European Journal of Human Genetics (2008) 16, 229–235

www.nature.com/ejhg

ARTICLE

Exploration of gene–gene interaction effects using entropy-based methods

Changzheng Dong[1,2,3], Xun Chu[2], Ying Wang[2], Yi Wang[4], Li Jin[4], Tieliu Shi[1], Wei Huang*[,2,5] and Yixue Li*[,1,6]

[1]*Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, the Chinese Academy of Sciences, Shanghai, China;* [2]*Chinese National Human Genome Center at Shanghai, Shanghai, China;* [3]*Graduate School of the Chinese Academy of Sciences, Beijing, China;* [4]*MOE Key Laboratory of Contemporary Anthropology and Center for Evolutionary Biology, Fudan University, Shanghai, China;* [5]*Rui Jin Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China;* [6]*Shanghai Center for Bioinformation Technology, Shanghai, China*

Gene–gene interaction may play important roles in complex disease studies, in which interaction effects coupled with single-gene effects are active. Many interaction models have been proposed since the beginning of the last century. However, the existing approaches including statistical and data mining methods rarely consider genetic interaction models, which make the interaction results lack biological or genetic meaning. In this study, we developed an entropy-based method integrating two-locus genetic models to explore such interaction effects. We performed our method to simulated and real data for evaluation. Simulation results show that this method is effective to detect gene–gene interaction and, furthermore, it is able to identify the best-fit model from various interaction models. Moreover, our method, when applied to malaria data, successfully revealed negative epistatic effect between sickle cell anemia and α^{+}-thalassemia against malaria.

European Journal of Human Genetics (2008) **16,** 229–235; doi:10.1038/sj.ejhg.5201921; published online 31 October 2007

Keywords: genetic interaction; epistasis; interaction model; entropy; logistic regression

Introduction

The advent of high-throughput genotyping technology has made the technology of whole-genome single-nucleotide polymorphisms (SNPs) scanning for susceptible genes easy and popular, resulting in the generation of mass genotyping data. Analysis of these data using statistical and data mining methods have led to the discoveries of many association or predisposing genes, yet few causative genes were determined.[1] In addition, it is hard to put forward genetic mechanisms for most common diseases. Therefore, many researches are focusing on factors affecting the power of association study.[2–4] One of the most important factors is gene–gene interaction[5,6] and the other is gene–environment interaction.[7] In this paper, we focus on gene–gene interaction, or the so-called epistasis.

As first advanced by Bateson,[8] epistasis has been defined as a phenomenon whereby the effects of a given gene on a biological trait are masked or enhanced by one or more genes.[9] Several studies[8–11] have provided evidences for the existence of gene–gene interaction or epistasis. Since gene–gene interactions may play a role in the mechanisms of complex diseases and weaken the major effects of single gene, the association study often turns out to be confusing and hard to explain.[12] Genetic interaction models that consider two-locus genotype combinations have been proposed; for example (Figure 1), the threshold model,

*Correspondence: Dr Y Li, Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, the Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China.
Tel: 86 21 64836199; Fax: 86 21 54920089; E-mail: yxli@sibs.ac.cn or
Dr W Huang, Rui Jin Hospital, School of Medicine, Shanghai Jiaotong University, 197, Rui Jin II Road, Shanghai, 200025, China.
Tel: 86 21 50801795; Fax: 86 21 50801795; E-mail: huangwei@chgc.sh.cn
Received 3 February 2007; revised 26 June 2007; accepted 16 August 2007; published online 31 October 2007

M1 (RR)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
aa	0	0	1

M3 (RD)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
aa	0	1	1

M7 (1L: R)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
aa	1	1	1

M11 (T)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	1
aa	0	1	1

M15 (Mod)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	1
aa	1	1	1

M27 (DD)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	1	1
aa	0	1	1

M78 (XOR)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	1
Aa	0	0	1
aa	1	1	0

M84 (Diagonal)

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	1
Aa	0	1	0
aa	1	0	0

Figure 1 Eight typical two-locus models. 1 denotes high-risk genotype combinations and 0 low-risk. M1, jointly recessive-recessive model (RR); M3, jointly recessive-dominant model (RD); M7, single-gene recessive model (1L: R); M11, threshold model (T); M15, modifying-effect model (Mod); M27, jointly dominant-dominant model (DD); M78, exclusive OR model (XOR); M84. diagonal model (Diagonal).

jointly recessive–recessive model and jointly dominant–dominant model.[13–15] Some of these models, such as the additive model, multiplicative model and heterogeneity model, can be presented as deviation formation.[15–17] Li and Reich[15] have enumerated all possible two-locus models, some of which had been reported with significant biological meaning.

