Genetic mapping of complex discrete human diseases by discriminant analysis*

LI Xia[1,2,3]**, RAO Shaoqi[2,4]**, Kathy L. MOSER[5], Robert C. ELSTON[2], Jane M. OLSON[2], GUO Zheng[3], ZHANG Tianwen[1] and ZHANG Qingpu[1]

(1. Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China; 2. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44109, USA; 3. Department of Mathematics, Harbin Medical University, Harbin 150086, China; 4. Center for Molecular Genetics, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, Ohio 44195, USA; 5. Department of Medicine, University of Minnesota, Minnesota 55455, USA)

Received August 28, 2001; revised November 5, 2001

Abstract The objective of the present study is to propose and evaluate a novel multivariate approach for genetic mapping of complex categorical diseases. This approach results from an application of standard stepwise discriminant analysis to detect linkage based on the differential marker identity-by-descent (IBD) distributions among the different groups of sib pairs. Two major advantages of this method are that it allows for simultaneously testing all markers, together with other genetic and environmental factors in a single multivariate setting and it avoids explicitly modeling the complex relationship between the affection status of sib pairs and the underlying genetic determinants. The efficiency and properties of the method are demonstrated via simulations. The proposed multivariate approach has successfully located the true position(s) under various genetic scenarios. The more important finding is that using highly densely spaced markers (1～2 cM) leads to only a marginal loss of statistical efficiency of the proposed methods in terms of gene localization and statistical power. These results have well established its utility and advantages as a fine-mapping tool. A unique property of the proposed method is the ability to map multiple linked trait loci to their precise positions due to its sequential nature, as demonstrated via simulations.

Keywords: **categorical traits, IBD linkage analysis, discriminant analysis.**

A complex disease trait refers to a phenotype that does not follow simple Mendelian segregation attributable to a single gene locus, but instead, can be caused by multiple disease loci, their interactions, polygenic inheritance and environmental effects. Genetic analysis of such a trait is complicated by its discrete phenotypic nature (usually binary), for which a linear relationship between the observable phenotypes and the underlying genetic effects does not exist. Current univariate methods for genetic mapping of complex disease traits analyze one marker (or an off-marker position) at a time, which may generate two problems. First, it does not take into account the correlated structure of multiple linked markers purely due to linkage between them, resulting in correlated test statistics. Hence, it tends to produce a flat test statistic profile and a wide empirical confidence interval of the estimated trait location. Second, a complex trait can be caused by effects of multiple linked trait loci and their interactions. A univariate method might generate a false peak of test statistic at a position between two close trait loci, usually called a 'ghost' trait locus.

Methods for genetic mapping of a categorical disease include model-based method[1], model-free method[2], association analysis based on linkage disequilibrium[3] and novel multivariate pattern recognition techniques[4,5]. Although the model-based approaches have been very successful in mapping hundreds of disease-predisposing genes, it becomes difficult to do a good model-based linkage analysis of complex human diseases, for which the modes of inheritance are usually unknown prior to substantial genetic analyses. Model-free approaches, in which no assumption is made about the mode of inheritance of the trait under study, are popular in practice due to their simplicity. Typical model-free methods in human genetics are the affected-sib-pair (ASP) and the extended affected relative pair tests. This group of methods and the transmission/disequilibrium test (TDT) have one disadvantage: they do not make full use of all the available data, information from non-affected individuals being discarded due to the very nature of the

* Supported by the National Natural Science Foundation of China (Grant Nos. 39970397, 30170515) and the National Center of Human Genome Research of USA (Grant No. HG01577)

** To whom correspondence should be addressed. E-mail: shaoqir@yahoo.com, lixia6@yahoo.com

methods.

According to the ways of modeling the relationship between the outward discrete disease phenotypes and the underlying genetic effects, the approaches for genetic mapping of a disease can be divided into two classes. A linear model assumes a linear relationship and the analysis is performed as if the discrete phenotype were continuous. A popular example is (new) Haseman-Elston (H-E) regression[1]. A non-linear model such as a generalized linear model takes the non-linear relationship and discrete nature[6,7] of phenotypes into account. Statistically it is desirable but modeling a sophisticated genetic architecture in disease manifestations can be prohibitively complex and computational demand is high, especially when random polygenic and common environmental effects are included[8].

