Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 250 (2008) 362–374
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

An entropy-based approach for testing genetic epistasis underlying complex diseases

Guolian Kang[a], Weihua Yue[b], Jifeng Zhang[c], Yuehua Cui[a], Yijun Zuo[a], Dai Zhang[b,*]

[a]Department of Statistics and Probability, East Lansing, Michigan State University, MI 48824, USA
[b]Key Laboratory for Mental Health, Ministry of Health, Institute of Mental Health, Peking University, Beijing 100083, China
[c]Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China

Received 11 June 2007; received in revised form 29 September 2007; accepted 1 October 2007
Available online 6 October 2007

Abstract

The genetic basis of complex diseases is expected to be highly heterogeneous, with complex interactions among multiple disease loci and environment factors. Due to the multi-dimensional property of interactions among large number of genetic loci, efficient statistical approach has not been well developed to handle the high-order epistatic complexity. In this article, we introduce a new approach for testing genetic epistasis in multiple loci using an entropy-based statistic for a case-only design. The entropy-based statistic asymptotically follows a χ^2 distribution. Computer simulations show that the entropy-based approach has better control of type I error and higher power compared to the standard χ^2 test. Motivated by a schizophrenia data set, we propose a method for measuring and testing the relative entropy of a clinical phenotype, through which one can test the contribution or interaction of multiple disease loci to a clinical phenotype. A sequential forward selection procedure is proposed to construct a genetic interaction network which is illustrated through a tree-based diagram. The network information clearly shows the relative importance of a set of genetic loci on a clinical phenotype. To show the utility of the new entropy-based approach, it is applied to analyze two real data sets, a schizophrenia data set and a published malaria data set. Our approach provides a fast and testable framework for genetic epistasis study in a case-only design.

Keywords: Case-only design; Complex diseases; Entropy; Genetic epistasis; Genetic network

1. Introduction

One of the major goals of modern biomedical research is to understand the function of genes underlying physical manifestation of an organism (Macdonald and Long, 2005). To achieve this objective, studies have long-time been framed within the content of discovering associations between genetic markers and biological phenotypes (Judson et al., 2002). Due to unknown disease etiologies and complex heterogeneities, however, searching for causative genes underlying complex diseases has not been quite successful. The complicated functional mechanism among genes underlying complex diseases presents great challenges. For example, Strohman (2002) recently reported that human disease phenotypes are influenced not only by DNA variations but also by self-organizing networks and system dynamics. Understanding how genomic information underlies the development of complex human diseases, thus, has been one of the greatest challenges in the 21st century (Zhao et al., 2006). The availability of complete human sequence information, relatively cost-efficient high-throughput genotyping technologies, and powerful statistical methods, provide promising future in unravelling the genetic secrets of complex human diseases.

It is well known that most human diseases are complex which are typically caused by multiple factors, including main effects of multiple genes, complicated gene–gene as well as gene–environment interactions (Zhao et al., 2006). Gene–gene interaction, or epistasis, plays a pivotal role in shaping an organism development, as well as contributing to a complex disease. Examples of epistasis have been identified. For example, a group of scientists recently found that the interaction in mutations between RET and

*Corresponding author. Tel.: +86 10 82801937; fax: +86 10 62078246.
E-mail address: daizhang@hsc.pku.edu.cn (D. Zhang).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.10.001

EDNRB genes was associated with Hirschsprung diseases (Carrasquillo et al., 2002). For another example, studies employing model organisms such as *Drosophilla melanogaster* and *Saccharomyces cerevisiae* (yeast) have suggested that epistasis occurs frequently, involving multiple loci (genes), and in some cases produces effects as large as the main effects at individual loci (Brem et al., 2005). While identifying gene interactions using biological techniques is time and money-consuming, statistical approaches have been proven to be alternative efficient tools for elucidating the interaction mechanism in a variety of settings (Nelson et al., 2001; Ritchie et al., 2001; Moore and Hahn, 2002; Soares et al., 2005).

Case-control design has been the most commonly used design in genetic association studies. It can also be applied to test gene–gene interaction. Recent research has shown that gene–gene interaction can also be tested with a case-only design (Yang et al., 1999). One of the advantages of case-only design is that it requires fewer sample size than a case-control design for testing genetic epistasis (Gauderman, 2002), which consequently eliminates estimation biases when selecting controls in a case-control design and can potentially save resources. One commonly applied approach for testing epistasis in a case-only design is to apply a simple χ^2 test (Gauderman, 2002). When a particular disease is associated with several clinical phenotypes (i.e., clinical symptoms) which are the results of complex interactions among a group of genes, the χ^2 test can not serve as the purpose to identify which set of disease loci contributing to a clinical phenotype. For instance, the schizophrenia disease is a complex disorder which shows several clinical phenotypes such as delusions or conceptual disorganization (Andreasen, 2000). The variation of each clinical phenotype could be explained by a unique set of genes functioning in a complex epistatic way. The χ^2 test may not be appropriate for testing such an association. To enhance the power of testing epistasis under a case-only design framework, in this paper, we introduce a new entropy-based approach in a case-only design for testing gene–gene interactions under the assumption of linkage equilibrium (LE) among loci. The entropy is commonly used in information theory to measure the uncertainty of random variables (Shannon, 1948) and has been applied to different subjects in biology (Ackerman et al., 2003; Hampe et al., 2003; Jawaheer et al., 2002; Kang and Zuo, 2007; Mihalek et al., 2004; Zhao et al., 2005) and other fields (Kang et al., 2007; Kubat et al., 2007). An entropy measure represents a non-linear transformation of interested variables. We derive an entropy-based statistic to test gene interactions. The entropy-based test is further extended to test the association between a set of disease loci and a specific clinical phenotype. We deal with multiple interactive loci as a locus-system, with joint genotypes as its microstates and clinical phenotypes as its macrostates, a similar idea presented in statistical physics (Cover and Thomas, 1991). The interactions among multiple loci can be defined as the deviance of the entropy at one locus-system from that of the same locus-system assuming independence. The type I error rate and power of the new entropy approach are evaluated through Monte Carlo simulations and are compared with the standard χ^2 test. Finally, we apply the new approach to two real data sets. The advantages and limitations of the entropy-based approach are discussed.

2. Methods

2.1. Entropy and entropy epistasis measure

The Shannon entropy (S) of a discrete random variable X is defined as

$$S(X) = -\sum_i p(x_i)\log p(x_i), \tag{1}$$

where $p(x_i) = Prob(X = x_i)$. Define a set of genes or loci as a genetic-locus system which is referred to as a "microstate", a term used in statistical physics to measure an individual state of a system, such as a particular path taken by a random walk (Greiner et al., 1995). Correspondingly, a particular disease status can be defined as a "macrostate" which is the physical manifestation of the complex interactions among a group of disease genes (or microstate). Therefore, by measuring the entropy of a set of genes, we can test the epistasis associated with a particular disease.

It is well known that the equilibrium state of a system is the state that all microstates in this system have the same chance to occur (Greiner et al., 1995). For a genetic-locus system, an equilibrium state refers to independence among disease loci. Under the assumption of linkage equilibrium, the system is likely to maintain an equilibrium state with maximum entropy. A deviation from the equilibrium state (independence) represents a gain in order structure, which consequently results in decreased entropy compared to the equilibrium state. Therefore, if the difference between the state of independence and the state due to interaction in entropy for a set of loci is significant, a conclusion of significant interaction can be reached. This is the basic idea that motivates us to derive the following entropy-based test procedure. To illustrate the idea, we first start with two loci. A generalization is given later. We refer to a subset of disease loci as a locus system with joint genotypes as its microstates. Assume two loci in linkage equilibrium with M_1 having two allele A and a, M_2 having two alleles B and b. At locus M_1, there are three possible genotypes denoted as AA (2), Aa (1) and aa (0) with frequencies denoted as p_2^1, p_1^1 and p_0^1. Similarly there are three genotypes with frequencies p_2^2, p_1^2 and p_0^2 at locus M_2. Here the superscript denotes the identification of each locus and the subscript denotes different genotypes at a locus. The genotypic combination at these two loci forms 3^2 joint genotypes expressed as $h_1 = 22$, $h_2 = 21$, $h_3 = 20$, $h_4 = 12$, $h_5 = 11$, $h_6 = 10$, $h_7 = 02$, $h_8 = 01$ and $h_9 = 00$. The frequencies of these nine joint genotypes are denoted by p_1, p_2, p_3, p_4, p_5,

p_6, p_7, p_8 and p_9, respectively. Thus, under the null hypothesis of no interaction, the frequency of joint genotype is just a product of their corresponding genotypes. For example, the frequency of the joint genotype $h_1(=22)$ is expressed as $q_1 = p_2^1 p_2^2$. So, the entropy under the null hypothesis of no interaction for two disease loci is defined as $S_{ind} = -\sum_{i=1}^{9} q_i \log q_i$. And the observed entropy for two disease loci is $S_{observe} = -\sum_{i=1}^{9} p_i \log p_i$.