Statistical[18] and data mining methods are the mainstream in the current analysis approaches. One of the most familiar methods is Cordell's unified stepwise regression procedure,[5] which can be applied to the additive, multiplicative and heterogeneity model. It is the most familiar method for analyzing interaction effects. Millstein *et al*[19] developed an interaction testing framework called FITF, which is also based on stepwise regression. Recently, Zhao *et al*[20] proposed an LD-based measure between two unlinked loci and the method was proved to be powerful under some two-locus disease models. Evans *et al*[21] investigated the performance of two simple two-stage strategies. Moore *et al*[22] used attribute interaction[23] to select potential interacting SNPs and construct interaction graph. While these methods are applicable and useful, they often could not distinguish which two-locus model was proper for the interaction effects but could only tell that interaction exists under certain genetic data. Multilocus statistics such as S-statistic[24] and data mining approaches, such as multifactor-dimensionality reduction (MDR),[25] restricted partitioning method,[26] combinatorial partitioning method,[27] dynamic algorithm (DA),[28] decision tree[29] and random forest,[30] are powerful to reduce data dimension and to get a set of SNPs that can interpret the results best. Taking the popular MDR[25] as an example, it considers each possible genotype combination of SNPs as high- or low-risk combinations, repeats the calculation from single SNP to multi-SNP combinations and performs cross-validation to get combinations with maximum cross-validation consistency and minimum prediction error at different dimensions. However, as Moore *et al* pointed out, the combination results of MDR were still hard to interpret. These methods focus on data reduction by mathematical methods but ignore how to interpret the resulting interaction effects from the point of view of biology or genetics. Thus the results are often confusing as statistical significance may not correspond to biological or genetic significance.[6,31] For example, if we know two SNPs have significant statistical interaction, how do they interact in biology or genetics? Does each allele of SNPs act? Do dominant or recessive effects exist?

To solve these problems, we developed a novel entropy-based method called ESNP2 (entropy-based SNP–SNP interaction method) integrating two-locus genetic models. In our ESNP2 system, there are two options: ESNP2-S (ESNP2-standard option) and ESNP2-Mx (ESNP2-model option). The former aims to detect two-locus interactions, whereas the latter, ESNP2-Mx, tests the interaction against various two-locus genetic models and gets the best-fit model. A program implemented by Java for ESNP2 algorithm can be downloaded from our website (http://www.biosino.org/papers/esnp2/).

Methods

Entropy is defined as follows:[32]

$$H(p,1-p) = -p\log p - (1-p)\log(1-p)$$

$$p = \frac{N_{\text{case}}}{N_{\text{case}} + N_{\text{control}}}$$

where N_{case} denotes the number of cases in the population and N_{control} the number of controls in the population.

ESNP2-S

Our entropy-based method has three steps (Figure 2).

First step: compute entropy of initial data set (H_0)

Given an initial data set (or sample), D (case + control), p and $(1-p)$ are the proportions of case and control in the data set, respectively. According to p, we can compute the entropy of the data set:

$$H_0 = H(D) = H(p,1-p) = -p\log p - (1-p)\log(1-p)$$

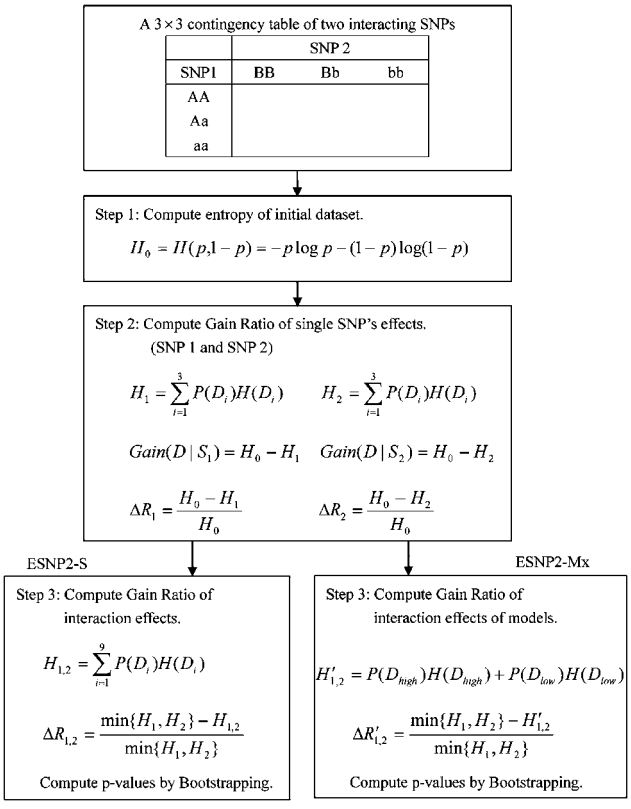

Figure 2 Workflows of ESNP2-S and ESNP2-Mx.