In this study, we propose a novel multivariate approach to mapping disease loci in human genomic studies with an intention to overcome the weakness inherent in current methods. We integrate a standard stepwise discriminant analysis into a sibpair linkage study. The rationale underlying this approach is that if a disease locus is tightly linked to a (some) molecular marker(s), the differential marker IBD distributions among the affected groups of sibpairs can be observed because of genetic effects of a disease locus on phenotypic manifestations and its tight linkage to nearby markers. Characteristics of this multivariate approach include (i) it uses information both from affected and unaffected sibs; (ii) it avoids explicitly modeling the relationships between the outward ordinal (or binary) phenotypes and the underlying genetic effects (major gene and polygenic background); (iii) due to the sequential testing properties of stepwise discriminant analysis, it enjoys robustness to an assumption on number of trait loci involved and can control multiple trait loci background.

1 Statistical method

Consider a sibpair linkage study of a categorical disease trait. Each sib can take any possible ordinal value, say, $c\,(c = 1, 2, \cdots, C)$. We define an affection group (a specific combination of two ordinal values of a sib pair) as: $G_i\,(i = 1, 2, \cdots, K)$. Thus, the total number of mutually exclusive groups (K) is:

$$K = C + \binom{C}{2}.$$

$C = 2$ corresponds to a binary human disease trait and G_i might be defined as:

G_2 = concordant affected, both sibs in a sib pair are affected;

G_1 = discordant, only one in a sib pair is affected;

G_0 = concordant unaffected, no sibs in a sib pair are affected,

so that we have a population consisting of three mutually exclusive groups. Next, for each sib pair we define a discriminant vector ($\boldsymbol{X}$) which can include the following feature variables: (i) the estimated proportions of alleles shared IBD by the sib pair at L markers along a chromosome (segment); and (ii) other covariates (potential confounding factors for a linkage study), for example, polygenic inheritance and common environment (aliasing with so-called 'household' effect), mother effects (including maternal genetic and environmental effects, and mitochondrial effects) and other epidemiological factors (gender, race, age and so on). In a linkage study, we are searching for a marker or cluster of markers (tightly linked to an unobserved trait locus) whose IBD (the feature variable) distributions among the disease affected groups lead to the best-fit partition (grouping) to the observed one.

Suppose that we have K disease affected groups and M feature variables (L marker IBD variates plus M-L covariates). Let N_1, N_2, $\cdots$, N_K be sample sizes for $G_i\,(i = 1, 2, \cdots, K)$. Then, the feature vector data for N ($N = \sum N_i$) sib pairs can be expressed as:

$$\begin{array}{cccc} \boldsymbol{x}_1^{(1)}, & \boldsymbol{x}_2^{(1)}, & \cdots, & \boldsymbol{x}_{N_1}^{(1)} \\ \boldsymbol{x}_1^{(2)}, & \boldsymbol{x}_2^{(2)}, & \cdots, & \boldsymbol{x}_{N_2}^{(2)} \\ \cdots, & \cdots, & \cdots, & \cdots, \\ \boldsymbol{x}_1^{(K)}, & \boldsymbol{x}_2^{(K)}, & \cdots, & \boldsymbol{x}_{N_K}^{(K)} \end{array}.$$

Denote by $\boldsymbol{\mu}_i\,(i = 1, 2, \cdots, K)$ the mean vector corresponding to the population $G_i\,(i = 1, 2, \cdots, K)$, and assume that $\boldsymbol{x}^{(i)}$ is identically and independently distributed and follows a multivariate normal distribution with a constant variance-covariance matrix,

[1] Elston, R. C. et al. Statistical Analysis for Genetic Epidemiology, Beta4.0-1. The Department of Epidemiology and Biostatistics, Case Western Reserve Universiry, Cleveland, 2000.

