J Hum Genet (2008) 53:739–746
DOI 10.1007/s10038-008-0307-x

ORIGINAL ARTICLE

Two-stage designs to identify the effects of SNP combinations on complex diseases

Guolian Kang · Weihua Yue · Jifeng Zhang · Marianne Huebner · Handi Zhang · Yan Ruan · Tianlan Lu · Yansu Ling · Yijun Zuo · Dai Zhang

Received: 31 October 2007 / Accepted: 21 May 2008 / Published online: 27 June 2008

Abstract The genetic basis of complex diseases is expected to be highly heterogeneous, with many disease genes, where each gene by itself has only a small effect. Based on the nonlinear contributions of disease genes across the genome to complex diseases, we introduce the concept of single nucleotide polymorphism (SNP) synergistic blocks. A two-stage approach is applied to detect the genetic association of synergistic blocks with a disease. In the first stage, synergistic blocks associated with a complex disease are identified by clustering SNP patterns and choosing blocks within a cluster that minimize a diversity criterion. In the second stage, a logistic regression model is given for a synergistic block. Using simulated case–control data, we demonstrate that our method has reasonable power to identify gene–gene interactions. To further evaluate the performance of our method, we apply our method to 17 loci of four candidate genes for paranoid schizophrenia in a Chinese population. Five synergistic blocks are found to be associated with schizophrenia, three of which are negatively associated (odds ratio, OR < 0.3, $P < 0.05$), while the others are positively associated (OR > 2.0, $P < 0.05$). The mathematical models of these five synergistic blocks are presented. The results suggest that there may be interactive effects for schizophrenia among variants of the genes neuregulin 1 (*NRG1*, 8p22-p11), *G72* (13q34), the regulator of G-protein signaling-4 (*RGS4*, 1q21-q22) and frizzled 3 (*FZD3*, 8p21). Using synergistic blocks, we can reduce the dimensionality in a multi-locus association analysis, and evaluate the sizes of interactive effects among multiple disease genes on complex phenotypes.

Keywords SNP pattern · Synergistic block · Association study

Guolian Kang and Weihua Yue contributed equally to this work.

G. Kang · J. Zhang
Key Laboratory of Systems and Control,
Institute of Systems Science,
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences,
Beijing 100080, China

W. Yue · H. Zhang · Y. Ruan · T. Lu · Y. Ling · D. Zhang (✉)
Key Laboratory for Mental Health, Ministry of Health,
Institute of Mental Health, Peking University,
51 Hua Yuan Bei Road, Beijing 100083, China
e-mail: daizhang@hsc.pku.edu.cn

Present Address:
G. Kang
Department of Biostatistics,
The University of Alabama at Birmingham,
Birmingham, AL 35294, USA

M. Huebner · Y. Zuo
Department of Statistics and Probability,
Michigan State University, East Lansing, MI 48824, USA

Introduction

Genetic association analyses need to address genetic and phenotypic heterogeneities for complex diseases. Investigating associations between marker genotypes and disease phenotypes for only one locus at a time without considering combinations of (unlinked) loci may capture only a small proportion of the total combined effect of all disease loci. Thus new methods are needed that allow the joint analysis of multiple loci for association analysis (Hoh and Ott 2003; Kang et al. 2008). For a complex disease, individual loci with small effects may not be sufficient to identify a genetic association with a clinical syndrome

(Ritchie et al. 2001). However, the combined effect of multiple loci with minor or modest effect sizes might confer additive or multiplicative genetic contributions (Ritchie et al. 2001). Genotypic combinations at some loci that contain several alleles may be specific to certain classes of cases. Multiple loci with minor or modest effect sizes and that are located in different or the same chromosomes could form a joint functional unit for disease susceptibility. The problem is to identify these loci and quantify the interaction among these loci. Several algorithms are available to examine SNP combinations for complex diseases (Ritchie et al. 2001, 2003; Goodman et al. 2006; Onay et al. 2006; Hahn et al. 2003). These methods include dimension reduction, which combines all possible SNP–SNP interactions and chooses the set of SNPs that minimizes the classification error of cases and controls (Ritchie et al. 2003; Hahn et al. 2003). Some analyses use classification scoring functions to identify subsets of SNPs likely associated with disease risk (Goodman et al. 2006). Multivariate logistic regression and bootstrap analyses can be used to select SNP–SNP interactions via stepwise regression (Onay et al. 2006).

