PLISKA

STUDIA MATHEMATICA
BULGARICA

ПЛИСКА

БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

Pliska Stud. Math. Bulgar. **18** (2007), 293–314

PLISKA
STUDIA MATHEMATICA
BULGARICA

A TEST OF ASSOCIATION BETWEEN QUALITATIVE TRAIT AND A SET OF SNPS

Radoslav Nickolov Valentin Milanov

In this article, we propose a novel candidate-gene association test that utilizes a set of tightly linked single nucleotide polymorphisms (SNPs). This is a powerful likelihood ratio test based on Gibbs random field model. We use simulation studies to evaluate the type I error rate of our proposed test, and compare its power with that of other candidate-gene association tests. The simulation results show that our proposed test has correct type I error rate, and is more powerful than the other tests in most cases considered in our simulation studies.

1. Introduction

There are about $30,000$ genes in the human genome. With the recent report of the sequence that constitutes our genome and improving biological technology, we are now in a position to begin detecting the genes that predispose humans to complex diseases: cancer, diabetes, hypertension, obesity, etc. This represents a formidable challenge for modern science.

The search for genetic factors that influence the human diseases is based on studying the observed correlations between genetic polymorphisms (markers) and disease. The statistical methods for disease gene mapping are usually divided into two broad categories: linkage analysis methods and association methods (population- and family-based).

2000 *Mathematics Subject Classification:* 62P10, 92D10, 92D30, 62F03
Key words: Case-control study; Genotypes; Gibbs distribution; Likelihood ratio test.

These methods all are based on one biological phenomenon: recombination (crossing-over). It is exploited for determining of the closeness (genetic distance) between two loci. Loci which are close to each other will rarely be separated by a recombination; they are in linkage. In family pedigrees, recombinations may be seen directly. On the other hand, the consequences of recombinations in past generations can be observed, in population samples, in the form of linkage disequilibrium (LD, non-random association of alleles at adjacent loci), [25].

Linkage analysis methods, [24], test for a relationship between disease and alleles transmitted at a given marker within families. They utilize a relatively small number of polymorphic markers (microsatellites) throughout the human genome. With the help of linkage methods approximately, 1, 500 single-gene human diseases have been identified to date (Online Mendelian Inheritance in Man (OMIM); http://www.ncbi.nlm.nih.gov/Omim/mimstats.html). These disease genes are rare, with large effects and known modes of inheritance. However, the success of linkage analysis in detection of genetic factors for complex diseases has been limited. Complex diseases are common, with unknown modes of inheritance and arise as a result of multiple mechanisms: common alleles with small to moderate effects, rare alleles with moderate to large effects, complex gene-gene and gene-environmental interactions, etc.

Association methods utilize statistical association of genetic markers to traits, either because of the direct effect of the polymorphism on the phenotype (”direct” approach [29]), or due to linkage disequilibrium of a marker locus to a close disease locus (”indirect” approach [6]). It was suggested that association study methods may be more powerful than linkage analysis methods, [30]. For common complex diseases, strong hypotheses about the roles of specific variants are generally not available, so the indirect approach is preferred, [50, 20]. Indirect association methods employ a dense map of polymorphic markers to test candidate genes, scan all genes in candidate regions found by linkage analyses, and even scan all the genes throughout the genome, [3]. They usually utilize biallelic single nucleotide polymorphisms (SNPs) as markers. There are almost 10 million SNPs available in the public database dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) to date, [35, 37]. However, even though genome-wide association studies are feasible, the candidate-gene approach is still being preferred in practice, [4]. Known genetic markers in the candidate gene are genotyped, and their association to the disease is tested with statistical association analysis methods. This can be done in a case-control design, with a sample of affected individuals, or cases, and unaffected individuals, or controls.

Genetic linkage is the coinheritance of alleles of loci that reside close to each other in a chromosome. A chromosome passed on from one generation to the next is a mosaic of maternal and paternal chromosomes. The transition points are the points in which crossing over has taken place. The co-segregation of a disease with particular marker alleles is evidence of the close proximity of the disease gene. The genetic distance unit, the Morgan (M), is based on the observed number of recombination events between loci, [24]. Since recombination events can be recognized only on the basis of haplotypes (a haplotype is the specific set of alleles observed on a single chromosome) passed from parents to children, linkage analysis can be carried out solely in pedigrees.

The coinheritance of alleles on haplotypes of tightly linked loci leads to associations between these alleles in the population, known as linkage disequilibrium (LD). Unlike linkage, linkage disequilibrium can be detected in population samples. A pair of loci is said to be in linkage disequilibrium when, in a sample of individuals, their joint haplotype frequencies deviate from those expected under independence. Consider, as an example, two closely spaced loci 1 and 2 on a chromosome with alleles A, a, and B, b. Suppose $p\,(A)$, $p(a)$ and $p(B)$, $p(b)$ are the frequencies of these alleles in the population. At equilibrium, the frequency $p(AB)$ of the AB haplotype should be equal to the product, $p\,(A)\,p(B)$, of the allele frequencies of A and B, respectively. If this holds, then $p(Ab) = p(A)p(b)$, $p(aB) = p(a)p(B)$, and $p(ab) = p(a)p(b)$ as well. Any deviation from these values imply linkage disequilibrium.

Linkage disequilibrium can be generated by several different mechanisms such as random genetic drift, mutation, selection, and population admixture and stratification. On one hand, new mutations, selection, drift and population admixture constantly create new LD between nearby loci, and on the other hand, recombinations break it down over the generations, [33, 40].

The relative strength of these forces determines the overall level of the LD and the distances to which it extends. LD varies widely between regions of the genome and, in some part of the genome, between populations. Pairs of loci that are tens of kilobases apart might be in complete LD, whereas nearby pairs of loci from the same region might be in very weak LD. Despite the apparent complexity of observed patterns, recent studies have proposed that the genome is composed of blocks of DNA conserved as a group over long regions of chromosome, up to 0.1 cM in length, see for example [42]. It is believed that these findings will have large impacts on genetic association studies, [47]. A haplotype map describing the block structure has recently been constructed, [37, 38].