i.e.,

$$\boldsymbol{x}^{(i)} \sim N(\boldsymbol{\mu}_i, \sum).$$

A Wilks's ratio Λ, equivalent to a likelihood ratio test statistic[9], which is used to test the effects of the feature variates on separation of disease affected groups, is constructed as:

$$\Lambda = \frac{\det \boldsymbol{W}}{\det \boldsymbol{T}},$$

where det is determinant, $\boldsymbol{W} = \{w_{ij}\}$ the within group covariance matrix and $\boldsymbol{T} = \{t_{ij}\}$ total covariance matrix. A function of Λ, $-(N-(M-K-1)/2-1)\ln\Lambda$, is a random variable asymptotically following a chi-square distribution with $M(K-1)$ degrees of freedom under the null hypothesis $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_K$, no differences among the mean vectors for K disease affected groups.

To assess the contribution from each feature variable, we use a stepwise discriminant analysis procedure[10], which combines forward selection and backward elimination by user defined criteria (with the same p-value of 0.05 for inclusion and exclusion in this study). If the feature variable is a marker IBD, then assessment of it is equivalent to detection of linkage to a putative trait locus. Suppose now that $x_1, x_2, \cdots, x_p$ have been selected from a total of M feature variates at the s-th step. To assess the individual contribution of each variable, say, x_r, from the remaining $M-p$ feature variates, we partition $\boldsymbol{W}$ and $\boldsymbol{T}$ corresponding to subset $x_1, x_2, \cdots, x_p$ and x_r as

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix}, \quad \boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{T}_{21} & \boldsymbol{T}_{22} \end{bmatrix},$$

where $\boldsymbol{W}$, $\boldsymbol{T}$, $\boldsymbol{W}_{ij}$ and $\boldsymbol{T}_{ij}(i, j = 1, 2)$ have similar meanings as described previously. $\boldsymbol{W}_{11}$ and $\boldsymbol{T}_{11}$ are $p \times p$ matrices. $\boldsymbol{W}_{21}$ and $\boldsymbol{T}_{21}$ are $1 \times p$ matrices (vectors). $\boldsymbol{W}_{22}$ and $\boldsymbol{T}_{22}$ are 1×1 matrices (scalars). The Wilks' statistic for p feature variates is

$$\Lambda_p = \frac{\det \boldsymbol{W}_{11}}{\det \boldsymbol{T}_{11}}$$

and for $p+1$ variates, it is

$$\Lambda_{p+1} = \Lambda_p \frac{\det(\boldsymbol{W}_{22} - \boldsymbol{W}_{21}\boldsymbol{W}_{11}^{-1}\boldsymbol{W}_{12})}{\det(\boldsymbol{T}_{22} - \boldsymbol{T}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{T}_{12})}.$$

Hence,

$$\frac{\Lambda_p}{\Lambda_{p+1}} - 1 = \frac{\det(\boldsymbol{T}_{22} - \boldsymbol{T}_{21}\boldsymbol{T}_{11}^{-1}\boldsymbol{T}_{12}) - \det(\boldsymbol{W}_{22} - \boldsymbol{W}_{21}\boldsymbol{W}_{11}^{-1}\boldsymbol{W}_{12})}{\det(\boldsymbol{W}_{22} - \boldsymbol{W}_{21}\boldsymbol{W}_{11}^{-1}\boldsymbol{W}_{12})}.$$

Under the assumption of multivariate normality, we have

$$F_{1r} = \left(\frac{\Lambda_p}{\Lambda_{p+1}} - 1\right)\frac{N-p-K}{K-1} \sim F(K-1, N-p-K).$$

The variable x_r with the largest values of the partial F-statistic is added to the current subset consisting of p feature variates, provided that it exceeds the specified critical value at $\alpha = 0.05$.

We use a similar method in the backward elimination step. First, consider deleting a single variable from a set of p feature variables. For each variable, the partial F-statistic is computed to test whether it provides additional information over the remaining $p-1$ variates. The variate with the smallest partial F-statistic is eliminated first, provided that the statistic does not exceed a specified critical value at $\alpha = 0.05$. The stepwise discriminant procedure alternates forward selection and backward elimination. The procedure stops when none of the included variables can be taken out and no further feature variables can be taken in. We are cautious about the nominal p values given in SAS stepwise discriminant procedure. Extensive empirical simulations prove that the theoretical p value by SAS is liberal. Hence, we resort to a permutation technique to obtain the empirical thresholds.