Next, we generalize the entropy measure to multiple loci, each with two alleles. Suppose there are $s \geqslant 2$ disease loci, with each locus consisting of three possible genotypes denoted as 0, 1 or 2. This locus system with s disease loci forms 3^s joint genotypes denoted as $h^i = (h_1^i, h_2^i, \ldots, h_s^i) \in \{0,1,2\}^s$, where h_k^i $(= 0, 1$ or $2)$ denotes the genotype of the kth $(1 \leqslant k \leqslant s)$ disease locus at the ith joint genotype, $1 \leqslant i \leqslant 3^s$.

Let us define

$$x(i,j,k) = \begin{cases} 1 & \text{if } h_k^i = j, \\ 0 & \text{if } h_k^i \neq j, \end{cases} \tag{2}$$

where $j = 0, 1$ or 2. Under the null hypothesis of no interaction, the frequency of the ith joint genotype can be expressed as the product of the marginal genotype frequencies,

$$q_i = p_{h^i} = \prod_{k=1}^{s} p_{(k,2)}^{x(i,2,k)} p_{(k,1)}^{x(i,1,k)} p_{(k,0)}^{x(i,0,k)}, \tag{3}$$

where $q_i = p_{h^i}$ denotes the frequency of the ith joint genotype under the null hypothesis of no interaction, $p_{(k,0)} = 1 - p_{(k,2)} - p_{(k,1)}$, $p_{(k,g)}$ denotes the frequency of genotype g at the kth disease locus.

The entropy under the null hypothesis of no interaction for s disease loci is expressed as $S_{ind} = -\sum_{i=1}^{3^s} q_i \log q_i$. Note that missing joint genotypes with frequencies of zero do not contribute to the entropy S. Also for rare genotype, q_i approaches zero which leads to $q_i \log q_i$ approaching zero, and consequently rare genotypes do not contribute to the entropy measure S too.

Similarly, the observed entropy of a set of SNP markers can be expressed as $S_{observe} = -\sum_{i=1}^{3^s} p_i \log p_i$, where p_i is the observed frequency of the ith joint genotype. If there are no interactions among multiple disease loci, the difference between the state of independence and the state of interaction in entropy will be zero.

2.2. The entropy-based statistic for epistasis test

We focus our attention in this section to derive an entropy-based statistic for testing interactions among multiple disease loci. A measure for the system's deviation from the equilibrium state (i.e., no interaction) is given as

$$\Delta S = S_{ind} - S_{observe}. \tag{4}$$

Therefore, rejection of the null hypothesis $H_0 : \Delta S = 0$ indicates interaction among disease loci.

To quantify the magnitude of interaction among multiple disease loci in a normalized scale, we define a new measure (I) among these s loci as $I = 1 - \frac{S_{observe}}{S_{ind}}$. Clearly, absence of interaction among s loci, i.e., $S_{ind} = S_{observe}$, leads to $I = 0$ which provides the same information as $\Delta S = 0$. A plot of I against a measure such as allele frequency can provide useful information about a systematic derivation from independence to interaction in entropy at a locus system. It can be shown that $2n^D \Delta S \sim \chi^2_{3^s-2s-1}$ holds under the null hypothesis of no interaction, where n^D is the sample size for affected individuals. This statistic $2n^D \Delta S$ is referred to as the "entropy-based statistic". Let $L_{observe}$ denote the likelihood of the observed joint genotypes of a locus system, and L_{ind} denote the likelihood of the joint genotypes defined by the marginal frequencies under no interaction, where $L_{observe}$ and L_{ind} can be derived based on a multinomial distribution. Then it can be shown that (see Appendix A for details) $\Delta S = \frac{1}{n^D} \log \frac{L_{observe}}{L_{ind}}$. Following the result of Wilks (1962), we get $2n^D \Delta S \sim \chi^2_{3^s-2s-1}$.

2.3. The relative entropy-based test for testing the association between disease loci and clinical phenotypes

Generally, a complex disease may show many symptoms in clinic. For example, conceptual disorganization and delusions are two symptoms for a person diagnosed to have schizophrenia. These clinical symptoms are referred to as clinical phenotypes in this section, and may be associated with different sets of genes as well as interactions among these genes. In this section, we derive a relative entropy-based test statistic to quantify and further test the association between clinical phenotypes and disease loci.

From the previous section, we know that the joint genotypes among a set of disease loci in a locus system can be considered as microstates. Based on the entropy theory, a clinical phenotype can be viewed as a macrostate corresponding to a disease locus system. Therefore, we can link these two states and treat a macroscopical clinical phenotype as a functional result of many loci or genes (many microstates). Study shows that many schizophrenia individuals with positive syndromes almost share the same positive clinical symptoms (Andreasen et al., 1994). But how to test which set of genes (or loci) interact to contribute to an associated clinical phenotype remains unclear in literatures.

A clinical phenotype is often measured as categorical data to indicate the severity of a symptom. For example, a score of 0 might indicate no symptom and high scores might indicate severe cases. It also could be measured as binary with 0 and 1 indicating no symptom and symptom, respectively. Our testing model will be derived based on binary clinical phenotypes. For a categorical clinical phenotype, we can code it as 1 if a score is greater than or equal to 1, and 0 otherwise.

Consider s disease loci and a clinical phenotype Ψ. To test the association, we start with a subset with w disease loci as a locus system. Let $\{h^i\}_{i=1}^{3^w}$ be 3^w joint genotypes in this locus system, K^D be the sample size with a clinical phenotype Ψ, and k_i^D be the number of individuals with joint genotype h^i in K^D.

The relative entropy of an associated clinical phenotype Ψ is defined as

$$S(\Psi)=\begin{cases}\dfrac{-\sum_{i=1}^{W}\frac{k_i^D}{K^D}\log\frac{k_i^D}{K^D}}{-\sum_{i=1}^{W}\frac{1}{W}\log\frac{1}{W}} & \text{if } W>1,\\ 0 & \text{if } W=1,\end{cases} \tag{5}$$

where $W\leqslant 3^w$ is the number of actually observed joint genotypes in this locus system, w is the number of loci in this locus-system. The maximum entropy is obtained when all joint genotypes are uniformly distributed with frequency of $1/W$. Therefore, if all joint genotypes in this locus system are evenly distributed (this system is in the equilibrium state), then $S(\Psi)=1$ which corresponds to the null hypothesis of no association. Otherwise, if a single joint genotype is far more frequent than other joint genotypes, then $S(\Psi)$ approaches 0. Therefore, the smaller the relative entropy, the higher the chance that a clinical phenotype is associated with these w set of disease loci.

In fact, the relative entropy of a clinical phenotype is related to the likelihood L of a multinomial distribution of joint genotypes in this locus system. The relationship between L and $S(\Psi)$ can be shown as (see Appendix B for details) $L=\mathrm{e}^{-K^D S(\Psi)\log W}$. Clearly, the likelihood distribution of the locus system increases as the relative entropy of a clinical phenotype decreases. This satisfies the maximum likelihood principle. So, we identify a subset of loci with maximum likelihood value (or the minimum relative entropy) as a set of loci that interact to contribute to this clinical phenotype Ψ.

To test the association among w loci associated with a clinical phenotype Ψ, we introduce a new relative entropy-based statistic which has the form

$$E^P=2K^D(1-S(\Psi))\log W. \tag{6}$$

It can be easily shown that the test statistic E^P is asymptotically distributed as a central χ^2_{W-1} distribution under the null hypothesis of no association.

To search for subsets of disease loci associated with a clinical phenotype, we use a sequential forward selection approach. The algorithm starts with one disease locus to compute the test statistic E^P for all possible combinations at the w loci and test the association. There might be several set of combinations significantly associated with a clinical phenotype. Those significant combinations in the earlier step are then kept in the model and another locus is added to the model for association test. The search is stopped when an additional adding has no significant contribution to the test statistic. Finally, we may have one or more locus combinations significantly associated with a clinical phenotype. This can be explained by the locus heterogeneity of a complex disease in which the interaction pattern among different loci might have different contributions to a clinical phenotype. We can then identify the most commonly appeared loci among a set of significant combinations. This set of loci is defined as the most functional ones interacting with other loci to contribute to a clinical phenotype. A sequential interactive network structure can be displayed using a tree diagram.