There are various measures that can be utilized to summarize the magnitude of LD between two loci (markers), see for example [19, 33, 43, 44, 42].

One of them is D'. In the notation of the example considered above, let $D = p(AB) - p\,(A)\,p(B)$. D is dependent on allele frequencies in the population. Its maximum and minimum values are $D_{\max} = \min(p\,(A)\,p(b), p(a)p(B))$ and $D_{\min} = \max(-p\,(A)\,p(B), -p(a)p(b))$. Then D' is defined as

$$D^{'} = \begin{cases} \frac{D}{D_{\max}}, & D > 0 \\ \\ \frac{D}{-D_{\min}}, & D < 0 \end{cases} .$$

A value of 0 means no disequilibrium. If $|D'| = 1$, only two or three of the four possible haplotypes are present. If $|D'| < 1$, all haplotypes are present.

Indirect genetic association studies rely on the extent of LD between neutral (without effect on the gene product) markers and disease susceptibility loci. One will be able to find markers in LD with a disease locus if the effect of the disease susceptibility locus on the observed phenotype is strong enough and if there exists a sufficient amount of LD between the disease locus and markers neighboring it. Among other factors that influence the outcome of an association study are the frequency of the disease allele(s), the frequency of the markers' allele(s) and the spacing of the markers.

In this article we propose a test of association between a set of tightly linked biallelic markers typed in a candidate gene and qualitative trait of interest (case/control status). We model the probability of observing a multi-marker genotype by Gibbs distribution. This leads to a model similar to that of Potts (with interaction terms only) in statistical mechanics. Based on it, we construct a likelihood ratio test, called Potts Likelihood Ratio (PLR).

Using simulations we study the type I error rate and the power of our proposed test PLR. We compare them, under various disease models, with those of two well known and widely used candidate-gene association tests: Pearson's χ^2 goodness-of-fit test (GOF), [8, 1, 33, 43], and the generalized two-sample Hotelling's T^2 test, [41]. We use Monte Carlo permutation method to evaluate the p-values of the above three tests. Our simulation results show that the PLR has correct type I error rate. The power comparisons are performed under two scenarios, which include three one-locus disease models, and high LD haplotype structure in one candidate gene. The results show that the PLR is the most powerful test among the three tests in most cases considered. Therefore, PLR might be potentially useful for association studies of candidate genes with high LD haplotype structure.

2. Methods

2.1. Potts Likelihood Ratio Test

Consider a candidate gene. To test its association with a trait of interest, n unrelated individuals are sampled: n_1 of them affected called cases, and n_2 normal called controls. We assume that the underlying population is either homogeneous or that the cases and the controls are properly matched for race, ethnicity, etc., to avoid spurious association due to population substructure and (or) admixture, [33, 40]. Let the subjects be genotyped at a set $S = \{1, 2, \ldots, m\}$ of m biallelic markers (SNPs for example) in the candidate gene. We shall denote the alleles of each marker by 0 and 1. Thus, at each marker locus there are three possible genotypes 0/0, 0/1 and 1/1. Therefore, the observed i^{th} individual's multi-marker genotype ($i = 1, 2, \ldots, n$) can be written as a m-dimensional numerical vector $x^i = (x^i_1, x^i_2, \ldots, x^i_m)$, where the genotype x^i_s at each marker is coded as

$$x^i_s = \begin{cases} 0, & \text{if the genotype is} \quad 0/0 \\ 1, & \text{if the genotype is} \quad 0/1 \\ 2, & \text{if the genotype is} \quad 1/1 \end{cases} .$$

This set $S = \{1, 2, \ldots, m\}$ is actually a one dimensional lattice, and the markers in S could be considered as sites in that lattice. Then a multi-marker genotype $x = (x_1, x_2, \ldots, x_m)$ is the configuration of the observed values of a family of discrete random variables $\mathbf{X} = \{X_s\}_{s \in S}$ indexed by S. The random variables in $\mathbf{X}$ take values in the set $\{0, 1, 2\}$.

We propose to model the probability of observing a multi-marker genotype with the Gibbs distribution

$$P(x) = \frac{1}{Z_m} \exp(-H(x)),$$

where

$$Z_m = \sum_x \exp(-H(x))$$

is the partition function, and $H(x)$ is the energy. The summation in Z_m is over all 3^m possible multi-marker genotypes. The origins of the Gibbs distribution can be found in physics and statistical mechanics literature (there, it is also known as Boltzmann distribution), where it is used for modeling the equilibrium states of large physical systems, [13]. The Gibbs distribution was successfully applied by Majewski, Li and Ott ([22]) in the context of the multipoint affected-sib-pair linkage analysis in genetics.

The energy $H(x)$ may have different forms. For example, in the 3-states Potts model, [45, 21], the energy is

$$H(x) = -J \sum_{s=1}^{m-1} \delta(x_s, x_{s+1}) - h \sum_{s} x_s,$$

where $\delta(x_s, x_{s+1})$ is the Kronecker $\delta-$function giving value 1 when $x_s = x_{s+1}$ and 0 otherwise. J is the interaction strength (one and the same for all neighboring pairs x_s, x_{s+1}) and h is the external magnetic field. Another model is

$$H(x) = -\sum_{s=1}^{m-1} j(s)\delta(x_s, x_{s+1}) - \sum_{s=1}^{m} h(s)x_s,$$

where $j(s)$ is the interaction strength between x_s, x_{s+1} and $h(s)$ is the external field acting on x_s.