2 Simulation studies

To investigate the efficiency and properties of the proposed discriminant analysis, four Monte Carlo simulation experiments are carried out. Factors considered include (i) heritability of the trait locus/loci (0.1～0.9), which is defined to be the ratio of segregation variance at the trait locus (loci) over total phenotypic variance of the underlying liability (as explained later) for the ordinal disease phenotype; (ii) marker density (1～20 cM marker spacings); (iii) categorical nature of the observations (ordinal versus binary); (iv) disease prevalence (5%～90%); and (v) two-linked trait loci with various distances between them (20～90 cM).

Only single chromosome segments of different lengths (5, 10, 25, 50, and 100 cM for 1, 2, 5, 10 and 20 cM marker spacings, respectively), covered by six evenly spaced codominant markers each having eight equivalent alleles, are simulated. A diallelic trait locus is simulated in the middle of the interval between the second and third marker locus in the scenarios of one trait locus. The two alleles of the trait locus are equally frequent. For simplicity, simulated pedigrees consist of only nuclear families (two generations) each having four full-sibs unless indicated otherwise. The total number of progeny, N, is fixed at 800. As a result, the total number of sib pairs is 1200. Under each design, the simulation is repeated 100 times. The global statistical power determined by counting the number of runs that have the largest F statistic among finally included marker IBD variates in the stepwise discriminant procedure greater than the empirical threshold at $\alpha = 0.05$ or 0.01, is obtained by simulating 500 replicates under the null model of no trait locus (loci) segregating. The local statistical power is determined similarly but now we confine a region between the two flanking markers of the true trait locus (loci). The standard deviations calculated over the 100 replicates represent the standard errors of parameter estimates and test statistics.

The simulation of genetic values of the trait locus for a liability starts by randomly assigning two alleles to each parent drawn from a random-mating population. A dominant effect is simulated as an interaction between two parental alleles. The liability of each offspring is the sum of its genetic value, the overall mean and its residual error sampled from $N(0, 1)$. Effects of various covariates and polygene on disease liability are not simulated. A set of fixed thresholds truncate the underlying liability into mutually exclusive and exhaustive intervals, which is then translated into observable categorical scores with the desired categorical distributions.

2.1 Heritability of the disease locus (Experiment 1)

A single diallelic trait locus is simulated at 15 cM on a chromosome of 50 cM. Six evenly spaced codominant markers with eight equally frequent alleles (one marker at every 10 cM) on the chromosome make up the linkage group. Four fixed thresholds (-1.25, -1.0, -0.5 and 0.5) truncate the underlying liability into five categorical scores with incidences of 10.56%, 5.32%, 14.99%, 38.30% and 30.85%, respectively. The mean and standard deviations of the location, F-statistic and R^2, and statistical power, obtained from 100 simulations under each of five trait locus attributed heritabilities, are given in Table 1.

Table 1. Effects of heritability of the disease locus on statistical efficiency of the proposed discriminant analysis approach, in terms of statistical power and gene localization, averaged over 100 replicates

Heritability	Location (cM)	Partial R^2	Test statistic (F)	Global statistical power[a] (%)		Local statistical power[a] (%)	
				$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
0.9	15.2 (5.2)	0.189 (0.021)	21.10 (3.03)	100	100	100	100
0.4	15.3 (6.3)	0.044 (0.012)	3.90 (1.11)	99	98	97	96
0.3	15.9 (8.4)	0.032 (0.009)	2.83 (0.84)	81	72	67	62
0.2	17.6 (13.4)	0.024 (0.007)	2.09 (0.59)	48	37	31	24
0.1	22.4 (15.5)	0.019 (0.004)	1.61 (0.39)	13	7	6	3
0.0	26.8 (17.1)	0.016 (0.004)	1.38 (0.39)	5	1	1	0

a) See text for explanation. Standard errors are in parentheses.

