide variety models (46). The result is summarized in Table 2. We see that the large sample size of Study C led to a moderately significant false positive using linkage analysis only. When applied with the combined linkage and LD analysis (in Study A and B), false positive rate is much reduced while the true signal enhanced.

Recently, Abecasis and colleagues extended Fulker's method to accommodate more general models and multi-generation pedigrees (47;48). They showed that with a locus specific heritability of 20% and parent genotypes available, a sample of 360 families with 3 sibs per family can achieve a power above 70%, even when 25% of LD strength is lost. Further enhancement of power may be achieved by selective sampling schemes similar to that applied in the context of linkage analysis (36;49;50).

11. DISCUSSION

Genetic association studies are all based on the principle of "guilty by association", namely, one identifies excessive correlation of a particular genetic variant or variants (alleles) and a phenotype of interest. Association studies based on the linkage disequilibrium between closely linked disease and marker alleles coupled on the same haplotype can lead to a more efficient study and, perhaps hold the key to the analysis of complex diseases. In a sole LD mapping design, implicit assumptions are made about the sample population. Therefore, possible misspecification introduces considerable instability of observed LD strength around disease loci. Such variation becomes intractable because of the unknown past events in the population history. Family-based designs of LD association analysis can reduce the influence of some of these factors such as population admixture.

We reviewed methods currently available in the literature to perform family-based LD association analysis, and investigated the effect of several design factors on the efficiency of such a study. We did this from a practitioner's point of view and selected the TDT platform for our power calculation. Our analysis is by no means exhaustive and should be taken in the context specified. By testing a wide spectrum of models, we showed that the disease allele has to be frequent enough for the LD association studies to be fruitful. We also showed that when there is no way to tell if a SNP marker frequency is equal to that of the disease allele, one can perform analysis to get the range of frequencies under a given model for a desired power. The density of the SNP marker map is a tricky and debatable topic. We want to stress that it is perhaps more important to know which SNPs one uses, rather than how far they are placed apart. For example, it is known that when closely placed markers are in strong LD among themselves, the power to detect the disease locus will be compromised. Also, whether a SNP can result in a nonsynonymous change (coding SNPs) may have biological implications

and should be included even if some conditions such as common allele frequency were not satisfied.

There are many practical issues that we did not address here. These include the many extensions of TDT for using multi-allelic markers (42;51;52). We did not discuss in detail of how to handle various kinds of missing information in family-based LD association studies. Both score test extension of TDT and regression based methods have been developed to handle the missing information (13;48;49;53;54). We also did not address the important issue of multiple testing. This has become the bottleneck of global genomic analysis, be it traditional genome-wide linkage scan or the more recent genomic LD association analysis using highly dense SNP markers. Instead of applying simple adjustment such as Bonferroni correction, more sophisticated permutation or bootstrapping procedures have been used by many to derive more practical empirical significance levels (55). Alternatively, a sequential multiple-decision procedure (SMDP) can be applied as proposed by Province (56) to solve multiple testing problems in genome-wide linkage scans.

The multiple testing problem is closely related to the problem of whole-genome LD scanning versus candidate gene scanning. The whole-genome scanning approach has many attractive aspects, such as the global assessment of LD strength and highest resolution for pinpointing the location of disease variant. However, the time may not be ripe yet for such a grand design. The large number of potential false-positives resulting from a whole-genome high-density scan may require either prohibitively large sample sizes (57) or huge cost in chasing false leads. Apparently, additional information is needed to help reduce the false positive findings. One solution is the candidate gene SNP scanning approach. Instead of an unconditional scanning of the whole genome, dense SNP markers are placed in clusters around candidate regions implicated by known positional significance (e.g., by linkage analysis) or physiological function (15). Such a design allows simultaneous analysis of association of multiple tightly linked markers in a small region to the QTL by haplotype analysis (51;58). Further investigation of such a design in the context of dense SNP scans at a large number of candidate genes is certainly warranted. Recently, a public effort has been launched to devise a haplotype map to identify segments of the chromosome that have virtually no recombinant activities so only flanking SNPs of such segments will be necessary for scanning. Success of this project will substantially reduce the necessary number of markers (SNPs) for genome-wide LD scanning. By scanning candidate genes this way by haplotype analysis, not only one can pinpoint a disease mutation site with high resolution but also can study collective actions of multiple variants within and across candidate genes that can lead to the etiology of the complex disease (59).