3. Results

3.1. Monte Carlo simulation

In this section we conduct Monte Carlo simulations to demonstrate the entropy test approach. For more information about simulations based on entropy, readers are referred to the literatures (Carlacci and Chou, 1990; Zhang and Chou, 1992, 1995). We consider two disease loci (i.e. $s=2$) and assume that they are in LE. Let A and B be the two risk alleles at the first and second disease loci, with frequencies P_A and P_B, respectively. Let g_A and g_B denote the genotypes at the first and second disease loci, respectively, i.e., $g_A, g_B\in\{0,1,2\}$. These two-locus genotypes are simply denoted by g_Ag_B with frequency denoted as $P_{g_Ag_B}$. Let $f_{g_Ag_B}$ be the penetrance for genotype g_Ag_B. Then, the disease prevalence is defined by

$$\begin{aligned}P(D)=&P_A^2P_B^2f_{00}+2P_A^2P_BP_bf_{01}+P_A^2P_b^2f_{02}\\&+2P_AP_aP_B^2f_{10}+4P_AP_aP_BP_bf_{11}+2P_AP_aP_b^2f_{12}\\&+P_a^2P_B^2f_{20}+2P_a^2P_BP_bf_{21}+P_a^2P_b^2f_{22},\end{aligned} \tag{7}$$

where $f_{g_Ag_B}$ can be obtained by the disease model (Table 1). So, the frequencies of genotype g_Ag_B in cases (Appendix C) can be expressed as

$$P(g_Ag_B|D)=\frac{f_{g_Ag_B}P_{g_Ag_B}}{P(D)},\quad 0\leqslant g_A,\ g_B\leqslant 2. \tag{8}$$

Table 1
Two-locus models of disease

	BB	*Bb*	*bb*
Model 1: *two-locus multiplicative interaction*			
AA	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^2$	γ
Aa	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	γ
aa	γ	γ	γ
Model 2: *two-locus threshold interaction*			
AA	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
Aa	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
aa	γ	γ	γ

Model 1 refers to a two-locus multiplicative. In this model, the odds of disease have a baseline value (γ) and increase multiplicatively once there is at least one disease allele at each disease locus. Model 2 refers to a two-locus threshold model. In this model, the odds of disease also have a baseline value (γ) unless a disease allele is present at each locus. Once this threshold is reached, the odds of disease increase to ($\gamma(1+\theta)$), θ is genotypic effect (6).

A total of 10,000 simulations are performed. Genotype data for multiple loci ($s>2$) can be simulated in similar way.

3.1.1. Distributions of the test statistics and interaction measure

In the previous section, we have shown that when the sample size approaches infinity to apply asymptotic theory, the distribution of the entropy-based statistic $2n^D\Delta$ under the null hypothesis asymptotically follows a central χ^2 distribution. To examine the small sample performance of the entropy-based test statistic under the null hypothesis of no interaction, we generate 150 cases at random with genotypes simulated according to the distribution of (8).

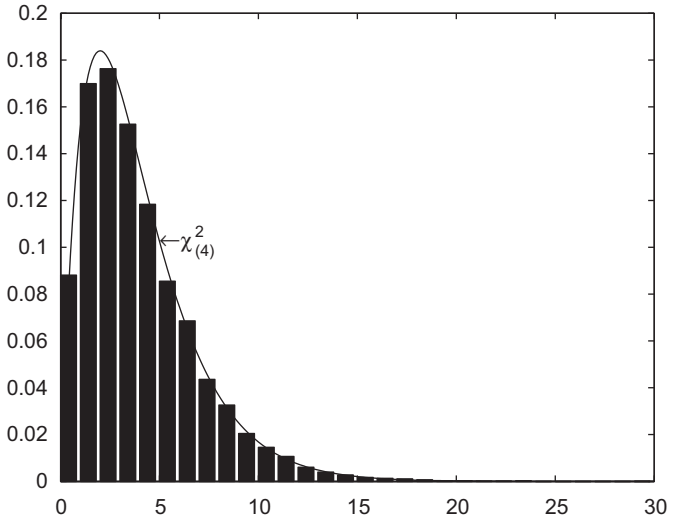

Fig. 1. Null distributions of the test statistic $2n^D\Delta S$ from simulated 150 case individuals. $\chi^2_{(4)}$ indicates χ^2 distribution with 4 df.

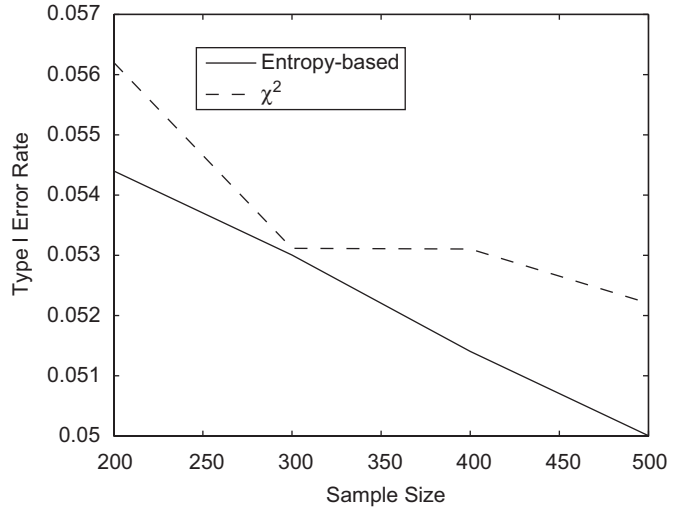

Fig. 2. Type I error rates of the test statistics $2n^D\Delta S$ and χ^2 in testing interaction between two disease loci. The type I error is calculated as the percentage of the number of simulations in which no significant interactions are detected. The result is obtained from 10,000 simulations.

Fig. 1 plots the histograms of the test statistic $2n^D\Delta S$ for a two-SNP interaction model. It is clear that the distribution of $2n^D\Delta S$ is similar to the asymptotic χ^2 distribution with 4 degree of freedom. We further compare the estimated type I error rate of the entropy-based test and the standard χ^2 test under different sample sizes for testing interaction (Fig. 2). The plot shows that the estimated type I error rates (at the significant level 0.05) of the proposed entropy-based test are close to the nominal level as sample size increases. Also, the new test has slightly better type I error control than the standard χ^2 test.

Fig. 3 plots the degree of interaction (I) between two loci as a function of allele frequencies (A) and genotypic effects (B). The dashed line refers to the multiplicative model and the solid line is for the threshold model (Table 1). The

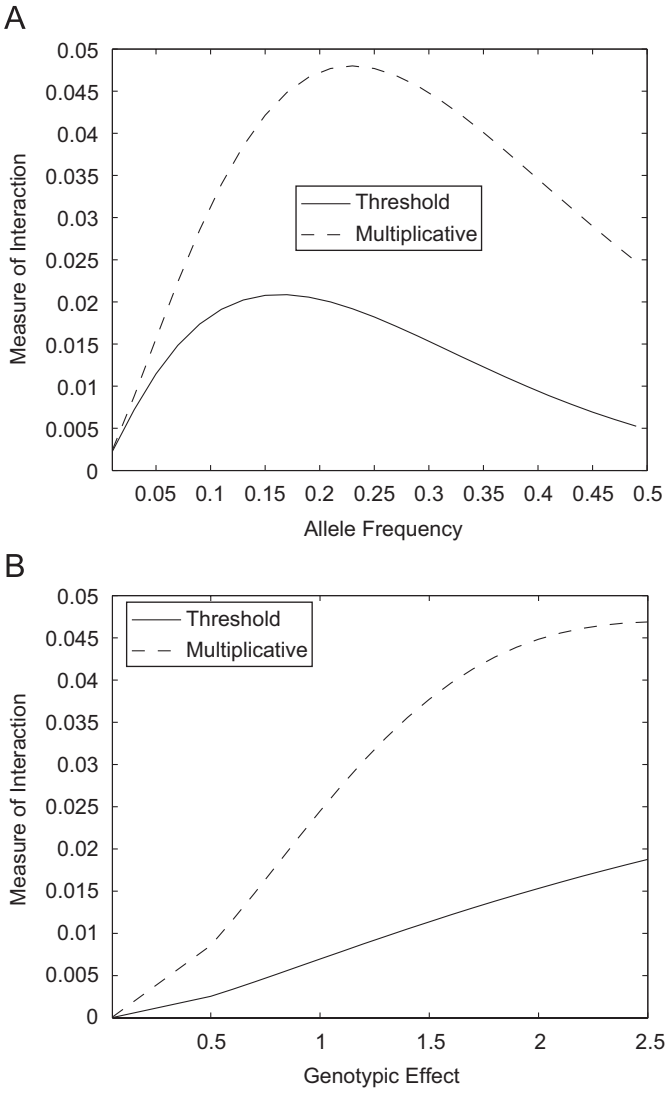

Fig. 3. Measure of interaction between two loci as a function of allele frequencies at two loci under two genetic models, where the baseline and the genotype effect are 0.01 and 2, respectively (A); and as a function of genotypic effects at two loci under the conditions that the minor allele frequencies at two loci are 0.3 and the baseline effect is 0.01 (B).