The proposed multivariate approach has successfully found the true position, with reasonable biases under the low trait locus attributed heritability ($h^2 = 0.10$) due to the interference of false positives. With heritability increasing, both global and local statistical powers are dramatically increased and the mean of estimated location is closer to the true position. It is obvious that heritability plays an important role in separating disease affected groups via marker IBD information. The proportion of marker IBD variances among sib pairs explained by disease affection grouping (R^2) increases with the magnitude of heritability. Larger standard errors for the F-statistic under higher heritabilities are due to scaling effects.

2.2 Marker density (Experiment 2)

The same design as for Experiment 1 is used except that heritability of a trait locus is fixed at 0.30. Five levels of marker densities, from sparsely to densely distributed markers, are investigated. As shown in Table 2, marker density influences the performance of the proposed approach in the sibpair linkage analysis of an ordinal trait. There may be an optimal marker density (2 ~ 10 cM marker spacings) in which genetic mapping by our approach reaches its maximal power and minimal estimation errors occur. If so, this would provide a guide to a multi-stage genetic mapping design. Using highly densely spaced markers (1 ~ 2 cM) will lead to little loss of statistical efficiency in terms of localization and global statistical power. These results might well establish its utility and show their advantages as a fine-mapping tool.

Table 2. Effects of marker density on statistical efficiency of the proposed discriminant analysis approach, in terms of statistical power and gene localization, averaged over 100 replicates

Marker distance[a] (cM)	Location[b] (cM)	Partial R^2 [b]	Test statistic[b] (F)	Global statistical power (%)		Local statistical power (%)	
				$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$
20 (30)	32.8 (18.1)	0.026 (0.007)	2.30 (0.62)	60	49	51	44
10 (15)	15.8 (9.5)	0.031 (0.008)	2.70 (0.72)	82	75	69	63
5 (7.5)	7.8 (4.5)	0.032 (0.009)	2.83 (0.86)	80	73	66	61
2 (3)	3.2 (2.1)	0.035 (0.011)	3.12 (0.97)	86	82	60	58
1 (1.5)	1.7 (1.3)	0.037 (0.011)	3.27 (1.04)	84	81	46	44

a) The true location of the trait locus are in parentheses for this column. b) Standard errors are in parentheses.

2.3 A binary trait and disease prevalence (Experiment 3)

Theoretically or suggested by empirical simulations, genetic mapping for a binary disease is less efficient than that for an ordinal disease using the generalized linear model-based approaches[6,7]. However, this argument may not be applied to a sibpair linkage study using a discriminant analysis approach in that we work on a second moment form (group) of the marginal phenotypes instead of directly modeling the relationship between the marginal phenotypes and the underlying genetic effects. Moreover, in a discriminant analysis, affected groups obtained from the marginal ordinal phenotypes of a sib pair are no longer a response variable but instead serve as an explanatory variable for marker IBD distributions. These characteristics may lead to very different conclusions from other approaches to modelling marginal phenotypes.

Table 3. Effects of prevalence of the disease locus on statistical efficiency of the proposed discriminant analysis approach for genetic mapping of complex binary human diseases, in terms of statistical power and gene localization, averaged over 100 replicates

Prevalence (%)	Location (cM)	Partial R^2	Test statistic (F)	Global statistical power (%)		Local statistical power (%)	
				$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$
90	23.5 (17.6)	0.004 (0.002)	2.50 (1.41)	5	4	0	0
70	20.6 (14.2)	0.008 (0.004)	4.76 (2.37)	49	41	34	28
50	16.6 (9.6)	0.012 (0.006)	7.44 (3.94)	73	67	62	57
30	14.8 (8.0)	0.017 (0.008)	10.38 (5.00)	94	92	80	78
20	15.6 (8.2)	0.015 (0.007)	8.92 (4.82)	80	76	71	68
10	18.1 (15.0)	0.006 (0.003)	3.59 (2.08)	26	18	15	13
5	21.7 (14.9)	0.006 (0.003)	3.51 (1.84)	25	17	15	13

Standard errors are in parentheses.