We note that in the first model above the external field h is only one, whereas in the second model the external fields $h(s)$ are local, one for each site s. If $\sum_{s=1}^{m} h(s)x_s$ is not present in $H(x)$ the model is called interactions only model.

Versions of the Potts model (set in a Bayesian framework) have been applied in epidemiology for spatial disease mapping [15], and in statistical genetics for spatial modeling of haplotype associations [39].

The above models, as they are, are not quite appropriate for modeling a set of markers in a gene because of the following reasons. First, the interaction strength parameter J is one and the same for all pairs of adjacent sites. However, if the interaction between two sites is to reflect the level of linkage disequilibrium between the markers, which may vary considerably from pair to pair, then we might lose information. Also, including external fields in $H(x)$ may increase the information, but will also increase the number of parameters. Second, the Kronecker's delta function in $H(x)$ gives value 1 when the genotypes at two adjacent markers are the same type, and 0 otherwise, which again will poorly capture the information carried by the nine possible two-marker genotypes at these markers.

Here, we let the energy $H(x)$ be similar to that in the one dimensional nearest-neighbor, interactions only, Potts model (see for example [45, 21]),

$$H(x) = -\sum_{s=1}^{m-1} j(s)G(x_s, x_{s+1}),$$

where the parameter $j(s)$ represents the interaction (correlation, linkage disequilibrium) between the markers s and $s+1$, and

$$G(x_s, x_{s+1}) = \begin{cases} 0, & \text{if the genotype at } (s, s+1) \text{ is} \quad 0\ 0 \\ 1, & \qquad - //- \qquad\qquad 1\ 0 \\ 2, & \qquad - //- \qquad\qquad 2\ 0 \\ 3, & \qquad - //- \qquad\qquad 0\ 1 \\ 4, & \qquad - //- \qquad\qquad 1\ 1 \\ 5, & \qquad - //- \qquad\qquad 2\ 1 \\ 6, & \qquad - //- \qquad\qquad 0\ 2 \\ 7, & \qquad - //- \qquad\qquad 1\ 2 \\ 8, & \qquad - //- \qquad\qquad 2\ 2 \end{cases} .$$

The sum in the partition function $Z_m = \sum_x \exp(-H(x))$ runs over all 3^m possible multi-marker genotypes. Hence, the number of terms in the summation increases exponentially with the number of markers, and the calculation may be very time intensive. Since the energy of our model do not include more than nearest-neighbor interactions, the partition function can be calculated recursively, in linear time, using the algorithm in [28] (see also [22]).

When a disease-associated mutation arises on a single chromosome, alleles at the nearby linked markers are initially in complete linkage disequilibrium with the mutation, [24]. Over time, this linkage disequilibrium decays by recombination at a rate determined by the map distance between the disease mutation and the marker, [27]. However, in small regions such as a gene, even after many generations from the mutation, substantial linkage disequilibrium might still exist. Thus, the linkage disequilibrium structure for a set of markers within a disease-associated gene observed in a sample of cases is expected to be different from that exhibited in a sample of controls. In other words, we expect the interaction strength (represented by $j(s)$, $s = 1, 2, , \ldots, m-1$) between the markers in the set S to be different in cases and controls.

Thus, when we model a set of markers with the Gibbs distribution, testing the null hypothesis of no association between the candidate gene against the alternative hypothesis of association is equivalent to testing that the parameters of the model for cases are equal to the parameters for controls. The null hypothesis of no association between gene and disease is then equivalent to the hypothesis

$$H_0 : j_{\text{cases}}(s) = j_{\text{controls}}(s), \;\; s = 1, 2, \ldots, m-1.$$

We call the likelihood ratio statistic for testing H_0 versus its alternative

$$H_a : \text{at least one pair is different}$$

Potts Likelihood Ratio test statistics, and denote it by PLR:

$$PLR = -2(\ln L(\{\hat{j}_{\text{all}}(s)\}) - \ln L(\{\hat{j}_{\text{cases}}(s)\}) - \ln L(\{\hat{j}_{\text{controls}}(s)\})),$$

where $\hat{\jmath}_{\text{all}}(s)$, $\hat{\jmath}_{\text{cases}}(s)$ and $\hat{\jmath}_{\text{controls}}(s)$ are the maximum likelihood estimators of the parameters based on all the data (pooled sample), the cases only and the controls only, respectively. Under the null hypothesis, as the sample size $n \to \infty$, PLR has, approximately, a χ^2 distribution with $m-1$ degrees of freedom. The maximum likelihood estimators are found numerically by Powell's method, [26]. The same method was used by Majewski, Li and Ott in [22].

To assess the statistical significance of an association, one can utilize the χ^2 approximation. However, the use of asymptotic results may not be appropriate for at least two reasons. First, for the asymptotic distribution to hold, the sample size should be large. Second, the number of degrees of freedom increases with the number of markers typed in the candidate gene, making the tail of the χ^2 distribution thicker, thereby leading to eventual loss of power to detect association. Thus, the use of the empirical distribution of the test statistic under the null hypothesis is desirable. This situation is encountered often in applied statistics, and in genetic association studies in particular (see for example [10, 9, 46, 23, 48]). Since, in most cases, obtaining of the exact null distribution of a particular test statistic is computationally infeasible, one estimates it through Monte Carlo methods in which a finite number of random permutations are performed, [14, 9]. We estimate the empirical p-values using the following randomization-permutation strategy, [9, 23]. Suppose that the observed value of the test statistic under consideration is t_0. Then, we 1. randomly shuffle the case and control status among the individuals; 2. calculate the value t_i of the test statistic using the permuted data; 3. repeat steps 1 and 2 N times; and 4. estimate the empirical p-value with the proportion of times the value of the test statistic calculated from the permuted data set is more extreme than that for the observed data: $\hat{p} = \frac{\#\{t_i > t_0\}}{N}$. In our studies we use $N = 1,000$ permutations to estimate the empirical p-values.