12. ACKNOWLEDGEMENT

We thank Dr. HW Deng and WM Chen for sharing their computer program *TDT-PC*. This work is supported partly by grants GM28719 and GM63340 from the National Institute of General Medical Sciences, NIH.

13. REFERENCES

1. Rao, D. C. and Michael A. Province. Genetic Dissection of Complex Traits. Adv Genet 42. 2001. San Diego, Academic Press.
2. Risch N. and K. Merikangas: The future of genetic studies of complex human diseases. *Science* 273, 1516-7 (1996)
3. Jorde L. B.: Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* 56, 11-4 (1995)
4. Feder J. N., A. Gnirke, W. Thomas, Z. Tsuchihashi and others: A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13, 399-408 (1996)
5. Hastbacka J., A. de la Chapelle, M. M. Mahtani, G. Clines, M. P. Reeve-Daly, M. Daly, B. A. Hamilton, K. Kusumi, B. Trivedi, A. Weaver, and a. l. et: The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* 78, 1073-87 (1994)
6. Jorde L. B.: Linkage disequilibrium and the search for complex disease genes. *Genome Res* 10, 1435-44 (2000)
7. Graham J. and E. A. Thompson: Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am J Hum Genet* 63, 1517-30 (1998)
8. Hastbacka J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver, and E. Lander: Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* 2, 204-11 (1992)
9. Kaplan N. L., W. G. Hill, and B. S. Weir: Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet* 56, 18-32 (1995)
10. Rannala B. and M. Slatkin: Likelihood analysis of disequilibrium mapping, and related problems. *Am J Hum Genet* 62, 459-73 (1998)
11. Xiong M. and S. W. Guo: Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *Am J Hum Genet* 60, 1513-31 (1997)
12. Collins A., C. Lonjou, and N. E. Morton: Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci U S A* 96, 15173-7 (1999)
13. Teng J. and N. Risch: The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res* 9, 234-41 (1999)
14. Ott, Jurg. Analysis of Human Genetic Linkage. 91. Baltimore, The Johns Hopkins University Press.
15. Gu C. and D. C. Rao: Optimum study designs. *Adv Genet* 42, 439-57 (2001)
16. Terwilliger J. D.: On the resolution and feasibility of genome scanning approaches. *Adv Genet* 42, 351-91 (2001)
17. Falk C. T. and P. Rubinstein: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 51 (Pt 3), 227-33 (1987)
18. Zhao H.: Family-based association studies. *Stat Methods Med Res* 9, 563-87 (2000)
19. Chen W. M. and H. W. Deng: A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genet Epidemiol* 21, 53-67 (2001)
20. Lunetta K. L., S. V. Faraone, J. Biederman, and N. M.