We simulate a binary trait with different incidence rates of 5%~90% (Table 3). Two important conclusions can be made from this simulation experiment: (i) greater power can be obtained for a binary disease trait than for an ordinal trait. Under the same heritability of a trait locus ($h^2=0.30$), the power (94% at $\alpha=0.05$) for a binary disease trait with prevalence of 30% is larger than the corresponding ordinal trait (81%, see Table 1 for cross reference); (ii) performance of the proposed approach is not even with respect to symmetry of the marginal binary phenotype. The optimal performance in terms of localization and statistical power is attained at a prevalence of 30%.

2.4 A binary trait that is controlled by two-linked trait loci (Experiment 4)

A unique property of the proposed method is the ability to map multiple-linked trait loci to their precise positions due to its sequential nature. We demonstrate this property by simulating two-linked trait loci with various distances between them. For this purpose, a binary trait with an incidence of 30%, controlled by two trait loci located on a single chromosome of length of 100 cM covered by eleven evenly spaced codominant markers (each at 10 cM), is simulated. For convenience, we assume that the two trait loci contribute an equal genetic variance to the variability of the underlying liability. No epistasis between the two loci is simulated. Note that under the assumption of no epistasis and gender specific genetic heterogeneity the covariance between the two trait loci due to linkage is

$$\frac{1}{2}[(\alpha_1-\alpha_2)_1(\alpha_1-\alpha_2)_2](1-2r),$$

where r is the recombination fraction between the two trait loci and $(\alpha_1-\alpha_2)_q$ represents the effect of gene substitution at the qth $(q=1,2)$ trait locus. It is evident that this covariance depends on the recombination fraction between the two loci. The closer the two trait loci, the larger the covariance.

We fixed the total trait locus attributed heritability for the liability to be 0.50. The gene substitution effects are appropriately scaled under each distance option. Distances between the two trait loci are 90, 60, 40, 20 cM, respectively. Under the assumption of equal genetic contribution of the two trait loci, the corresponding gene substitutions are 0.9264, 0.8766, 0.8306 and 0.7738, respectively, which translate the marginal genetic variance of each trait locus to be 0.4290 ($h^2=0.21$), 0.3842 ($h^2=0.19$), 0.3449 ($h^2=0.17$) and 0.2994 ($h^2=0.15$). The two mean trait-locations are determined by averaging the locations of the two markers with the largest F-statistics (included in the final subset during the stepwise discriminant analysis). The statistical power is obtained by counting the number of replicates with an F-statistic larger than the critical value at $\alpha=0.05$.

Table 4. Effects of the distance between two QTLs on statistical efficiency of the proposed discriminant analysis approach for genetic mapping of complex binary human diseases, in terms of statistical power and gene localization, averaged over 100 replicates

Distance between two QTLs[a)]	Location (cM)		Partial R^2		Test statistic (F)		Global statistical power (%)		Local statistical power (%)	
	QTL1	QTL2	QTL1	QTL2	QTL1	QTL2	QTL1	QTL2	QTL1	QTL2
							$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$
							($\alpha=0.01$)	($\alpha=0.01$)	($\alpha=0.01$)	($\alpha=0.01$)
90	7.1	90.4	0.015	0.015	8.84	9.12	85	81	80	68
(5, 95)	(7.3)	(9.2)	(0.007)	(0.008)	(4.48)	(5.04)	(82)	(79)	(77)	(66)
60	18.5	71.2	0.019	0.018	11.79	10.89	86	93	68	71
(15, 75)	(9.6)	(10.0)	(0.010)	(0.008)	(6.50)	(5.23)	(84)	(90)	(66)	(69)
40	27.7	64.20	0.021	0.021	13.25	12.67	86	93	72	80
(25, 65)	(8.0)	(9.9)	(0.014)	(0.011)	(9.13)	(6.71)	(81)	(80)	(69)	(69)
20	32.8	55.4	0.021	0.023	12.70	14.40	72	71	59	55
(35, 55)	(11.6)	(13.2)	(0.016)	(0.018)	(10.14)	(11.45)	(65)	(67)	(52)	(53)

a) The true locations of two trait loci are in parentheses for this column. Standard errors are in parentheses.