2.2. Other Tests Compared

In this study, we will compare the performance of the Potts Likelihood Ratio test with two well known and widely used candidate-gene association tests: Pearson's χ^2 goodness-of-fit test (GOF), [8, 1, 33, 43], and the generalized two-sample Hotelling's T^2 test, [41, 12].

Suppose that a set of m tightly linked biallelic markers (SNPs) are typed in a sample of cases and controls. There are 3^m possible multi-marker genotypes at these m markers. Suppose that L of them, $L \leq 3^m$, are present in the pooled sample of cases and controls. Denote these multi-marker genotypes by $g_1, g_2, \ldots, g_L$. The resulting data can be written in a $2 \times L$ contingency table. The statistical model underlying such data is that of two samples drawn from multinomial distributions. Pearson's χ^2 goodness-of-fit statistic provides a standard asymptotic test of the null hypothesis of no association in the $2 \times L$ contingency table, that is, the equality of the distributions of the multi-marker genotype frequencies in cases and controls. When the sample size is large, under the null hypothesis, GOF test statistic is asymptotically distributed as a χ^2 distribution with $L-1$ degree of freedom. The main drawback of GOF in its application to multi-marker genotype data is that the number of degrees of freedom increases with the increase of the number of markers considered. Hence, it loses power. Besides, asymptotic approximations may be rather poor when there is a small number of data in some cells.

Recently Xiong et al., [41], proposed the generalized T^2 statistic for case-control association studies of complex traits that simultaneously utilizes multiple biallelic (SNP) markers. This statistic is a corollary to that originally developed for multivariate analysis, [17], and is known in this context as two-sample Hotelling's T^2 statistic, see for example [14, 2]. Under the null hypothesis of no linkage disequilibrium between any marker in the set and a disease locus, the covariance matrix of the indicator variables for the marker genotypes of the cases, and the covariance matrix of indicator variables for the controls are equal. Hence, when the sample size is large enough, under the null hypothesis, T^2 is asymptotically distributed as a χ^2 distribution with m degrees of freedom, [2]. The generalized T^2 statistic is a promising statistic for association studies of complex diseases. It has been generalized, [11, 4], for use in high-resolution linkage disequilibrium mapping based on haplotype maps, microsatellite marker maps, and for family data [12]. In [4] was shown that the generalized T^2statistic is optimal or near optimal (in the case of one disease susceptibility locus, strong LD between markers and additive disease model) among wide class of statistics, in the generalized linear models framework, see also [5].

We evaluate the p-values of GOF and T^2 with the same randomization-permutation strategy employed for the PLR test.

3. Simulations

We use simulation studies to evaluate the performance of the proposed test and to compare its power with two other tests. We perform the simulation studies based on the haplotype frequencies in a gene taken from a study aiming to understand the impact of the apolipoprotein B (ApoB) polymorphism on cholesterol[16]. In this study, five intragenic single nucleotide polymorphisms were typed in 121 French nuclear families. ApoB gene has 10 haplotypes. The haplotype structure, haplotype frequencies, and SNP frequencies from the study are given in Table 1.

			SNP			
Haplotype	1	2	3	4	5	Frequency
1	1	1	1	1	1	0.180
2	0	0	1	1	1	0.214
3	1	1	0	1	0	0.194
4	0	0	1	0	1	0.100
5	1	1	0	1	1	0.277
6	0	1	0	1	1	0.006
7	0	0	0	1	1	0.014
8	1	1	1	0	1	0.006
9	1	1	0	0	1	0.004
10	0	1	1	1	1	0.005
Minor allele frequency	0.34	0.328	0.49	0.11	0.19	

Table 1: Haplotypes and haplotype frequencies of ApoB gene.

For each pair of SNPs, the haplotype frequencies from all SNPs were collapsed to obtain the four haplotype frequencies for the two SNPs under consideration, and the linkage disequilibrium measure D' was calculated. The linkage disequilibrium values for each pair of SNPs are given in Table 2. The five loci are in tight linkage disequilibrium, suggesting that recombination in this region is rare, and the haplotype structure is similar to that of a haplotype block. That was the reason for choosing ApoB gene for our simulation studies. Apo B gene was also used in other theoretical studies [32, 18, 34]

SNP	2	3	4	5
1	1.00	-0.88	0.86	-1.00
2		-0.91	0.87	-1.00
3			-0.93	1.00
4				-1.00

Table 2: Linkage diseequilibrium value (D') for each pair of SNPs of the ApoB gene.

In our simulations the trait values (case/control status) depend on the genotypes at disease susceptibility loci and disease models. Let D and d denote the two alleles at a disease susceptibility locus with frequencies p_D and $p_d = 1 - p_D$, respectively. Denote the penetrance of the genotypes DD, Dd and dd as follows:

$f_{DD} = P(\text{case}|DD)$, $f_{Dd} = P(\text{case}|Dd)$, and $f_{dd} = P(\text{case}|dd)$. Let D be the high risk allele at the disease locus. Then $f_{DD} \geq f_{Dd} \geq f_{dd}$. We consider three one-locus disease models namely the dominant ($f_{DD} = f_{Dd} \neq f_{dd}$), the recessive ($f_{DD} \neq f_{Dd} = f_{dd}$), and the additive ($f_{Dd} = 0.5(f_{DD} + f_{dd})$) models (see for example [33, 40, 49]). We choose one SNP as disease susceptibility SNP and choose the minor allele as high-risk allele. We consider two scenarios:

- Scenario I. The second SNP is the disease susceptibility locus itself. The mutant allele D is the allele denoted with 0 in Table 1, and its relative frequency 0.328 is taken as p_D.
- Scenario II. The fourth SNP is the disease susceptibility locus itself. The mutant allele D is the allele denoted with 0 in Table 1, and its relative frequency 0.11 is taken as p_D.