We use a two-stage approach that identifies multiple SNP patterns and evaluates their risk with a disease. The method allows the inclusion of any number of main effects together with the highest interaction term. At the first stage, synergistic blocks associated with a complex disease are identified. At the second stage, a logistic regression model is chosen to evaluate the interactive effects among different loci in a synergistic block. Logistic regression models for SNPs in synergistic blocks have better statistical power and more statistical significance than using a full model that includes all SNPs and their interactions, because of the reduced number of parameters.

This method is applied to a case–control schizophrenia study from a Chinese Han population to detect the effects of 17 loci of four candidate genes, the regulator of G-protein signaling-4 (*RGS4*, 1q21-q22), frizzled 3 (*FZD3*, 8p21), neuregulin 1 (*NRG1*, 8p22-p11), and *G72* (13q34), on the susceptibility to this disease, since these have been reported in other studies to be possible candidate genes for schizophrenia (Yue et al. 2006, 2007; Zhang et al. 2004; Harrison and Owen 2003).

Harrison and Weinberger proposed that schizophrenia might be a genetic disorder of the synapses. There may be a putative common effect of schizophrenia susceptibility genes on the plasticity and functioning of synapses and other neurodevelopmental processes. We hypothesized that the NRG1, G72, RGS4 and FZD3 may play a common role in the pathogenesis of neurodevelopment and plasticity in schizophrenia.

Methods

Identifying SNP combination patterns associated with the phenotype

We assume n marker loci with two alleles each which are in Hardy–Weinberg equilibrium (HWE). The genotypes for one single nucleotide polymorphism (SNP) marker are coded by 0, 1, or 2. Denote a combination pattern of n SNPs (an SNP pattern) as an n-dimension genotype vector $G = (a_1 a_2 \ldots a_n)$, where n is the number of marker loci, and a_i is the genotype of the ith SNP position. Denote $\pi^A(G)$ and $\pi^U(G)$ as the frequencies of the SNP pattern G in affected cases and unaffected controls, respectively. Then, the odds ratio of SNP pattern G (OR_G) is the ratio of the odds of SNP pattern G in the cases to that of SNP pattern G in the controls, i.e.,

$$\mathrm{OR}_G = \frac{\pi^A(G)}{1-\pi^A(G)} \Big/ \frac{\pi^U(G)}{1-\pi^U(G)},$$

when $\mathrm{OR}_G > 1$, the SNP pattern G can be positively associated with the disease; when $0 < \mathrm{OR}_G < 1$, the SNP pattern G can be negatively associated with it; when OR_G is very close to 1, the SNP pattern G cannot be associated with the disease.

The hypotheses used to test whether SNP pattern G is associated with a disease are defined as follows:

$$H_0 : \ln OR_G = 0, \quad H_1 : \ln OR_G \neq 0.$$

When the null hypothesis H_0 is rejected, then we can conclude that there is evidence for the association of G with the disease.

Let N^A and N^U be the numbers of cases and controls; $N^A(G)$ and $N^U(G)$ denote the number of subjects with the SNP pattern G among the cases and controls, respectively. Then, the log transform of the sample odds ratio, $\ln \hat{\mathrm{OR}}_G$ is given by

$$\ln \hat{\mathrm{OR}}_G = \ln\left[\frac{N^A(G)}{N^A - N^A(G)} \Big/ \frac{N^U(G)}{N^U - N^U(G)}\right].$$

This random variable has a large-sample approximate normal distribution with a mean of $\ln \mathrm{OR}_G$ and a standard deviation, referred to as the asymptotic standard error (ASE) (Agresti 1996), of

$$\mathrm{ASE}(\ln \hat{\mathrm{OR}}_G) = \sqrt{\frac{1}{N^A(G)} + \frac{1}{N^A - N^A(G)} + \frac{1}{N^U(G)} + \frac{1}{N^U - N^U(G)}}.$$

Therefore, the statistic

$$Z(G) = \frac{\ln \hat{\mathrm{OR}}_G}{\mathrm{ASE}(\ln \hat{\mathrm{OR}}_G)}$$

can be used to test the above hypothesis. If $|Z(G)| > u_{\alpha/2}$, (where $u_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution) then H_0 is rejected.