Second step: compute gain ratio of single SNP (ΔR_1 and ΔR_2)
The effects of a single SNP that may be involved in interaction can be estimated by computing its gain ratio. Considering each SNP splitting the initial data set D with its possible genotypes into several subsets, $S_1(D) = S_1\{D_{AA}, D_{Aa}, D_{aa}\}$. Each subset has its sub-entropy $H(D_i)$ and corresponding weighting coefficient $P(D_i)$, which is the proportion of a certain genotype or genotype combinations to the total data set.

$$H_1 = H(D|S_1) = \sum_{i=1}^{3} P(D_i)H(D_i)$$

$$\mathrm{Gain}(D|S_1) = H_0 - H_1$$

$$\Delta R_1 = \frac{H_0 - H_1}{H_0}$$

Gain refers to information gain, which is also called mutual information when considering H_1 as conditional entropy. It reflects the relation between SNP and disease status. As gain correlates with entropy of the initial data set that is determined by p, we normalize it to eliminate the effects of p and get gain ratio ΔR_1. In addition, ΔR_1 is a likelihood ratio that can be approximated to a χ^2-test (proof not shown here).

Similarly, we can compute the gain ratio of another SNP ΔR_2 involved in interaction.

Third step: compute gain ratio of SNPs' interactions ($\Delta R_{1,2}$)
Interacting SNPs split the initial data set D into nine subsets:

$$S_{1,2}(D) = S_{1,2}\{D_{AABB}, D_{AABb}, D_{AAbb}, D_{AaBB}, D_{AaBb}, D_{Aabb}, D_{aaBB}, D_{aaBb}, D_{aabb}\}$$

Its gain ratio ΔR_{12} can be computed as follows:

$$H_{1,2} = H(D|S_{1,2}) = \sum_{i=1}^{9} P(D_i)H(D_i)$$

$$\mathrm{Gain}(D|S_{1,2}) = (H_0 - H_{1,2}) - \max\{(H_0 - H_1), (H_0 - H_2)\} = \min\{H_1, H_2\} - H_{1,2}$$

$$\Delta R_{1,2} = \frac{\min\{H_1, H_2\} - H_{1,2}}{\min\{H_1, H_2\}}$$

$\Delta R_{1,2}$ measures interaction effects of SNP1 and SNP2 whenever marginal effects exist. Bootstrapping strategy[33] is performed to get P-values corresponding to $\Delta R_{1,2}$ with the initial data set D. The bootstrapping strategy resamples random samples of size ($N_{\mathrm{case}} + N_{\mathrm{control}}$) with replacement from the original data. Repeating the sampling procedure a large number of times and counting new $\Delta R_{1,2}$ larger than the original value generates P-values.

ESNP2-Mx
ESNP2-S aims to explore the interaction effects of two SNPs. It explores all the nine possible combinations of two SNPs independently. To bestow ESNP2 with biological or genetic meaning, we extend ESNP2-S to ESNP2-Mx. Mx is the abbreviation of Model x, such as M1, M11 and M27, representing the jointly recessive model, threshold model and jointly dominant model, respectively. It is a binary coding system used by Li and Reich.[15] Figure 1 shows eight classical interaction models. Two-locus genetic models are presented as penetrance table where 1 denotes high penetrance or high risk and 0 low penetrance or low risk. High-risk genotype combination leads to higher disease susceptibility in case than in control. In practice, risk can be evaluated by the ratio p (ie, if a certain SNP combination has a p larger than the average of the data set, it can be considered as a high-risk combination). By using ESNP2-S to calculate gain ratios and P-values for each of the possible models, we then get the best-fit model. If both ESNP2-S and ESNP2-Mx result in similar and significant P-values, we can conclude the data are fitting a certain interaction model.

Similar to ESNP2-S, ESNP2-Mx procedure has three steps except the third step (Figure 2). With ESNP2-Mx, the initial

data set D is divided into two subsets, namely high- and low-risk subsets: $S'_{1,2}(D) = S'_{1,2}\{D_{\text{high}}, D_{\text{low}}\}$.

$$H'_{1,2} = P(D_{\text{high}})H(D_{\text{high}}) + P(D_{\text{low}})H(D_{\text{low}})$$

$$\Delta R'_{1,2} = \frac{\min\{H_1, H_2\} - H'_{1,2}}{\min\{H_1, H_2\}}$$

where $H'_{1,2}$ is the entropy and $\Delta R'_{1,2}$ is the gain ratio.

According to different candidate interaction models, new gain ratio $\Delta R'_{1,2}$ can be calculated, followed by a bootstrapping strategy, to get the *P*-values. The procedure is similar to that of ESNP2-S.