Once the sample of cases and controls is simulated (as explained bellow), we calculate the values of the statistic under consideration for the set of all five SNPs. Then, we randomly choose one SNP and remove it from the data. Next, we calculate the statistic with the set of four SNPs left. Again, we reduce the number of SNPs to three and recalculate the statistic. In such a way, we not only have the possibility to compare the performance of the tests for different number of SNPs typed, but also make the comparison more fair. For Scenario I, we first exclude the fifth SNP and use SNPs 1,2, 3 and 4. Then we exclude the third SNP, and use SNPs 1,2 and 4. So, the sets are $\{1, 2, 3, 4, 5\}$, $\{1, 2, 3, 4\}$ and $\{1, 2, 4\}$. For Scenario II, we exclude the first SNP and then the third. So, the set of SNPs are $\{1, 2, 3, 4, 5\}$, $\{2, 3, 4, 5\}$ and $\{2, 4, 5\}$.

3.1. Data sets for assessing the power

To assess the power, in both scenarios considered, we set the disease prevalence to be 0.1 and vary the relative risk (f_{DD}/f_{dd}) of genotypes DD to dd from 2 to 5, in increments of 1.5. Given the prevalence, relative risk, disease allele frequency p_D and the disease model (dominant, recessive or additive), we calculate the probability of the genotypes at the disease susceptibility locus, under the condition that the individual is affected (case) or normal (control), assuming Hardy-Weinberg equilibrium. The probabilities used in our simulations are given in Table 3.

Next, in order to generate the haplotypes of an individual, we first generate the individual's genotype at the disease locus, given the affectation status, using the above conditional probabilities. If the individual's genotype is DD, we choose

Probability	Relative Risk	Scenario I			Scenario II		
		Dom.	Rec.	Add.	Dom.	Rece.	Add.
$P(DD\|case)$	2	0.139	0.194	0.162	0.020	0.024	0.022
	3.5	0.159	0.297	0.207	0.028	0.041	0.033
	5	0.168	0.376	0.232	0.033	0.058	0.042
$P(Dd\|case)$	2	0.569	0.398	0.498	0.324	0.193	0.265
	3.5	0.651	0.347	0.545	0.451	0.190	0.346
	5	0.690	0.308	0.572	0.535	0.187	0.408
$P(dd\|case)$	2	0.292	0.408	0.340	0.656	0.783	0.714
	3.5	0.190	0.356	0.248	0.521	0.769	0.621
	5	0.141	0.316	0.195	0.432	0.756	0.550
$P(DD\|control)$	2	0.104	0.098	0.102	0.011	0.011	0.011
	3.5	0.102	0.087	0.097	0.010	0.009	0.010
	5	0.101	0.078	0.094	0.010	0.007	0.009
$P(Dd\|control)$	2	0.427	0.446	0.434	0.182	0.196	0.188
	3.5	0.418	0.451	0.429	0.168	0.196	0.179
	5	0.413	0.456	0.426	0.158	0.197	0.172
$P(dd\|control)$	2	0.469	0.456	0.464	0.807	0.793	0.801
	3.5	0.481	0.462	0.474	0.822	0.795	0.811
	5	0.486	0.467	0.480	0.832	0.796	0.819

Table 3: Conditional probabilities at the disease locus used in the simulations, under dominant (Dom.), recessive (Rec.), and additive (Add.) models.

two haplotypes with the disease mutation. We do this, using the distribution of the haplotype frequencies from Table 1. If there are many haplotypes bearing the mutation, we sample one of them according to their frequencies. If the individual's genotype is *Dd*, we sample one haplotype with, and one haplotype without the disease mutation, according to the haplotype frequencies in Table 1. If the individual's genotype is *dd*, we sample two haplotypes without the disease mutation, again according to the haplotype frequencies in Table 1. Proceeding in this way, we simulate the haplotypes for a sample of cases and controls. Finally, we convert the haplotype data into genotype data.

For each simulation scenario, we generate $1,000$ independent samples of 100 case and 100 controls, and for each sample, the p-values of the tests considered is estimated by $1,000$ permutations.

3.2. Data sets for assessing the type I error:

To assess the type I error rate, we generate the data under null hypothesis of no association between trait values and the multi-marker genotypes, as described above, by simply setting the relative risk equal to 1. For each simulation scenario, we generate $1,000$ samples of 100 case and 100 controls, and use $1,000$ permutations for each sample to evaluate the p-values of the tests.

4. Results

4.1. Type I Error Rates

The estimated type I error rates for the three tests (PLR, GOF, and T^2), under Scenario I and Scenario II,are given in Table 4. Two levels of statistical significance are considered: 0.05 and 0.01. For $1,000$ replicated samples, the standard errors for the type I error rate estimates were $\sqrt{(0.05 \times 0.095)/1,000} \approx 0.0069$ and $\sqrt{(0.01 \times 0.099)/1,000} \approx 0.00315$, for the nominal levels of 0.05 and 0.01, respectively. The 95% confidence intervals were $(0.0365, 0.0635)$ and $(0.0019, 0.0181)$, respectively. It is easy to see from Table 4 that the estimated type-I errors of PLR are similar to these of the other tests, and are not statistically significantly different from the nominal levels. However, even though all three tests have reasonable type I error rates they are slightly inflated. This might be improved by increasing the sample size and (or) the number of permutations.