When the number of candidate SNPs is very large, a genetic algorithm (GA) (Goldberg 1989) can be applied to elucidate the associated SNP patterns quickly. In a GA, we examine every SNP pattern G as a candidate solution to the problem of associated SNP patterns. The fitness of an SNP pattern is defined as

$$\mathrm{fitness}_G = |Z(G)| = \left| \frac{\ln \hat{\mathrm{OR}}_G}{\mathrm{ASE}(\ln \hat{\mathrm{OR}}_G)} \right|.$$

The sample odds ratio $\left(\hat{\mathrm{OR}}_G\right)$ of SNP pattern G is 0 (or ∞) if $N^A(G) = 0$ (or $N^U(G) = 0$). The slightly amended estimator can be expressed as (Agresti 1996)

$$\mathrm{fitness}_G^A = \left| \frac{\ln\left[\frac{N^A(G)+0.5}{N^A - N^A(G)+0.5} \Big/ \frac{N^U(G)+0.5}{N^U - N^U(G)+0.5}\right]}{\sqrt{\frac{1}{N^A(G)+0.5} + \frac{1}{N^A - N^A(G)+0.5} + \frac{1}{N^U(G)+0.5} + \frac{1}{N^U - N^U(G)+0.5}}} \right|.$$

The parameters in the genetic algorithm are population sizes, crossover probabilities and mutation probabilities (Goldberg 1989). By assigning different values to the parameters, we can get different SNP patterns that are significantly associated with the disease.

Patterns associated with clustering

Cluster analysis is used to group similar SNP patterns associated with the disease. An SNP synergistic block is the loci in the SNP pattern cluster.

The clustering of SNP patterns is based on a similarity measure (Duda and Schafer 2001). A matched vector of the associated SNP pattern G may be denoted by $Q^G = [q_1, q_2, \ldots, q_N]'$, where $q_i = 1$ if sample i has an associated SNP pattern G; otherwise, $q_i = 0$, $1 \le i \le N$, $N = N^A + N^U$. The similarity distance $d(G_1, G_2)$ between the associated SNP patterns G_1 and G_2 can be defined as

$$d(G_1, G_2) = 1 - \left[\frac{(Q^{G_1})^T Q^{G_2}}{\|Q^{G_1}\|_2^{1/2} \|Q^{G_2}\|_2^{1/2}} \right]^2,$$

(where $\|a\|_2$ represents the 2-norm of vector a), which is in fact the cosine measure between points. This distance measure is used to investigate the clustering of SNP patterns.

Grouping loci/SNPs into synergistic blocks

Based on the previous step, we cluster the SNP patterns into several SNP pattern clusters that include similar associated SNP patterns. Let R denote one SNP pattern cluster and N_R denote the number of associated SNP patterns in R. The set of loci considered in R is referred to as $\gamma(R)$. Let $G^i = \left(a_1^i a_2^i \ldots a_n^i\right)$ denote the ith associated SNP pattern in R, where $1 \le i \le N_R$. Then, the difference between the ith and the jth SNP pattern can be defined as $G^i - G^j = \left(a_1^i - a_1^j a_2^i - a_2^j \cdots a_n^i - a_n^j\right)$, where, for each locus $k \in \{1, 2, \ldots, n\}$, we get $a_k^i - a_k^j = \begin{cases} 1, & a_k^i \neq a_k^j; \\ 0, & a_k^i = a_k^j. \end{cases}$