baseline effect is defined as $\gamma = 0.01$ and the baseline genotypic effect is given as $\theta = 2$. The allele frequencies in Fig. 3A are the minor allele frequencies for alleles A and B. The results show that the interactions are stronger for the multiplicative model than the threshold model under different situations. The measure of interaction is a monotonic function of the genotypic effects. However, when the genotype effect is fixed, it shows a quadratic function of allele frequency. Therefore, the measure of interaction depends on both disease models and allele frequencies at two loci.

To check the asymptotic distribution of the test statistic E^P for testing association between a set of genetic loci and a clinical phenotype Ψ, we simulate 250 cases which have a common clinical phenotype. Fig. 6 plots the histograms of the test statistic E^P for testing association between two loci and a clinical phenotype Ψ under the condition that the number of present joint genotypes are 6 and 9, respectively. It can be seen that the distributions of the test statistic E^P are similar to the theoretical central χ^2 distribution.

3.1.2. Power evaluation

To evaluate the performance of the proposed statistic, we compare the power of entropy-based statistic with that of the standard χ^2. Fig. 4 plots the power to detect interaction for two loci as a function of sample size for fixed genotypic and baseline effects (A and B) and as a function of genotypic effect for fixed sample size and baseline effect (C and D). The minor allele frequency is fixed at 0.25 at two loci under the two genetic models, multiplicative and threshold. It can be seen that the sample size has a great impact on the testing power. The power increases as sample size increases for both the entropy-based and χ^2 test. For example, the power is about 50% when sample size is 100 and it increases to nearly 90% when sample size increases to 200 for the multiplicative model (Fig. 4A). Similar trend is also observed for the threshold model. We also observed a consistent trend that the entropy-based test outperforms the standard χ^2 test. The effect of genetic models on testing power can also be clearly seen from the plot in which the multiplicative model always has high power than the threshold model for fixed sample size and genotypic effect. For example, when sample size is fixed at 200, the multiplicative model has almost 90% power, while the threshold model only has 50% power.

Intuitively, the degree of genetic interaction measured by I should have a direct effect on the testing power, where we expect high power to test the interaction when I is large. Fig. 5 shows the effect of I on the power. We fix the allele frequency (0.10 in Fig. 5A and 0.30 in Fig. 5B) and the baseline effect, and change I by adjusting the genotypic effect. It is clearly seen that as the measure of interaction increases, the power increases. The multiplicative model displays higher power than the threshold model. These results are consistent with the underlying model since the multiplicative model has stronger interaction effect than the threshold model given positive genotypic effect (Table 1).

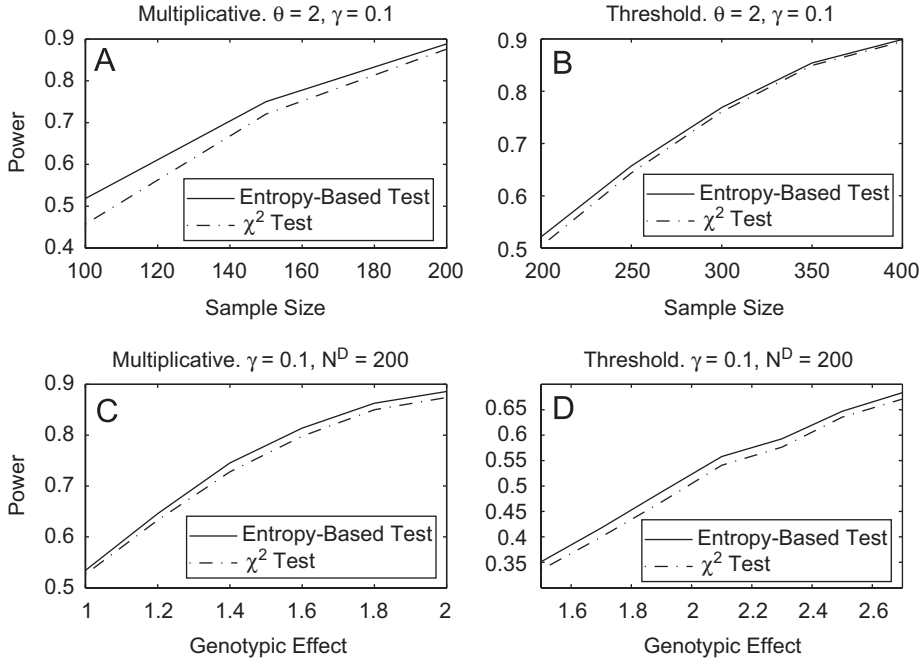

Fig. 4. Power of the test statistic $2n^D\Delta S$ as a function of sample sizes for fixed genotypic and baseline effects assuming multiplicative model (A) and threshold model (B), and as a function of genotypic effects for fixed baseline effect and sample size assuming multiplicative model (C) and threshold model (D) between two loci. The minor allele frequency at two loci in the disease population is assumed to be 0.25.

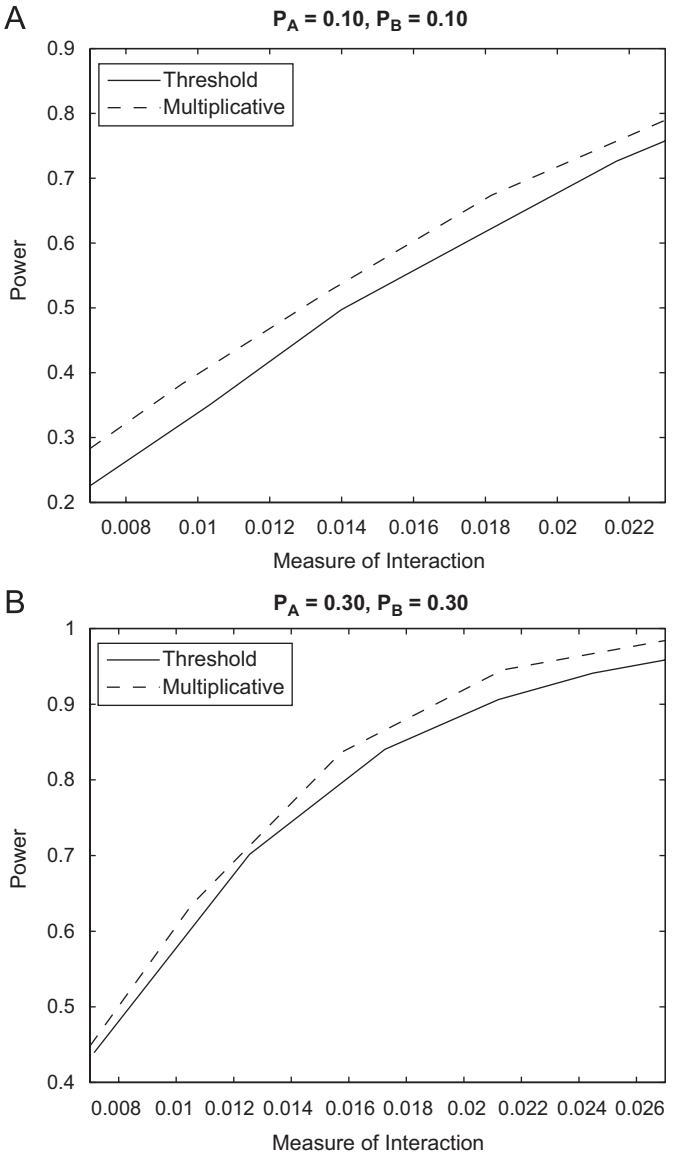

Fig. 5. Power of the test statistic $2n^D\Delta S$ as a function of the interaction measure between two loci under two genetic models, multiplicative and threshold. The minor allele frequency at two loci are 0.10 (A) and 0.30 (B). The significance level is 0.05 and the sample size is 200.