4.2. Power Comparisons

To compare the power of the three tests, we generate data by using the haplotype frequencies in the ApoB gene considered in our simulation. For each of the two scenarios, we consider three disease models. For each scenario and each disease

	Number of	Significance level 0.01			Significance level 0.05		
Scenario	markers	*PLR*	*GOF*	T^2	*PLR*	*GOF*	T^2
I	5	0.016	0.013	0.012	0.064	0.059	0.063
	4	0.014	0.014	0.017	0.060	0.056	0.060
	3	0.014	0.008	0.007	0.054	0.056	0.066
II	5	0.021	0.016	0.014	0.064	0.072	0.063
	4	0.012	0.018	0.015	0.057	0.066	0.060
	3	0.018	0.014	0.016	0.072	0.059	0.071

Table 4: Type I error rate comparisons of the three tests in simulations of 100 cases and 100 controls.

model, we consider three different values, 2, 3.5,and 5, of relative risks, and 3, 4, or 5 markers in the gene. We set the significance level to be 0.01, and 0.05. Thus, we perform the simulations under $2 \times 3 \times 3 \times 3 \times 2 = 108$ different cases. The results are summarized in Tables 5 and 6. Table 5 presents the power comparisons under the Scenario I, and Table 6 gives the power comparisons under the Scenario II.

Scenario I

As can be seen from Table 5, *PLR* is the most powerful test in all cases considered. It clearly outperforms T^2, especially in the case of 4 and 5 markers, under recessive and dominant models, and for low relative risk. Under the recessive model, five markers, and relative risk 3.5, the margin is the highest: 19%. Even under additive model *PLR* is more powerful than T^2 (margin up to 14%). In the case of recessive model T^2 has power less than that of *GOF*.

For 3 markers and relative risk 5, *GOF* has power close to that of *PLR*. For 4 and 5 markers the power of *GOF* is reduced due to the increase of the degrees of freedom.

Scenario II

Under this simulation scenario, the fourth marker is the disease susceptibility locus itself. Under the recessive model, its lower frequency, 0.11, leads to very high percentage of phenocopies (individuals who are affected without carrying a disease-causing allele) among the cases (see Table 3). This in turn leads to lower power of the tests. However, even in this case it can be seen from Table 6 that *PLR* is more powerful than the other two tests (margin up to 4%).

Under dominant and additive models (3 markers) the most powerful test is *PLR*, followed closely by T^2. In the case of 4 and 5 markers T^2 and *PLR* performed similarly without clear winner.

As in scenario I, our proposed test and T^2 are more powerful than the GOF test. Only in the case of 5 markers, dominant disease model, and relative risk 5, GOF has power comparable to that of PLR and T^2.

Comparing the power of the three tests shown in the two tables, we see that the PLR is the most powerful test among the three tests in all cases considered. Its power advantage is larger when the disease allele frequency is higher (Scenario I).

5. Discussion

Our proposed test targets the detection of population-level association between candidate gene and disease from a case-control sample. It is designed to exploit, simultaneously, set of tightly linked biallelic markers (SNPs), typed in the gene of interest. The goal is to optimally use the information on all markers. The proposed likelihood ratio test statistics, PLR, is built on a Gibbs random field model. Similar models have been used in many areas of science, such as statistical mechanics, spatial statistics, image analysis, biology, genetics, etc. for modeling interacting systems.

We evaluated the type I error rate and the relative power of the proposed test through simulations. In our simulation scenarios we assumed one disease susceptibility locus in the candidate gene of interest. The proposed test statistics had correct type I error rate, when permutation testing was employed.

We compared the power of our proposed test with the power of two others, the widely used Pearson's χ^2 goodness-of-fit test (GOF), and the generalized two sample Hotelling's T^2 test. We choose T^2, because it was shown in [4], that in certain cases, this test is optimal or near optimal, among wide class of statistics in the generalized linear models framework.

We found that our proposed test PLR is more powerful than GOF and T^2 in the cases considered. The difference in power increases with the number of markers, and is larger when the relative risk is low.

Our limited simulation studies were based on the haplotype structure of ApoB gene. We choose this gene, because all five SNP are in high linkage disequilibrium with each other mimicking haplotype block structure. We assumed that one of the genotyped markers is the disease locus itself. This is equivalent to the case in which one of the SNPs is in complete linkage disequilibrium with the unobserved disease locus. We expect our results to hold also in the cases when the disease locus is not genotyped and is in incomplete linkage disequilibrium with the observed SNPs, as long as the last are in a haplotype block. However,

Number of markers	Model	Relative Risk	Significance level 0.01			Significance level 0.05		
			PLR	GOF	T^2	PLR	GOF	T^2
5	Dom.	2	0.31	0.14	0.21	0.55	0.31	0.40
		3.5	0.81	0.58	0.70	0.93	0.78	0.86
		5	0.93	0.87	0.86	0.98	0.96	0.93
	Rec.	2	0.10	0.05	0.05	0.25	0.15	0.16
		3.5	0.62	0.40	0.43	0.83	0.66	0.65
		5	0.92	0.83	0.82	0.98	0.94	0.94
	Add.	2	0.22	0.06	0.11	0.42	0.17	0.26
		3.5	0.74	0.32	0.59	0.91	0.54	0.80
		5	0.94	0.64	0.86	0.98	0.83	0.94
4	Dom.	2	0.33	0.21	0.24	0.58	0.38	0.42
		3.5	0.83	0.71	0.74	0.94	0.87	0.88
		5	0.95	0.93	0.89	0.96	0.98	0.95
	Rec.	2	0.11	0.06	0.06	0.27	0.19	0.18
		3.5	0.65	0.52	0.49	0.84	0.73	0.68
		5	0.94	0.90	0.86	0.98	0.97	0.94
	Add.	2	0.22	0.08	0.14	0.43	0.21	0.30
		3.5	0.77	0.41	0.64	0.93	0.64	0.84
		5	0.95	0.76	0.89	0.99	0.89	0.95
3	Dom.	2	0.36	0.29	0.28	0.60	0.49	0.49
		3.5	0.87	0.83	0.80	0.96	0.93	0.91
		5	0.95	0.97	0.91	0.99	0.99	1.00
	Rec.	2	0.11	0.08	0.08	0.28	0.23	0.19
		3.5	0.67	0.68	0.54	0.85	0.86	0.73
		5	0.94	0.97	0.89	0.99	0.99	0.96
	Add.	2	0.23	0.11	0.18	0.44	0.27	0.33
		3.5	0.81	0.60	0.70	0.94	0.80	0.88
		5	0.96	0.88	0.91	0.99	0.95	0.97

Table 5: Power comparisons of the three tests in simulations based on ApoB gene; Scenario I. Sample size for cases and controls 100.