The diversity of loci considered between the ith and jth SNP patterns is expressed as

$$D_{ij}(W) = \left(G^i - G^j\right) W \left(G^i - G^j\right)',$$

where $W = (w_{ij})$ is a weight matrix that can take any of several forms, depending on the intended use of the ancillary information. The sum of $D_{ij}(W)$ can be used to describe the diversity of the loci considered in a SNP pattern cluster. In particular, the diversity of a block B can be defined as

$$D_B = \frac{1}{2N_R} \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} D_{ij}(W_B) = \frac{1}{2N_R} \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} \left(G^i - G^j\right) W_B \left(G^i - G^j\right)',$$

where $W_B = (w_{ij})$ with $w_{ij} = \begin{cases} 1, & i = j \text{ and } i \in \gamma(B); \\ 0, & \text{otherwise}, \end{cases}$

where $\gamma(B)$ is the set of loci considered in block B.

Similarly, the diversity of all loci in $\gamma(R)$ is defined as

$$D_R = \frac{1}{2N_R} \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} D_{ij}(W_R) = \frac{1}{2N_R} \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} \left(G^i - G^j\right) W_R \left(G^i - G^j\right)',$$

where $W_R = (w_{ij})$ with $w_{ij} = \begin{cases} 1, & i = j \text{ and } i \in \gamma(R); \\ 0, & \text{otherwise}. \end{cases}$

Therefore, we choose one subset B of $\gamma(R)$ as a synergistic block of this SNP pattern cluster. There are three conditions for choosing B:

(1) $B \in R_g$.
(2) Only choose significant SNP patterns whose fitness satisfy $Fitness(G_B) > u_{\alpha/2}$.
(3) Choose the value of B that minimizes $\frac{D_B}{D_R} + \frac{1}{Adapt(G_B)}$, where R_g represents all possible subsets of $\gamma(R)$, and G_B is one of the SNP patterns corresponding to block B.

The G_B that satisfies (2) and (3) is then an associated SNP pattern for this synergistic block B.

Suppose there are v loci in a synergistic block, i.e., v covariates. The multivariate logistic regression model (Agresti 1996) that includes all main effects and the highest order interaction term is

$$\log it(p) = \beta_0 + \beta_1 SNP_1 + \cdots + \beta_v SNP_v + \beta_{v+1} SNP_1 \times SNP_2 \times \cdots \times SNP_v,$$

where $SNP_i \in \{0,1\}$, $1 \le i \le v$, and $SNP_i = 1$ when the genotype of the ith locus is the same as that of the same locus in the synergistic block and 0 otherwise. Furthermore, $p = \Pr(\text{Affected}|(SNP_1, \ldots, SNP_v) \in \{0,1\}^v)$. In this model, we only consider the main effects and the interactions of all loci in a synergistic block.

Results

Application of the synergistic block algorithm to simulated data

Two data sets containing 100 replicates of 200 cases and 200 controls for ten unlinked biallelic loci were simulated using two two-locus interaction models as examples. The first and second loci out of ten were chosen as the disease loci with interaction effects. This number of replicates was selected to provide method validation and to enable exhaustive computational searches of all possible fourth-order SNP combinations to be performed. Hardy–Weinberg equilibrium was assumed. For the two-locus interaction disease models, the interaction effect was simulated using penetrance functions via two models. Model 1: P(Disease|AAbb) = 0.02, P(Disease|AaBb) = 0.2, P(Disease|aaBB) = 0.02, and P(Disease|others) = 0; Model 2: P(Disease|AAbb) = 0.2, P(Disease|AaBb) = 0.2, P(Disease|aaBB) = 0.2, and P(Disease|others) = 0, where A, a, B and b represent the alleles for the disease loci (Frankel and Schork 1996), with a population allele frequency of 0.3 in all cases. For the other SNPs, their population allele frequencies are drawn from the uniform distribution [0.1, 0.9]. Here, we chose the SNP combinations from among all of the fourth-order ones with fitnesses >3 as clusters when searching for a synergistic block.

The results are presented in Table 1. It includes the percentage of the time that the synergistic block associated with the disease was identified within 100 replicates, the average frequency in cases and controls, the average fitness with its standard deviation, and the average odds ratio for the SNP combination corresponding to the synergistic block. From these results, we know that for this interaction disease model, the synergistic block method has reasonable power to identify high-order gene–gene interactions.