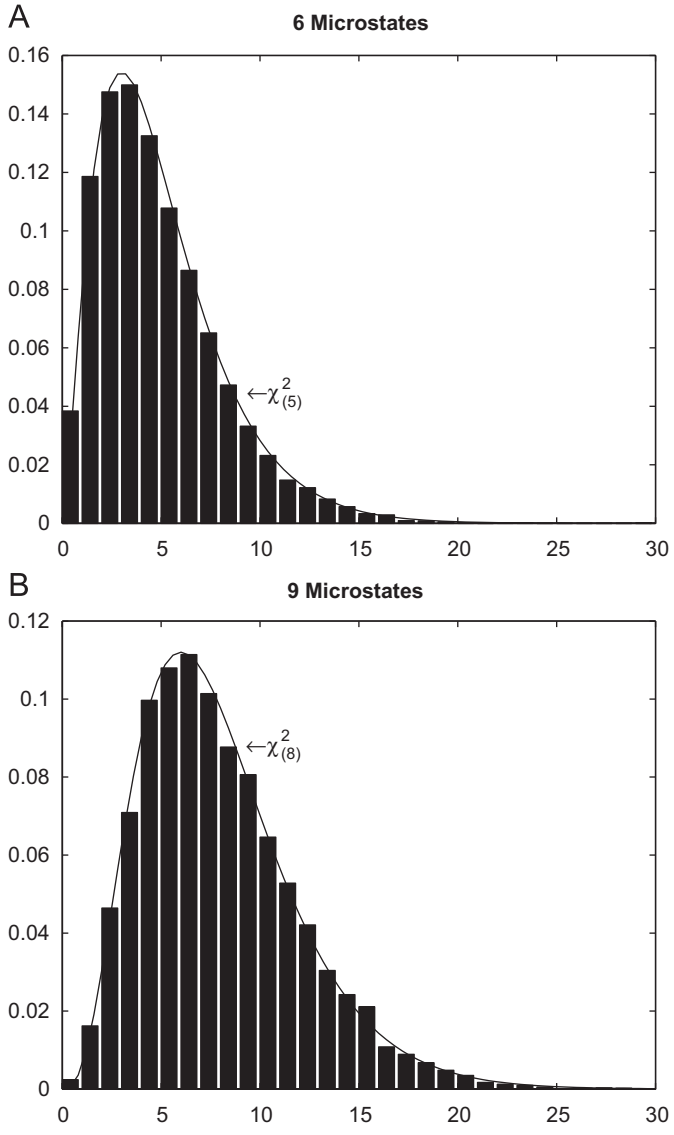

Fig. 6. Null distributions of the test statistic E^P with 250 simulated individuals having a clinical phenotype under the condition that the numbers of present microstates in two-locus system are 6 and 9, respectively. $\chi^2_{(5)}$ and $\chi^2_{(8)}$ indicate χ^2 distribution with 5 and 8 df.

In a short summary, our entropy-based test has better control of the type I error rate and has higher power compared to the standard χ^2 test for testing genetic interactions under a number of situations. Our simulation results also confirm that the asymptotic distribution of the proposed test statistic follows a χ^2 distribution when testing genetic interactions, as well as testing the association between disease loci and clinical phenotypes under the null hypothesis (see Fig. 6). The sample size, allele frequency, genotypic effect, degree of interaction and genetic models all have impacts on the power. The power increases as the degree of interaction increases and consistent higher power is observed for the multiplicative model than the threshold model.

3.2. Application to real data examples

3.2.1. Schizophrenia data

To show the utility of the proposed entropy-based test, we apply it to two real data sets. In the first data set, three genes, namely neuregulin 1 (*NRG1*, 8p22-p11, MIM 142445), *G72* (13q34, MIM 607408) and regulator of G-protein signaling-4 (*RGS4*, 1q21-q22, MIM 602516), are thought to converge functionally upon schizophrenia by influencing synaptic plasticity and the cortical microcircuitry (Harrison and Weinberger, 2005; Yue et al., 2007). A total of 13 SNPs, including 7 from gene *NRG1*, 3 from gene *G72* and 3 from gene *RGS4*, are genotyped in 339 schizophrenia patients and 339 matched controls in a

Chinese Han population. All SNPs are in Hardy–Weinberg Equilibrium (HWE). Prior reports suggest that variations in these genes might increase the risk of developing schizophrenia, and hence these three genes serve as candidate genes for an association study.

Seven of the 13 SNPs are identified to be associated with schizophrenia by using standard χ^2 test, including three functional polymorphic markers *NRG1* (rs3924999), *NRG1* (rs3735774) and *G72* (rs2391191); two intronic SNPs, *NRG1* (rs2919390) and *NRG1* (rs6988339); and one 3′UTR SNP, *RGS4* (rs10759); and one 5′UTR SNP, *NRG1* (SNP8NRG221533). Two functional polymorphic markers *NRG1* (rs3924999) and *NRG1* (rs3735774) are known to encode the glial growth factor (GGF2) and the sensory and motor neuron-derived factor (SMDF), respectively. The inactivation of SMDF cause marked neuronal abnormalities as it might act as "glial growth factors" (Kirov et al., 2005). All the three SNPs within gene *NRG1* are in linkage equilibrium (data not shown). Both analysis about interaction among SNPs and association analysis between disease loci and clinical phenotypes in the following focus on these seven SNPs within three genes. Table 2 shows the *P* values of the entropy-based statistic for testing the interactions between two SNPs within schizophrenia-associated gene *NRG1*. For comparison, Table 2 also includes the *P* values for standard χ^2-statistic. It is evident that *P* values of the entropy-based test are smaller compared to those of the standard χ^2 test.

Table 2
Interaction test between two SNPs at *NRG1*

Interaction marker pair	*P* value for $2n^D\Delta S$	*P* value for χ^2
rs3924999 and rs3735774	0.032	0.045
rs3924999 and rs2919390	0.002	0.002
rs3735774 and rs2919390	0.005	0.020
rs3735774 and rs6988339	0.014	0.037
rs2919390 and rs6988339	1.81E−05	2.69E−05

We start with interaction test for two SNPs and gradually increase the number of SNPs in the model. Table 3 shows the result of three-way and four-way interactions among these three genes, *NRG1*, *RGS4* and *G72*. The interaction results indicate that there is a complex interaction network structure among these three genes. The interaction effects between two functional polymorphic markers *NRG1* (rs3924999) and *NRG1* (rs3735774) are well characterized through the test (Table 3). Another significant finding is that there are significant interactions between gene *NRG1* and genes *RGS4* and *G72*, which has been identified and validated by real experiments (Kirov et al., 2005; Thaminy et al., 2003). These results further indicate the power and robustness of the proposed approach. We list all the significant disease loci combinations that contribute to a clinical phenotype (Supplementary). There are total 30 clinical phenotypes. The SNPs are listed sequentially as the order they significantly entered into the model. From the table, we can construct the interaction network that associates with a clinical phenotype. The most important functional loci can be easily seen from the table. For example, *NRG1* (rs3735774) is the most important one associated with the clinical phenotype "delusions" from a sequential point of view. To further clearly show the interaction network, we can construct interaction tree-diagrams for each clinical phenotype. Here we choose the first clinical phenotype to demonstrate the idea. Fig. 7 shows the interaction network in which *NRG1* (rs3735774) is the single SNP showing significance for a single SNP test. Then, we keep *NRG1* (rs3735774) in the model and add another SNP, which leads to *G72* (rs2391191) and *NRG1* (rs3924999) significantly. Keep adding more loci, we get the sequential tree structure. It can be easily seen from the graph that *NRG1* (rs3735774) is the major locus that interacts with other loci to affect the phenotype "delusions". We also see two paths containing the same set of loci, namely *NRG1* (rs3735774)–*NRG1* (rs3924999)–*G72* (rs2391191) and *NRG1* (rs3735774)–*NRG1* (rs3924999)–*G72* (rs2391191). This information indicates