Number of markers	Model	Relative Risk	Significance level 0.01			Significance level 0.05		
			PLR	GOF	T^2	PLR	GOF	T^2
5	Dom.	2	0.20	0.07	0.22	0.43	0.25	0.43
		3.5	0.84	0.65	0.88	0.96	0.83	0.97
		5	0.98	0.95	0.99	1.00	0.99	1.00
	Rec.	2	0.01	0.02	0.01	0.05	0.06	0.06
		3.5	0.02	0.02	0.03	0.09	0.08	0.08
		5	0.04	0.03	0.03	0.15	0.11	0.11
	Add.	2	0.07	0.03	0.07	0.17	0.11	0.17
		3.5	0.40	0.17	0.40	0.64	0.36	0.65
		5	0.75	0.52	0.80	0.91	0.72	0.92
4	Dom.	2	0.23	0.10	0.22	0.46	0.25	0.45
		3.5	0.89	0.66	0.90	0.97	0.85	0.98
		5	0.99	0.95	1.00	1.00	0.99	1.00
	Rec.	2	0.01	0.02	0.01	0.05	0.05	0.05
		3.5	0.03	0.02	0.02	0.09	0.08	0.09
		5	0.04	0.03	0.04	0.16	0.12	0.12
	Add.	2	0.07	0.03	0.07	0.17	0.10	0.19
		3.5	0.44	0.17	0.44	0.70	0.37	0.67
		5	0.82	0.52	0.82	0.94	0.73	0.93
3	Dom.	2	0.32	0.15	0.27	0.54	0.33	0.47
		3.5	0.94	0.80	0.93	0.98	0.93	0.98
		5	1.00	0.99	1.00	1.00	1.00	1.00
	Rec.	2	0.01	0.01	0.01	0.06	0.05	0.06
		3.5	0.03	0.03	0.02	0.10	0.09	0.09
		5	0.07	0.05	0.04	0.18	0.16	0.15
	Add.	2	0.10	0.04	0.08	0.24	0.14	0.21
		3.5	0.56	0.25	0.49	0.76	0.47	0.73
		5	0.89	0.68	0.87	0.97	0.84	0.96

Table 6: Power comparisons of the three tests in simulations based on ApoB gene; Scenario II. Sample size for cases and controls 100.

to achieve adequate power one will need larger sample size. Future work is needed to assess the performance of our proposed test with respect to sample size and to candidate genes with different haplotype structures.

In conclusion, our proposed test seems a promising approach for case-control candidate-gene association studies, in which a number of markers in high linkage disequilibrium with each other (a haplotype block) are genotyped in the gene of interest.

Acknowledgements

The authors would like to thank Dr. Shuanglin Zhang for his positive and helpful comments.

REFERENCES

[1] Agresti, A. Categorical Data Analysis, John Wiley & Sons, Inc., 1990.

[2] Anderson, T. W. An Introduction to Multivariate Statistical Analysis, Second Edition, John Wiley & Sons, Inc., 1984.

[3] Balding, David, J. A Tutorial on Statistical Methods for Population Association Studies *Nature Reviews/Genetics*, **7** (2006), 781–791.

[4] Chapman, J. M., Cooper, J. D., Todd, J. A., Clayton, D. G. Detecting Disease Associations Due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the determinants of Statistical Power *Human Heredity* **56** (2003), 18–31.

[5] Clayton, D., Chapman, J., Cooper, J. The Use of Unphased Multilocus Genotype Data in Indirect Association Studies *Genetic Epidemiology* **27** (2004), 415–428.

[6] Collins, F. S., Guyer, M. S. and Chakravarti, A. Variations on A Theme: Cataloging Human DNA Sequence Variation *Science* **278** (1997), 1580–1581.

[7] Cressie, N. *Statistics for Spatial Data*, John Wiley & Sons, Inc., New York, 1991.

[8] Cressie, N., Read, T.R.C. Multinomial Goodness-of-Fit Tests *Journal of the Royal Statistical Society* Series B **46(3)** (1984).

[9] Fallin, D. Haplotype-Based Approaches for Genetic Case-Control Studies, dissertation, Western Reserve University, Cleveland, 2000.

[10] Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D. and Shork, N. Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: application to APOE Locus Variation and Alzhaimer's Disease *Genome Research*, **11** (2001), 143–151.

[11] Fan, R. and Knapp, M. Genome Association Studies of Complex Diseases by Case-Control Designs *Am. J. Hum. Genet.* **72** (2003), 850–868.

[12] Fan, R., Knapp, M., Wjst, M., Zhao, C., and Xiong, M. *Annals of Human Genetics* **69** (2005), 187–208.

[13] Gallavotti, G. Statistical Mechanics: a Short Treatise, Springer-Verlag Berlin Heidelberg, 1999.

[14] Good, P. Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses, Springer-Verlag New York, Inc., 1994.

[15] Green, P., Richardson, S. Hidden Markov Models and Disease Mapping *J. Am. Stat. Assoc.* **97** (2002), 1055–1070.

[16] Hallman, D. M., Visviks, S., Steinmetz and J., Boerwinkle, E. The Effect of Variation in the Apolipoprotein B Gene on Plasmid Lipid and Apolipoprotein B Levels. I. A Likelihood-Based Approach to Cladistic Analysis *Ann. Hum. Genet.* **58** (1994), 35–64.

[17] Hotelling, H. The Generalization of Student's Ratio *Ann Math Stat.* **2** (1931), 360–378.