We also evaluated the type I error rate by simulating 100 data sets under the assumption that there is no interaction effect between unlinked loci on the disease. If there is also no main genetic effect on the disease, then we should not find any SNP synergistic block; if there is one locus with a main genetic effect on the disease, we will find this locus 70 times out of 100, and there is no multi-locus synergistic block (data not shown).

Application of the synergistic block algorithm to schizophrenia data

Four candidate genes, *RGS4*, *FZD3*, *NRG1*, *G72*, and seventeen SNPs that were genotyped and analyzed in this study are listed in Table 2. The SNPs SNP8NRG221533 and SNP8NRG243177 (*NRG1*) and rs2323019 and rs352203 (*FZD3*) are in linkage disequilibrium (Table 3). It is generally speculated that these four genes functionally converge to act upon schizophrenia by influencing synaptic plasticity and cortical microcircuitry (Harrison and Weinberger 2005). Seventeen SNPs were genotyped across these four candidate genes in the Chinese Han population, which included 120 schizophrenia cases and 225 healthy controls. Prior reports have suggested that variations in the genes may be associated with increased risk of paranoid schizophrenia (Yue et al. 2006, 2007; Zhang et al. 2004; Harrison and Owen 2003).

Table 1 Simulation results for 200 cases and 200 controls in two disease models*

Disease model	Synergistic block	Proportion of times identified (%)	Average freq. in cases (%)	Average freq. in controls (%)	Average fitness (standard deviation)	Average OR
Model 1	[AaBb]	100	0.5471	0.0733	8.8448 (0.4886)	16.2827
Model 2	[aaBB]	87	0.1682	0.0370	6.9197 (0.5140)	10.6588
	[AaBb]	13	0.3950	0.0635	3.8688 (0.5276)	6.2720

Model 1: P(Disease|AAbb) = 0.02, P(Disease|AaBb) = 0.2, P(Disease|aaBB) = 0.02, and P(Disease|others) = 0; Model 2: P(Disease|AAbb) = 0.2, P(Disease|AaBb) = 0.2, P(Disease|aaBB) = 0.2, and P(Disease|others) = 0, where A, a, B and b represent the alleles for the disease loci and P(A) = P(B) = 0.3 in these two models

* Here we choose $u_{\alpha/2} = 3$ as a threshold for the selection of associated SNP patterns

Using a genetic algorithm, we identified 652 significantly associated SNP patterns using 120 cases and 225 controls. The *P* values for the permutation tests were at most 0.01 after adjusting for multiple testing using Benjamini–Hochberg's algorithm (Benjamini and Hochberg 1995).

These SNP patterns were then clustered, producing five (two positively and three negatively) associated SNP pattern clusters. Five synergistic blocks (Table 4) were identified. Logistic regression models were used to quantify the size of the effect. Synergistic block 1, including the polymorphisms *NRG1* (rs3735774), *NRG1* (rs 2919390), and *RGS4* (rs12753561), is associated with schizophrenia (OR 6.74, $P < 0.001$). The interaction between these three SNPs is statistically significant ($P = 0.0014$). Synergistic block 2, including the polymorphisms *RGS4* (rs12753561), *FZD3* (rs2241802) and *FZD3* (rs2323019), is associated with schizophrenia (OR 2.0948, $P < 0.001$). The interaction between these three SNPs is statistically significant ($p = 0.0221$). Synergistic block 3, including the polymorphisms *NRG1* (SNP8NRG221533) *NRG1* (rs3735774) and *NRG1* (rs6988339), is associated with schizophrenia (OR 0.2014, $P < 0.001$). The interaction between these three SNPs is statistically significant ($P = 0.0017$). Similar results were obtained for synergistic blocks 4 and 5.

Information on the models is shown in Table 5. These results indicate that the interactions of alleles at different loci located on different or the same chromosomes may significantly influence complex human diseases.