Table 3
Interaction tests among *NRG1*, *RGS4* and *G72*

Schizophrenia-associated SNPs with interactive effects		$2n^D\Delta S$	*P* value
Three SNPs			
NRG(SNP8NRG221533),	*NRG1* (rs3924999), *NRG1* (rs2919390)	38.130	0.0085
G72 (rs2391191),	*NRG1* (rs3924999), *NRG1* (rs2919390)	33.293	0.0313
G72 (rs2391191),	*NRG1* (rs2919390), *NRG1* (rs6988339)	43.028	0.0020
NRG1 (rs3924999),	*NRG1* (rs3735774), *NRG1* (rs2919390)	42.686	0.0022
NRG1 (rs3924999),	*NRG1* (rs2919390), *NRG1* (rs6988339)	54.194	5.41E−05
NRG1 (rs3735774),	*NRG1* (rs2919390), *NRG1* (rs6988339)	50.338	1.98E−04
NRG1 (rs3924999),	*NRG1* (rs2919390), *RGS4* (rs10759)	36.109	0.0149
NRG1 (rs3924999),	*NRG1* (rs6988339), *RGS4* (rs10759)	33.075	0.0331
NRG1 (rs2919390),	*NRG1* (rs6988339), *RGS4* (rs10759)	41.346	0.0034
Four SNPs			
NRG(SNP8NRG221533),	*NRG1* (rs3924999), *NRG1* (rs2919390), *NRG1* (rs6988339)	103.27	0.0092
NRG(SNP8NRG221533),	*NRG1* (rs2919390), *NRG1* (rs6988339), *RGS4* (rs10759)	98.949	0.0193
NRG1 (rs3924999),	*NRG1* (rs2919390), *NRG1* (rs6988339), *RGS4* (rs10759)	107.200	0.0045

that the final interaction sets are not affected by the order of loci added to the model. It is interesting to note that for all three-SNP interacting patterns detected, the highest entropy is always obtained with the order of 4–3–2. This information shows that SNP 4 (*NRG1*-rs3735774) is the most important one in determining most clinical phenotypes of schizophrenia disease followed by SNP 3 (*NRG1*-rs3924999) and SNP 2 (*G72*-rs2391191). If we slightly change the selection procedure by only keeping those patterns with highest entropy measure at each selection step, this pattern will be the one left in the end. Since SNP 4 and 3 are two known functional SNPs in determining schizophrenia disease, this piece of information provides an indirect support of the approach.

3.2.2. *Malaria data*

The second data set is related to a birth cohort study that recorded the incidence of hospital admission with malaria and severe malaria from Kilifi District Hospital on the coast of Kenya in Africa (Williams et al., 2005). A total of 2104 children were genotyped for both hemoglobin (Hb) and α^+-thalassemia genes to test their interaction. The data set was analyzed by using a Poisson regression analysis performed by Williams et al. (2005). We particularly choose this data set in a purpose to see if our approach gives similar answer. We applied the entropy-based statistic to this data set to test interaction between the Hb and α^+-thalassemia genes. The results are summarized in Table 4. For comparison, Table 4 also lists *P* values obtained by using Poisson regression analysis (Williams et al., 2005). Both approaches end up with the same interaction result between hemoglobin (Hb) and α^+-thalassemia genes. The *P* values of the entropy-based statistic are comparable to those of the Poisson regression analysis.

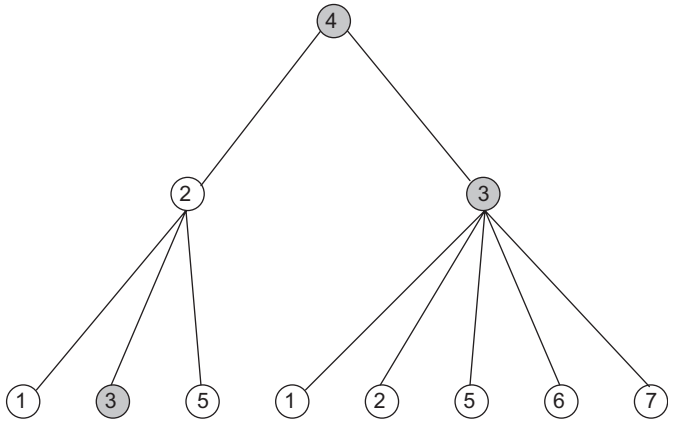

Fig. 7. The diagram of the interaction network that contributes to a positive clinical symptom, delusions of schizophrenia, where 1 = *NRG1* (SNP8NRG221533), 2 = *G72* (rs2391191), 3 = *NRG1* (rs3924999), 4 = *NRG1* (rs3735774), 5 = *NRG1* (rs2919390), 6 = *NRG1* (rs6988339), 7 = *RGS4* (rs10759); 3 and 4 are two functional polymorphic markers.

4. Discussion

Complex diseases may be linked to more than one chromosomal region and may be associated with more than one gene. They are likely to be controlled by a complex genetic mechanism, with minor-to-moderate effect size per gene (Schaid et al., 2005). To consider these problems, multiple-stage strategies should be applied. First, we can identify the main effects of genes with moderate or large effect size; then we may focus on the complex interaction detection among genes with small effect size. Subsequently, we try to detect gene–environment interactions. The purpose of this article is to present a new framework for the identification of interactions among multiple disease loci and mapping the association between disease loci and clinical phenotypes.

There are two ways to increase the power of an association test. One method is to reduce the degrees of freedom, and another way is to identify appropriate mathematical forms that can be used to develop test statistic with high power (Zhao et al., 2005). The standard χ^2 test is conducted based on the linear transformations of genotype frequencies and hence is not the uniform most powerful test (Tzeng et al., 2003). In contrasts, the entropy-based test is based on the non-linear transformation of genotype frequencies which amplifies the difference in genotype frequencies between equilibrium (independence) and non-equilibrium (interaction) state of a genetic locus system and consequently leads to substantial power increase (Zhao et al., 2005). From a statistical physics point of view, entropy measures the degree of the

Table 4
Interaction test between genes Hb and α^+-thalassemia

Hb	α^+-Thalassemia	Malaria admission		*P* value		Severe malaria		*P* value	
		No. of cases	No. of controls	Wald Test[a]	Entropy-based test	No. of cases	No. of controls	Wald Test[a]	Entropy-based test
Hb*AA*	$\alpha\alpha/\alpha\alpha$	168	458			67	559		
	$-\alpha/\alpha\alpha$	187	680			53	814		
	$-\alpha/\alpha\alpha$	56	246			17	285		
Hb*AS*	$\alpha\alpha/\alpha\alpha$	6	107	0.026	0.038	0	113	0.0012	0.0043
	$-\alpha/\alpha\alpha$	9	141			2	148		
	$-\alpha/\alpha\alpha$	10	36			5	41		

non-structure of a system. The difference of allele frequencies among affected and unaffected populations reflects the degree of non-structure of a complex disease, which indicates that there may be an association between a locus and a disease (Skol et al., 2006). Moreover, differences among genotype combination frequencies as a system between the observed data and the one assuming no interaction reflect the degree of the non-structure change for a complex disease, which reflects the perturbations of the underlying genetic factors conferring a disease. Alternately, it suggests that there exists interactions among multiple loci. Therefore, we can test the interactions among multiple loci by comparing the entropy difference of multiple loci between the observed entropy and the entropy assuming no interaction.

Such a novel approach using entropy-based test for interaction detection displays several advantages in a case-only design. First, interactions among multiple loci can be characterized by the entropy of a locus system. Therefore, entropy-based statistics for detection of interaction among multiple loci have good biological interpretation. Second, the new model can consider the nonlinear transformation of frequencies of joint genotypes in a genetic system. This might explain why the entropy-based statistic for detection of interaction among multiple loci has higher power than the traditional χ^2 test. Third, since we deal with the interactive loci as a locus system, we can simply detect the association of multiple loci with a clinical phenotype and identify the most functional ones that interact with other loci to contribute to a clinical phenotype.

We conduct computer simulations to investigate the statistical behavior of the new approach. The results show that the null distribution of the proposed entropy-based test asymptotically follows a central χ^2 distribution. The entropy-based test outperforms the standard χ^2 test in terms of type I error rate control and testing power. We further apply the proposed entropy-based statistic to two real data sets and genetic interactions are detected among the candidate genes. The P values obtained using our approach are smaller than that of the χ^2 test and are comparable to the Poisson regression analysis. Here we focus our discussion on the schizophrenia data set.