[18] Jannot, A-S., Essioux, L. and Clerget-Darpoux, F. Association in multifactorial trait: how to deal with rare observations? *Human Heredity* **58** (2004), 73–81.

[19] Jorde, L. B. Linkage Disequilibrium and the Search for Complex Disease Genes *Genome Research* **10** (2000), 1435–1444.

[20] Kruglyak, L. Prospects for Whole-Genome Linkage Disequilibrium Mapping of Common Disease Genes *Nature Genetics* **22** (1999), 139–144.

[21] Liu, J. S. Monte Carlo Strategies in Scientific Computing, Springer-Verlag New York, 2001.

[22] Majewski, J., Li, H. and Ott, J. The Ising Model in Phisics and Satistical Genetics *Am. J. Hum. Genet.* **69** (2001), 853–862.

[23] Nettleton, D. and Doerge, R. W. Accounting for Variability in the Use of Permutation Testing to Detect Quantative Trait Loci", *Biometrics* **56** (2000), 52–58.

[24] Ott, J. Analysis of Human Genetic Linkage, Third edition. The John Hopkins Press, Baltimore, 1995.

[25] Ott, J. and Hoh, J. Statistical Approaches to Gene Mapping", *Am. J. Hum. Genet.* **67** (2000), 289–294.

[26] Press, W., Teukolsky, S. A., Vetterling, W. T., Flannery, B.P. Numerical Recipes in C, Cambridge University Press, New York, 1990.

[27] Rannala, B. and Slatkin, M. Methods for Multipoint Disease Mapping Using Linkage Disequilibrium. *Genetic Epidemiology* **19(1)** (2000), S71–S77.

[28] Reichl, L. E. A modern Course in Statistical Physics, University of Texas Press, Austin, TX, 1980.

[29] Risch, N. Searching for Genetic Determinants in The New Millenium *Nature* **405** (2000), 847–856.

[30] Risch, N. and Merikangas, K. The Future of Genetic Studies of Complex Human Diseases *Science* **273** (1996), 1516–1517.

[31] Sasieni, P. D. From Genotypes to Genes: Doubling the Sample Size *Biometrics* **53** (1997), 1253–1261.

[32] Seltman, H., Roeder, K. and Devlin, B. Transmission/Disequlibrium Test Meets Measured Haplotypes Analysis: Family-Based Association Analysis Guided by Evoliution of Haplotypes *Am. J. Hum. Genet.* **68** (2001), 1250–1263.

[33] Sham, P. Statistics in Human Genetics, Oxford University Press, Inc., New York, 1998.

[34] Sha, Q., Dong, J., Jiang, R., and Zhang, S. Tests of Association Between Quantative Traits and Haplotypes In A Reduced-Dimensional Space *Annals of Human Genetics* **69** (2005), 715–732.

[35] Sherry, S. T. et al. dbSNP: The NCBI Database of Genetic Variation *Nucleic Acids Res.* **29** (2001), 308–311.

[36] Stram, D., Pearce, C., Bretsky, P., Freedman, M., Nirschhorn, J., Altshuler, D., Kolonel, L., Henderson, B., Thomas, D. Modeling and E-M Estimation of Haplotype-Specific Relative Risks from Genotype Data for a Case-Control Study of Unrelated Individuals *Human Heredity* **55** (2003), 179–190.

[37] The International HapMap Consortium The International HapMap Project", *Nature* **426** (2003), 789–796

[38] The International HapMap Consortium. A haplotype map of the human genome *Nature* **437** (2005), 1299–1320.

[39] Thomas, D., Stram, D., Conti, C., Molitor, J. Bayesian Spatial Modeling of Haplotype Associations *Human Heredity* **56** (2003), 32–40.

[40] Thomas, D. Statistical Methods in Genetic Epidemiology, Oxford University Press, 2004.

[41] Xiong, M., Zhao, J. and Boerwinkle, E. Generalized T^2 Test for Genome Association Studies *Am. J. Hum. Genet.* **70** (2002), 1257–1268.

[42] Wall, J. D. and Pritchard, J. K. Haplotype Blocks and Linkage Disequilibrium in The Human Genome *NatureReviews/Genetics* **4** 587–597.

[43] Weir, B. S. Genetic Data Analysis II, Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, 1996.

[44] Weiss, K. M., Clark, A. G. Linkage Disequilibrium and The Mapping of Complex Human Traits *Trends in Genetics* **18(1)** (2002), 19–24.

[45] Wu, F. Y. The Potts Model *Reviews of Modern Physics* **54(1)** (1982), 235-268.

[46] Zaykin, D., Westfal, P., Young, S., Karnoub, M., Wagner, M., Ehm. M. Testing Association of Statistically Infered Haplotypes with Discrete and Continuous Traits in Sample of Unrelated Individuals *Human Heredity* **53** (2002), 79–91.

[47] Zhang, K., Calabrese, P., Nordborg, M. and Sun, F. Haplotype Block Structure and Its Applications to Association Studies: Power nad Study Designs *Am. J. of Hum Genet.* **71** (2002), 1386–1394.

[48] Zhao, J. H., Curtis, D., Sham, P. C. Model-Free Analysis and Permutation Tests for Allelic Associations *Human Heredity* **50** (2000), 133–139.

[49] Ziegler, A., König, I. R. A Statistical Approach to Genetic Epidemiology: Concepts and Applications, WILEY-VCH, Weinheim, 2006.

[50] Zondervan, K. and Cardon, L. The Complex Interplay among Factors That Influence Allelic Association *Nature Reviews / Genetics* **5** (2004), 89-100.

Radoslav Nickolov, Ph.D.
Department of Mathematics and Computer Science
Fayetteville State University
1200 Murchison Road
Fyetteville, NC 28301
e-mail: `rnickolov@uncfsu.edu`