Table 2 Genes and SNP ID numbers

Gene name	SNP ID numbers*
NRG1 (8p22-p11) (1–7)	SNP8NRG221533, SNP8NRG243177, rs3924999, rs3735774, rs2954041, rs2919339, rs6988390
G72 (13q34) (8–10)	rs2391191, rs778294, rs947267
RGS4 (1q21-22) (11–13)	rs10759, rs2344671, rs12753561
FZD3 (8p21) (14–17)	rs2241802, rs2323019, rs352203, rs880481

* SNP ID numbers of the gene polymorphisms in the NCBI GenBank database (http://www.ncbi.nlm.nih.gov/SNP) and in the deCODE genetics database (http://www.decode.com/nrg1/markers)

Discussion

In the present study, we identify synergistic blocks as being a genetic factor in complex disease, and use a two-stage approach to detect the effects of synergistic blocks on

Table 3 The linkage disequilibrium information for four genes

r^2	SNP8NRG2 43177	rs3924999	rs3735774	rs2954041	rs2919339	rs6988390
(A) *NRG1*						
SNP8NRG221533	0.651	0.000	0.007	0.008	0.002	0.001
SNP8NRG243177	–	0.000	0.003	0.019	0.006	0.000
rs3924999	–	–	0.020	0.058	0.005	0.005
rs3735774	–	–	–	0.009	0.006	0.014
rs2954041	–	–	–	–	0.046	0.193
rs2919339	–	–	–	–	–	0.091

r^2	rs778294	rs947267
(B) *G72*		
rs2391191	0.036	0.157
rs778294	–	0.076

r^2	rs2344671	rs12753561
(C) *RGS4*		
rs10759	0.004	0.050
rs2344671	–	0.016

r^2	rs2323019	rs352203	rs880481
(D) *FZD3*			
rs2241802	0.292	0.181	0.153
rs2323019	–	0.591	0.177
rs352203	–	–	0.138

Table 4 The synergistic blocks extracted from the 17 SNPs of *NRG1*, *G72*, *RGS4* and *FZD3*

Synergistic block	Associated SNP pattern*	Freq. of case	Freq. of cont.	Fitness	OR	*P* value
Synergistic block 1 = [*NRG1* (rs3735774) *NRG1* (rs2919390) *RGS4* (rs12753561)]	(* * * 1 * 0 * * * * * * 0 * * * *)	0.1583	0.0267	3.9848	6.7369	3.4×10^{-5}
Synergistic block 2 = [*RGS4* (rs12753561) *FZD3* (rs2241802) *FZD3* (rs2323019)]	(* * * * * * * * * * * * 0 1 0 * *)	0.4333	0.2622	3.2117	2.0948	6.6×10^{-4}
Synergistic block 3 = [*NRG1* (SNP8NRG221533) *NRG1* (rs3735774) *NRG1* (rs6988339)]	(1 * * 0 * * 0 * * * * * * * * *)	0.0667	0.2578	3.9894	0.2014	3.3×10^{-5}
Synergistic block 4 = [*NRG1* (SNP8NRG243177) *G72* (2391191) *FZD3* (rs2323019)]	(* 1 * * * * * 1 * * * * * * 0 * *)	0.1333	0.3378	3.9520	0.2947	3.9×10^{-5}
Synergistic block 5 = [*FZD3* (rs2241802) *FZD3* (rs352203)]	(* * * * * * * * * * * * * * 0 * 0 *)	0.0417	0.2400	4.1072	0.1349	2.0×10^{-5}

* The order of SNPs are (SNP8NRG221533, SNP8NRG243177, rs3924999, rs3735774, rs2954041, rs2919339, rs6988390, rs2391191, rs778294, rs947267, rs10759, rs2344671, rs12753561, rs2241802, rs2323019, rs352203, rs880481), where the genotypes at an individual locus are represented by 0 or 1, and * denotes that the genotype of the locus is arbitrary