Laird: Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet* 66, 605-14 (2000)
21. Risch N. and J. Teng: The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling . *Genome Res* 8, 1273-88 (1998)
22. Spielman R. S., R. E. McGinnis, and W. J. Ewens: The transmission/disequilibrium test detects cosegregation and linkage. *Am J Hum Genet* 54, 559-60; discussion 560-3 (1994)
23. Kelsey, J. L. Thompson W. D. Evans A. S. Methods in Observational Epidemiology. 86. New York, Oxford University Press .
24. Chapman N. H. and E. M. Wijsman: Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am J Hum Genet* 63, 1872-85 (1998)
25. Devlin B. and N. Risch: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311-22 (1995)
26. Bengtsson B. O. and G. Thomson: Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* 18, 356-63 (1981)
27. Hill W. G. and B. S. Weir: Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* 54, 705-14 (1994)
28. Lewontin R. C.: The interaction of selection and linkage. I. Gneral considerations; heterotic models. *Genetics* 49, 49-67 (1964)
29. Olson J. M. and E. M. Wijsman: Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus. *Am J Hum Genet* 55, 574-80 (1994)
30. Guo S. W.: Linkage disequilibrium measures for fine-scale mapping: a comparison. *Hum Hered* 47, 301-14 (1997)
31. Morton N. E. and A. Collins: Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A* 95, 11389-93 (1998)
32. Tu I. P. and A. S. Whittemore: Power of association and linkage tests when the disease alleles are unobserved. *Am J Hum Genet* 64, 641-9 (1999)
33. Kidd J. R., A. J. Pakstis, H. Zhao, R. B. Lu, F. E. Okonofua, A. Odunsi, E. Grigorenko, B. B. Tamir, J. Friedlaender, L. O. Schulz, J. Parnas, and K. K. Kidd: Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet* 66, 1882-99 (2000)
34. Nickerson D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. Stengard, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C. F. Sing: DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet* 19, 233-40 (1998)
35. Kruglyak L.: Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22, 139-44 (1999)
36. Allison D. B.: Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 60, 676-90 (1997)
37. George V., H. K. Tiwari, X. Zhu, and R. C. Elston: A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* 65, 236-45 (1999)
38. Zhu X. and R. C. Elston: Transmission/disequilibrium tests for quantitative traits. *Genet Epidemiol* 20, 57-74 (2001)
39. Rabinowitz D.: A transmission disequilibrium test for quantitative trait loci . *Hum Hered* 47, 342-50 (1997)
40. Sun F., W. D. Flanders, Q. Yang, and M. J. Khoury: Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am J Epidemiol* 150, 97-104 (1999)
41. Self S. G., G. Longton, K. J. Kopecky, and K. Y. Liang: On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* 47, 53-61 (1991)
42. Schaid D. J.: General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13, 423-49 (1996)
43. Fulker D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt: Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64, 259-67 (1999)
44. Sham P. C., S. S. Cherny, S. Purcell, and J. K. Hewitt: Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66, 1616-30 (2000)
45. Gu C. P. M. A. R. D. C.: Precision mapping of human QTLs by combined linkage/disequilibrium analysis. *Genetic Epidemiology* 17, 227 (1999)
46. Province M. A. and D. C. Rao: General purpose model and a computer program for combined segregation and path analysis (SEGPATH): automatically creating computer programs from symbolic language model specifications. *Genet Epidemiol* 12, 203-19 (1995)
47. Abecasis G. R., L. R. Cardon, and W. O. Cookson: A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66, 279-92 (2000)
48. Allison D. B., M. Heo, N. Kaplan, and E. R. Martin: Sibling-based tests of linkage and association for quantitative traits. *Am J Hum Genet* 64, 1754-63 (1999)
49. Abecasis G. R., W. O. Cookson, and L. R. Cardon: The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am J Hum Genet* 68, 1463-74 (2001)
50. Gu C., A. Todorov, and D. C. Rao: Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genet Epidemiol* 13, 513-33 (1996)
51. Clayton D. and H. Jones: Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 65, 1161-9 (1999)
52. Rice J. P., R. J. Neuman, S. L. Hoshaw, E. W. Daw, and C. Gu: TDT with covariates and genomic screens with mod scores: their behavior on simulated data. *Genet Epidemiol* 12, 659-64 (1995)
53. Martin E. R., S. A. Monks, L. L. Warren, and N. L. Kaplan: A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 67, 146-54 (2000)
54. Tu I. P., R. R. Balise, and A. S. Whittemore: Detection of disease genes by use of family data. II. Application to nuclear families. *Am J Hum Genet* 66, 1341-50 (2000)
55. Boehnke M. and C. D. Langefeld: Genetic association mapping based on discordant sib pairs: the discordant-

alleles test. *Am J Hum Genet* 62, 950-61 (1998)
56. Province M. A.: A single, sequential, genome-wide test to identify simultaneously all promising areas in a linkage scan. *Genet Epidemiol* 19, 301-22 (2000)
57. Abel L. and B. Muller-Myhsok: Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. *Am J Hum Genet* 63, 664-7 (1998)
58. Zhao H., S. Zhang, K. R. Merikangas, M. Trixler, D. B. Wildenauer, F. Sun, and K. K. Kidd: Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 67, 936-46 (2000)
59. Daly M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander: High-resolution haplotype structure in the human genome. *Nat Genet* 29, 229-32 (2001)

Key Words: Linkage Disequilibrium, Genetic Association, Snp, Complex Disease, Family-Based Sample, Power Analysis, Review

Send correspondence to: Dr. Chi C. Gu, Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 S. Euclid Avenue, St. Louis, MO63110, Tel: 314-362-3642, Fax: 314-362-2693, E-mail: gc@wubios.wustl.edu