Schizophrenia is a putative neuro-developmental disorder with a glutamatergic transmission abnormality. The combined effects of gene variability, including variations among *NRG1*, *G72* and *RGS4*, on schizophrenia could influence synaptic plasticity and the cortical microcircuitry via *N*-methyl-D-aspartate (NMDA) receptors (Harrison and Weinberger, 2005; Lewis and Levitt, 2002). Several functional studies suggested an interaction between *NRG1* and *RGS4* or *G72*. For example, *RGS4* interacts with ErbB3, which may be of relevance, since ErbB3 is an *NRG1* receptor with down-regulated mRNA expression in schizophrenic brains (Thaminy et al., 2003). In the present study, through using of the entropy, we provide additional support for the contributions of *NRG1*, *G72* and *RGS4* variants to schizophrenia. These findings also validate our new approach. The genetic interaction network constructed based on the entropy test provides very useful and informative information on understanding how genes interact to contribute to a clinical phenotype, and which locus is the most important one. Further lab verification is need to validate the result. For example, we may mutate SNP *NRG1* (rs3735774) within gene *NRG1* to see if there is still phenotype "delusions". For the 30 clinical phenotypes, most interaction networks contain the two functional SNPs, *NRG1* (rs3924999) and *NRG1* (rs3735774). These results confirm the importance of these two loci associated with the schizophrenia disease.

In a conclusion, we develop a measure of interactions among multiple loci and introduce a new entropy-based statistic to test interactions among multiple loci. We explore allele and loci heterogeneity, identify the relationships among disease genes and clinical phenotypes by introducing the information theory into genetics. Our approach has potential to integrate both clinical phenotype and interactions among multiple loci into genome-wide association analysis of complex human diseases. However, like all population-based analysis for association studies, the entropy-based statistic for testing interaction among multiple loci also has its limitations, (1) it does not consider the environmental effects as well as the gene–environment interactions; (2) it will lose power at the present of pleiotropy; and (3) it is only valid under the assumption of linkage equilibrium among testing loci. It deserves a more close investigation about these limitations in the future.

Conflict of interest statement

The authors declare no conflicts of interest in this work.

Acknowledgments

The authors wish to thank Michael Boehnke and James Stapleton for their valuable suggestions which have improved the presentation of the manuscript. This work was supported in part by grants from the National Natural Science Foundation of China 30530290 (to D. Zhang), 30400149 (to W. Yue), 60334040 (to J. Zhang), the National High-Tech Research and Development Program of China 2006AA02Z195 (to D. Zhang), 2007AA02Z423 (to W. Yue), the National Key Project Grant 2007CB512301 (to W. Yue), and the National Science Foundation DMS 0707031 (to Y. Cui), 0234078 (to Y. Zuo).

Appendix A. $\Delta S = \frac{1}{n^D}\log\frac{L_{observe}}{L_{ind}}$

We consider a locus system with s bi-allelic loci each with genotypes 0, 1, 2 and with $h^i = (h^i_1, h^i_2, \ldots, h^i_s)$ as the ith joint genotype ($h^i_k \in \{0,1,2\}, 1 \leqslant k \leqslant s$) with frequency p_i in cases. Denote $n^D = \sum_{i=1}^{3^s} n_i$ the number of cases, where n_i

represents the number of subjects with the ith joint genotype in cases, that is, $p_i = \frac{n_i}{n^D}$. Denote $p_{(k,2)}$ and $p_{(k,1)}$ be, respectively, the marginal frequencies of genotypes 2 and 1 at locus $1 \leqslant k \leqslant s$. The frequency of the ith joint genotype under no interaction is given by product of the marginal genotype frequencies:

$$q_i = \prod_{k=1}^{s} p_{(k,2)}^{x(i,2,k)} p_{(k,1)}^{x(i,1,k)} p_{(k,0)}^{x(i,0,k)}, \tag{A.1}$$

where

$$x(i,j,k) = \begin{cases} 1, & h_k^i = j, \\ 0, & h_k^i \neq j. \end{cases}$$

Define $\Delta_i = p_i - q_i$. We can easily get $\sum_{i=1}^{3^s} \Delta_i = 0$. Denote $L_{observe}$ and $S_{observe}$ be the likelihood and entropy of a system for the observation and L_{ind} and S_{ind} be for the case under no interaction. Then

$$\begin{aligned} \frac{1}{n^D} \log L_{observe} &= \sum_{i=1}^{3^s} \frac{n_i}{n^D} \log p_i \\ &= \sum_{i=1}^{3^s} p_i \log p_i = -S_{observe}, \end{aligned} \tag{A.2}$$

$$\begin{aligned} \frac{1}{n^D} \log L_{ind} &= \sum_{i=1}^{3^s} \frac{n_i}{n^D} \log q_i = \sum_{i=1}^{3^s} (q_i + \Delta_i) \log q_i \\ &= -S_{ind} + \underbrace{\sum_{i=1}^{3^s} \Delta_i \log q_i}_{\Omega_1}, \end{aligned} \tag{A.3}$$

where

$$\begin{aligned} \Omega_1 &= \sum_{i=1}^{3^s} \Delta_i \log \left[\prod_{k=1}^{s} p_{(k,2)}^{x(i,2,k)} p_{(k,1)}^{x(i,1,k)} p_{(k,0)}^{x(i,0,k)} \right] \\ &= \sum_{i=1}^{3^s} \sum_{k=1}^{s} \Delta_i x(i,2,k) \log p_{(k,2)} \\ &\quad + \sum_{i=1}^{3^s} \sum_{k=1}^{s} \Delta_i x(i,1,k) \log p_{(k,1)} \\ &\quad + \sum_{i=1}^{3^s} \sum_{k=1}^{s} \Delta_i x(i,0,k) \log p_{(k,0)} \\ &= \sum_{k=1}^{s} \log p_{(k,2)} \left(\underbrace{\sum_{i=1}^{3^s} \Delta_i x(i,2,k)}_{\Omega_2} \right) \\ &\quad + \sum_{k=1}^{s} \log p_{(k,1)} \left(\underbrace{\sum_{i=1}^{3^s} \Delta_i x(i,1,k)}_{\Omega_3} \right) \end{aligned} \tag{A.4}$$

and

$$\begin{aligned} \Omega_2 &= \sum_{i=1}^{3^s} \Delta_i x(i,2,k) = \sum_{i=1}^{3^s} x(i,2,k)(p_i - q_i) \\ &= \sum_{i=1}^{3^s} x(i,2,k) p_i - \sum_{i=1}^{3^s} x(i,2,k) \prod_{l=1}^{s} (p_{(l,2)}^{x(i,2,l)} p_{(l,1)}^{x(i,1,l)} p_{(l,0)}^{x(i,0,l)}) \\ &= p_{(k,2)} - \sum_{i=1}^{3^s} p_{(k,2)} \prod_{l=1, l\neq k}^{s} (p_{(l,2)}^{x(i,2,l)} p_{(l,1)}^{x(i,1,l)} p_{(l,0)}^{x(i,0,l)}) \\ &= p_{(k,2)} - p_{(k,2)} \prod_{l=1, l\neq k}^{s} \left(\sum_{i=1}^{3^s} p_{(l,2)}^{x(i,2,l)} p_{(l,1)}^{x(i,1,l)} p_{(l,0)}^{x(i,0,l)} \right) \\ &= p_{(k,2)} - p_{(k,2)} \prod_{l=1, l\neq k}^{s} 1 = 0. \end{aligned} \tag{A.5}$$

Similarly, we obtain $\Omega_3 = 0$. Thus, we get $\frac{1}{n^D} \log L_{ind} = -S_{ind}$. If there are some disease loci with multiple alleles, by the similar argument, the results are also true.

Appendix B. The entropy of a clinical phenotype and the likelihood of a multinomial distribution

We consider m disease loci as a locus system and a clinical phenotype Ψ. The entropy of the clinical phenotype Ψ is calculated as

$$S(\Psi) = \begin{cases} -\dfrac{\sum_{i=1}^{W} \frac{k_i^D}{K^D} \log \frac{k_i^D}{K^D}}{\log W}, & W > 1, \\ 0, & W = 1, \end{cases} \tag{B.1}$$

where K^D is the number of cases with clinical phenotype Ψ, $\{k_i^D\}_{i=1}^{W}$ is the number of cases with joint genotype h^i in K^D and W is the number of present joint genotypes in this locus system. Also, $\{k_i^D\}_{i=1}^{W}$ follows a multinomial distribution. We have

$$L(\{k_i^D\}_{i=1}^{W}) = \prod_{i=1}^{W} \left(\frac{k_i^D}{K^D} \right)^{k_i^D}. \tag{B.2}$$

Taking the natural logarithm at both sides of the above equation, we get

$$\begin{aligned} \log(L(\{k_i^D\}_{i=1}^{W})) &= \sum_{i=1}^{W} k_i^D \log \frac{k_i^D}{K^D} = K^D \sum_{i=1}^{W} \frac{k_i^D}{K^D} \log \frac{k_i^D}{K^D} \\ &= -K^D S(\Psi) \log W. \end{aligned} \tag{B.3}$$

So,

$$L(\{k_i^D\}_{i=1}^{W}) = \mathrm{e}^{-K^D S(\Psi) \log W}. \tag{B.4}$$

Appendix C. The frequencies of genotypes in cases under two-locus threshold model

From the two-locus threshold model in Table 1, we get the penetrance of genotypes $AABB$, $AABb$, $AaBB$ and $AaBb$ is $\frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}$, the penetrance of genotypes $AAbb$, $Aabb$,

aaBb and *aabb* is $\frac{\gamma}{1+\gamma}$. So, the prevalence of disease is

$$P(D) = P_A P_B \frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}(1+P_a+P_b+P_aP_b) + \frac{\gamma}{1+\gamma}(P_a^2+P_b^2-P_a^2P_b^2), \quad \text{(C.1)}$$

where $P_a = 1-P_A, P_b = 1-P_B$.