Table 5 SNP interaction effects for logistic regression models

Parameter	Estimated value	Standard error	Chi-square	*P* value	OR	95% CI for EXP(B)	
						Lower	Upper
Model (1)							
NRG1 (rs3735774)	−1.6717	1.0689	2.4500	0.1178	0.1879	0.0231	1.5271
NRG1 (rs2919390)	−0.5971	0.2701	4.8900	0.0271	0.5504	0.3242	0.9345
RGS4 (rs12753561)	−0.0846	0.3105	0.0700	0.7853	0.9189	0.5000	1.6887
NRG1(rs3735774)**NRG1* (rs2919390)**RGS4* (rs12753561)	**3.7369**	**1.1712**	**10.1800**	**0.0014**			
Model (2)							
RGS4 (rs12753561)	−0.3622	0.3978	0.8300	0.3626	0.6961	0.3192	1.5181
FZD3 (rs2241802)	−0.3447	0.4559	0.5700	0.4495	0.7084	0.2899	1.7313
FZD3 (rs2323019)	−0.4621	0.4665	0.9800	0.3219	0.6300	0.2525	1.5718
RGS4(rs12753561)**FZD3* (rs2241802)* *FZD3* (rs2323019)	**1.2004**	**0.5247**	**5.2300**	**0.0221**			
Model (3)							
NRG1 (SNP8NRG221533)	−0.0120	0.2697	0.0000	0.9646	0.9881	0.5824	1.6763
NRG1 (rs3735774)	−0.5306	0.3612	2.1600	0.1419	0.5883	0.2898	1.1941
NRG1 (rs6988339)	0.1913	0.3117	0.3800	0.5394	1.2108	0.6573	2.2305
NRG1(SNP8NRG221533)**NRG1*(rs3735774)**NRG1* (rs6988339)	**−1.6435**	**0.5236**	**9.8500**	**0.0017**			
Model (4)							
NRG1 (SNP8NRG243177)	0.0063	0.2977	0.0000	0.9831	1.0063	0.5615	1.8036
G72 (2391191)	0.0673	0.2951	0.0500	0.8196	1.0696	0.5998	1.9073
FZD3 (rs2323019)	0.3268	0.4185	0.6100	0.4349	1.3865	0.6105	3.1489
NRG1(SNP8NRG243177)**G72* (2391191)* *FZD3* (rs2323019)	**−1.2835**	**0.4549**	**7.9600**	**0.0048**			
Model (5)							
FZD3 (rs2241802)	−0.1705	0.2907	0.3400	0.5576	0.8432	0.4770	1.4907
FZD3 (rs352203)	0.1094	0.2928	0.1400	0.7086	1.1156	0.6285	1.9804
FZD3(rs2241802)**FZD3* (rs352203)	**−1.9424**	**0.5934**	**10.7200**	**0.0011**			

Note: The values in boldface and italic type are significant ones that indicate interaction effects for the disease

paranoid schizophrenia. Cluster analysis and logistic regression models are used to identify genetic associations of synergistic blocks with the phenotype. The approach is applied to detect the individual and interactive effects of four candidate genes, *NRG1*, *G72*, *RGS4*, and *FZD3*, on paranoid schizophrenia. The results suggest associations between these four genes and schizophrenia and intergenic interaction effects among these four genes on schizophrenia. However, because of the potentially data-driven nature of these conclusions and the limited multi-locus interaction

model used in the simulation part, additional studies are required to confirm the validity of the present method in future studies.

Screening individual and interactive effects of disease genes in complex diseases is a feasible approach using the synergistic block-detecting method. These results further support previous findings about the interactive effects among *NRG1*, *G72*, *RGS4* and *FZD3*, especially via glutamatergic transmission mediated by the ErbB3 and *N*-methyl-D-aspartate (NMDA) receptors (Harrison and Weinberger 2005), the Wnt pathway, or other processes associated with neurodevelopment and plasticity.

In this study, we found that the target SNPs interacted, and we introduced the concept of synergistic blocks. Since there are several disease loci in every synergistic block, we can address the dimensionality problem in multi-locus association analysis, and evaluate the sizes of the interactive effects between loci and their contributions to the disease. For genes that may play a role via similar neuropathological mechanisms, such as the effect of *NRG1*, *G72*, *RGS4* and *FZD3* on schizophrenia via synaptic function or other neurodevelopmental processes, synergistic blocks can be used to identify groups of loci that are specific to the disease and to quantify the interaction between genes or loci.