The results, shown in Table 4, prove that the proposed multivariate approach can map multiple-linked loci to their precise positions even in the situation that two trait loci are close to each other (20 cM) and separated by only two markers. This nice property frees a concern of a 'ghost' quantitative trait locus, a well-known phenomenon for detection of closely linked trait loci by a single trait locus model. This has established its advantages of the proposed method as a genome screening tool as it is robust to the assumption of the number of trait loci involved, which is indeed unknown prior to a substantial genetic analysis. Furthermore, it saves the complex modeling work for a multiple trait loci model because it is essentially model-free in this aspect. It is interesting to note that the averaged R^2 for each trait locus of two closely linked trait loci (e.g. 20 and 40 cM) is higher than that for loosely linked loci although marginal genetic contributions under smaller distances between them are lower. Generally, it is difficult to detect multiple closely linked loci, which is demonstrated by a small decrease in statistical power.

3 Discussion

In this work, we have demonstrated the potential of a stepwise discriminant analysis as a screening tool in a linkage study. Because the analysis is carried out in a single run, the method does not suffer from the problems associated with multiple testing. Another important feature of the proposed method is that it is essentially model-free in that no specific relationship between the response variable and independent variables is assumed. Hence, it is not sensitive to model misspecifications as is a linear model or a generalized linear model.

A discriminant analysis has some analogies to a logistic regression. For example, both can be used to analyze discrete data. However, we point out a weakness in the application of logistic regression in a sibpair linkage study. The very act of taking cross-product in a sibpair regression has changed the relationship between the model components so that a threshold model implied in a logistic regression that is used to model the relationship between the re-defined phenotype (affection state of a sibpair) and its components might not be valid. Consequently, logistic regression might have a lower statistical power[5]. Though a bit surprising, it is logical that a discriminant analysis performs better than a logistic regression because it is generally inferior to a discriminant analysis when multivariate assumptions hold or are nearly true[9]. In discriminant analysis, we need distributional assumptions on the feature vector but not on the grouping variable, which is the key difference with logistic models.

We assumed in this study that the feature vectors for sib pairs are independent and follow multivariate normal distributions, both of which can be violated in the context of a sib pair linkage study, especially when large sibships are recruited. For example, the IBD vectors of sib pairs within a family tend to be correlated. Also, it may happen that some feature variables are not normally distributed. Moreover, the p-values reported by the stepwise discriminant analysis procedure of SAS are liberal. These issues deserve attention in future studies.

References

1 Terwilliger, J. D. Linkage analysis, model-based. In: Armitage, P. et al. ed. Encyclopedia of Biostatistics. West Sussex: John Wiley & Sons Ltd., 1998, Vol 3: 2279~2291.

2 Olson, J. M. Linkage analysis, model-free. In: Armitage, P. et al. ed. Encyclopedia of Biostatistics. West Sussex: John Wiley & Sons Ltd., 1998, Vol 3: 2291~2301.

3 Spielman, R. S. et al. The TDT and other family-based tests for linkage disequilibrium and association. Am. J. Human Genet., 1996, 59: 983.

4 Lucek, P. R. et al. Neural network analysis of complex traits. Genet. Epidemiol., 1997, 14: 1101.

5 Li, X. et al. Locating the genes underlying a simulated complex disease by discriminant analysis. Genet. Epidemiol., 2001, 21 (1): S516.

6 Rao, S. et al. Mapping quantitative trait loci for ordered categorical traits in four-way crosses. Heredity, 1998, 81: 214.

7 Rao, S. et al. Strategies for genetic mapping of categorical traits. Genetica, 2000, 109: 183.

8 Yi, N. et al. Bayesian mapping of quantitative loci under the identity-by-descent-based variance component model. Genetics, 2000, 156: 411.

9 Mclachlan, G. J. Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley, 1992, 392~400.

10 SAS Institute Inc. SAS/STAT @ User' s Guide. Version 6, Fourth Edition, Volume 2, Cary, NC: SAS Institute Inc., 1989, 1493~1509.