Then, the frequencies of genotypes in cases under a threshold model are

	BB	*Bb*	*bb*
AA	$\frac{P_A^2P_B^2}{P(D)}\frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}$	$\frac{2P_A^2P_BP_b}{P(D)}\frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}$	$\frac{P_A^2P_b^2}{P(D)}\frac{\gamma}{1+\gamma}$
Aa	$\frac{2P_AP_aP_B^2}{P(D)}\frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}$	$\frac{4P_AP_aP_BP_b}{P(D)}\frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}$	$\frac{2P_aP_AP_B^2}{P(D)}\frac{\gamma}{1+\gamma}$
aa	$\frac{P_a^2P_B^2}{P(D)}\frac{\gamma}{1+\gamma}$	$\frac{2P_a^2P_BP_b}{P(D)}\frac{\gamma}{1+\gamma}$	$\frac{P_a^2P_b^2}{P(D)}\frac{\gamma}{1+\gamma}$

where $P(D)$ is referred to (C.1).

Under a two-locus multiplicative model, the frequencies of genotypes in cases can similarly be got.

Appendix D. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2007.10.001.

References

Ackerman, H., Usen, S., Mott, R., Richardson, A., Sisay-Joof, F., Katundu, P., Tayor, T., Ward, R., Molyneux, M., Pinder, M., et al., 2003. Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Biol. 4 (4), R24.

Andreasen, N.C., 2000. Schizophrenia: the fundamental questions. Brain Res. Rev. 31, 106–112.

Andreasen, N.C., Nopoulos, P., Schultz, S., Miller, D., Gupta, S., Swayze, V., Flaum, M., 1994. Positive and negative symptoms of schizophrenia: past, present, and future. Acta Psychiatr. Scand. Suppl. 384, 51–59.

Brem, R.B., Storey, J.D., Whittle, J., Kruglyak, L., 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436, 701–703.

Carlacci, L., Chou, K.C., 1990. Monte Carlo method applied in the search for low energy conformations of structures. Biopolymers 30, 135–150.

Carrasquillo, M.M., McCallion, A.S., Puffenberger, E.G., Kashuk, C.S., Nouri, N., Chakravarti, A., 2002. Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. Nat. Genet. 32, 237–244.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York, pp. 12–15.

Gauderman, W.J., 2002. Sample size requirement for association studies of gene–gene interaction. Am. J. Epidemiol. 155, 478–484.

Greiner, W., Neise, L., Stocker H. (Translator) 1995. Thermodynamics and Statistical Mechanics. Springer, New York, pp. 121–135.

Hampe, J., Schreiber, S., Krawczak, M., 2003. Entropy-based SNP selection for genetic association studies. Hum. Genet. 114, 36–43.

Harrison, P.J., Weinberger, D.R., 2005. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol. Psychiatry 10, 40–68.

Jawaheer, D., Li, W.T., Graham, R.R., Chen, W., Damle, A., Xiao, X.L., Monteiro, J., Khalili, H., Lee, A., Lundsten, R., et al., 2002. Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. Am. J. Hum. Genet. 73, 585–594.

Judson, R., Salisbury, B., Schneider, J., Windemuth, A., Stephens, J.C., 2002. How many loci does a genome-wide haplotype map require? Pharmacogenomics 3, 279–391.

Kang, G.L., Li, S., Zhang, J.F., 2007. Entropy-based models for interpreting life systems in traditional Chinese medicine. Evidence-based Complementary Alternative Med. doi:10.1093/ecam/nem026.

Kang, G.L., Zuo, Y.J., 2007. Entropy-based joint analysis for two-stage genome-wide association studies. J. Hum. Genet. 52, 747–756.

Kirov, G., O'Donovan, M.C., Owen, M.J., 2005. Finding schizophrenia genes. J. Clin. Invest. 115, 1440–1448.

Kubat, J.A., Chou, J.J., Rovnyak, D., 2007. Nonuniform sampling and maximum entropy reconstruction applied to the accurate measurement of residual dipolar couplings. J. Magn. Reson. 186, 201–211.

Lewis, D.A., Levitt, P., 2002. Schizophrenia as a disorder of neurodevelopment. Annu. Rev. Neurosci. 25, 409–432.

Macdonald, S.M., Long, A.D., 2005. Prospects for identifying functional variation across the genome. Proc. Natl. Acad. Sci. USA 102, 6614–6621.

Mihalek, I., Res, I., Lichtarge, O., 2004. A family of evolution-entropy hybrid methods for ranking protein residues by importance. J. Mol. Biol. 336, 1265–1282.

Moore, J.H., Hahn, L.W., 2002. A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. In: Pacitic Symposium on Biocomputations, Pac. Symp. Biocomput. 7, 53–64.

Nelson, M.R., Kardia, S.L.R., Ferrell, R.E., Sing, C.F., 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. 11, 458–470.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H., 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69, 138–147.

Schaid, D.J., Mcdonnall, S.K., Hebbring, S.J., Cunningham, J.M., Thibodeau, S.N., 2005. Nonparametric tests of association of multiple genes with human disease. Am. J. Hum. Genet. 76, 780–793.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.

Skol, A.D., Scott, L.J., Abecasis, G.R., Boehnke, M., 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat. Genet. 38, 209–213.

Soares, M.L., Coelho, T., Sousa, A., Batalov, S., Conceicao, I., Sales-Luis, M.L., Ritchie, M.D., Williams, S.M., Nievergelt, C.M., Schork, N.J., et al., 2005. Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. Hum. Mol. Genet. 14, 543–553.

Strohman, R., 2002. Maneuvering in the complex path from genotype to phenotype. Science 296, 701–703.

Thaminy, S., Auerbach, D., Arnoldo, A., Stagljar, I., 2003. Identification of novel ErbB3-interacting factors using the Split-Ubiquitin membrane yeast two-hybrid system. Genome Res. 13, 1744–1753.

Tzeng, J.Y., Devlin, B., Wasserman, L., Roeder, K., 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am. J. Hum. Genet. 72, 891902.

Wilks, S.S., 1962. Mathematical Statistics. New York, Wiley, pp. 418–418.

Williams, T.N., Mwangi, T.W., Wambua, S., Peto, T.E., Weatherall, D.J., Gupta, S., Recker, M., Penman, B.S., Uyoga, S., Macharia, A., et al., 2005. Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. Nat. Genet. 37, 1160–1162.

Yang, Q., Khoury, M.J., Sun, F., Flanders, W.D., 1999. Case-only design to measure gene–gene interaction. Epidemiology 10, 167–170.

Yue, W.H., Kang, G.L., Zhang, Y.B., Qu, M., Tang, F.L., Han, Y.H., Ruan, Y., Lua, T.L., Zhang, J.F., Zhang, D., 2007. Association of DAOA polymorphisms with schizophrenia and clinical symptoms or therapeutic effects. Neurosci. Lett. 416, 96–100.

Zhang, C.T., Chou, K.C., 1992. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophys. J. 63, 1523–1529.

Zhang, C.T., Chou, K.C., 1995. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. II. Correlative effect. J. Protein Chem. 14, 251–258.

Zhao, J.Y., Boerwinkle, E., Xiong, M.M., 2005. An entropy-based statistic for genomewide association studies. Am. J. Hum. Genet. 77, 27–40.

Zhao, J.Y., Jin, L., Xiong, M.M., 2006. Test for interaction between two unlinked loci. Am. J. Hum. Genet. 79, 831–845.