The synergistic block method described in this paper considered only ten SNPs in the simulation part and 17 SNPs in the real data set. Further numerical investigations will be needed when the number of SNPs to be examined is large, for examples 100 SNPs or 1000 SNPs.

Since genome-wide association studies (Risch and Merikangas 1996; Kang and Zuo 2007) are a priority, there is also the potential for synergistic blocks to be useful on a larger scale. However, we cannot use the present method to investigate genome-wide SNP data directly because of the large number of SNP combinations. One possible strategy is to break up large analyses into roughly independent modules of hundreds of tests (or SNPs) each (Seaman and Müller-Myhsok 2005). If we then detect a synergistic block for each group of SNPs, this synergistic block can be examined by our logistic model. As long as the correlation between the modules of SNPs is reasonably low, little power will be sacrificed by approximating in this way because the synergistic block has accounted for the correlation within the blocks.

Electronic database information

deCODE genetics, http://www.decode.com/nrg1/markers for SNPs and microsatellite markers in *NRG1*.

GenBank, http://www.ncbi.nlm.nih.gov/SNP/ for *NRG1*, *G72*, *RGS4*, *FZD3*.

Acknowledgments We would like to thank the anonymous referees for very helpful comments on the early draft. This work was supported in part by grants from the National Natural Science Foundation of China (No. 30530290, 30400149, 60334040), the National High Technology Research and Development Program of China (No. 2006AA02Z195, 2007AA02Z423), The National Basic Research Program of China (No. 2007CB512301), The National Science Foundation of America (No. DMS 0234078), and the Strategic Partnership Grant of the Michigan Foundation.

References

Agresti A (1996) An introduction to categorical data analysis. Wiley, New York

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57:289–300

Duda RO, Schafer DW (2001) Pattern classification. Wiley, New York

Frankel WN, Schork NJ (1996) Who's afraid of epistasis? Nat Genet 14:371–373

Goldberg DE (eds) (1989) Genetic algorithms in search, optimization, and machine learning, vol 77. Addison-Wesley, New York

Goodman JE, Mechanic LE, Luke BT, Ambs S, Chanock S, Harris CC (2006) Exploring SNP–SNP interactions and colon cancer risk using polymorphism interaction analysis. Int J Cancer 118:1790–1797

Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. Bioinformatics 19:376–382

Harrison PJ, Owen MJ (2003) Genes for schizophrenia? recent findings and their pathophysiological implications. Lancet 361:317–319

Harrison PJ, Weinberger DR (2005) Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol Psychiatry 10:40–68

Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4:701–709

Kang GL, Yue WH, Zhang JF, Cui YH, Zuo YJ, Zhang D (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. J Theor Biol 250:362–374

Kang GL, Zuo YJ (2007) Entropy-based joint analysis for two-stage genome-wide association studies. J Hum Genet 52:747–756

Onay VU, Briollais L, Knight JA, Shi E, Wang Y, Wells S et al (2006) SNP–SNP interactions in breast cancer susceptibility. BMC Cancer 6:114

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150–157

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

Seaman SR, Müller-Myhsok B (2005) Rapid simulation of *p*-values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet 76:399–408

Yue WH, Kang GL, Zhang YB, Qu M, Tang FL, Han YH, Ruan Y, Lu TL, Zhang JF, Zhang D (2007) Association of DAOA polymorphisms with schizophrenia and clinical symptoms or therapeutic effects. Neurosci Lett 416:96–100

Yue WH, Liu ZH, Kang GL, Yan J, Tang FL, Ruan Y, Zhang JF, Zhang D (2006) Association of G72/G30 polymorphisms with early-onset and male schizophrenia. Neuroreport 17:1899–1902

Zhang YB, Yu X, Yuan YB, Ling YS, Ruan Y, Si TM, Lu TL, Wu SP, Gong XH, Zhu ZJ, Yang JZ, Wang F, Zhang D (2004) Positive association of the human Frizzled 3 (FZD3) gene haplotype with schizophrenia in Chinese Han population. Am J Med Genet 129B:16–19