Results

Data simulation

To validate the effect of ESNP2-S and ESNP2-Mx on detecting the different association intensities between SNPs and disease, we constructed a simulated data set with respect to certain parameters. Odds ratio (OR) was used as a parameter to define the relationship between disease and two loci. For various interaction models, we set OR to be 1.2, 1.5 and 1.8 ordinally, the sample sizes of case and control to be 1000 and simulation times to be 100 to get the median of the results. Eight classical interaction models were selected as shown in Figure 1. First, the total number of high- and low-risk genotype combinations were randomly generated for case and control according to the above parameters. Then we got the number of all possible genotype combinations with the collapsibility of OR and certain interaction models.

The resulting simulation data were then analyzed by ESNP2-S and ESNP2-Mx, respectively, with bootstrapping times set to be 10^4 by which we could get the maximal precision of 10^{-4} (we marked it as 0 instead of $<10^{-4}$). Even longer bootstrapping times can be consumed if higher precise results are needed. The familiar logistic stepwise regression has also been performed to estimate the effects deviated from additive model in log scale, which is considered to be interaction effects. Single-gene effects have been estimated by the minimum *P*-value of single-point association results of two genes (Figures 3 and 4). As a control, one single-gene effect model (M7) was included.

As the value of the parameter OR increases, both single-gene effects and interaction effects enhance rapidly as shown by most models. Significant single-gene effects can be observed except in models M78 and M84, whereas M7 is a model of single-gene effects that is used as a control. According to M7 data, as all of the three models initially considered the single gene effect by default, they subsequently failed to detect any interaction effects. For models

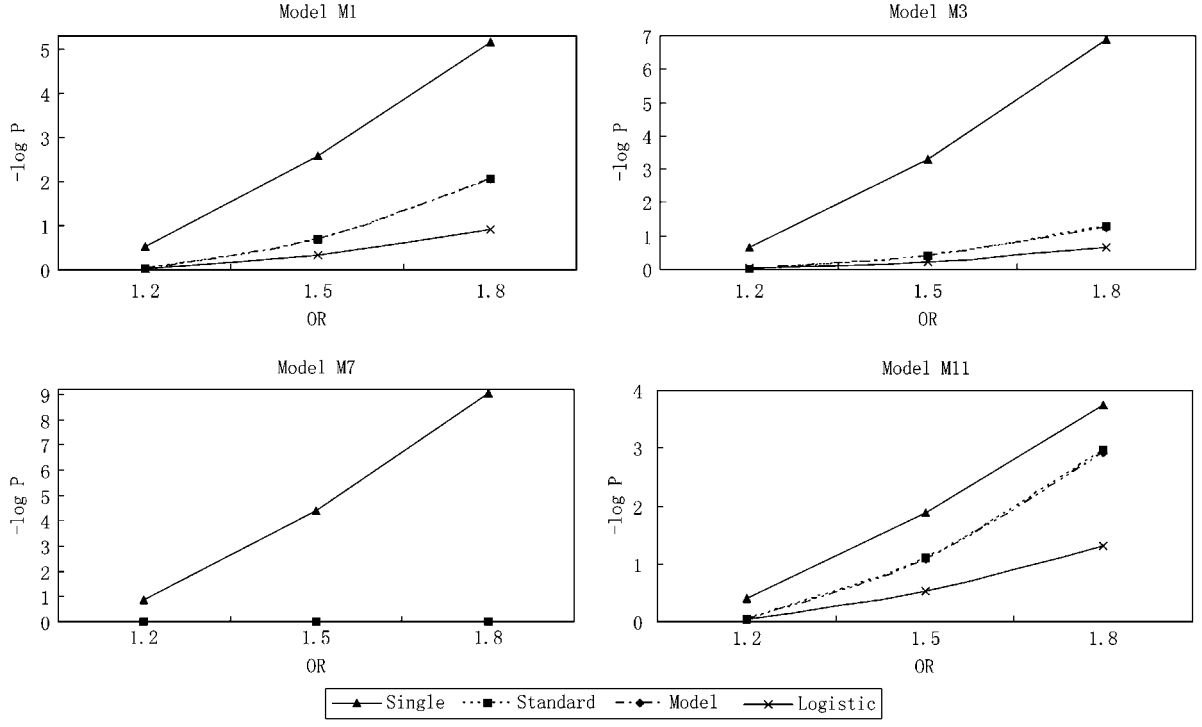

Figure 3 *P*-value curves generated by single association, ESNP2-S, ESNP2-Mx and logistic regression for simulated data (M1, M3, M7 and M11). Single, minimum *P*-value of two genes' association results representing maximum single-gene effects; standard, interaction effects computed by ESNP2-S; model, two-locus model effects computed by ESNP2-Mx; logistic, interaction effects computed by logistic regression representing effects deviation from additive effects. Standard curve overlaps with model curve entirely.

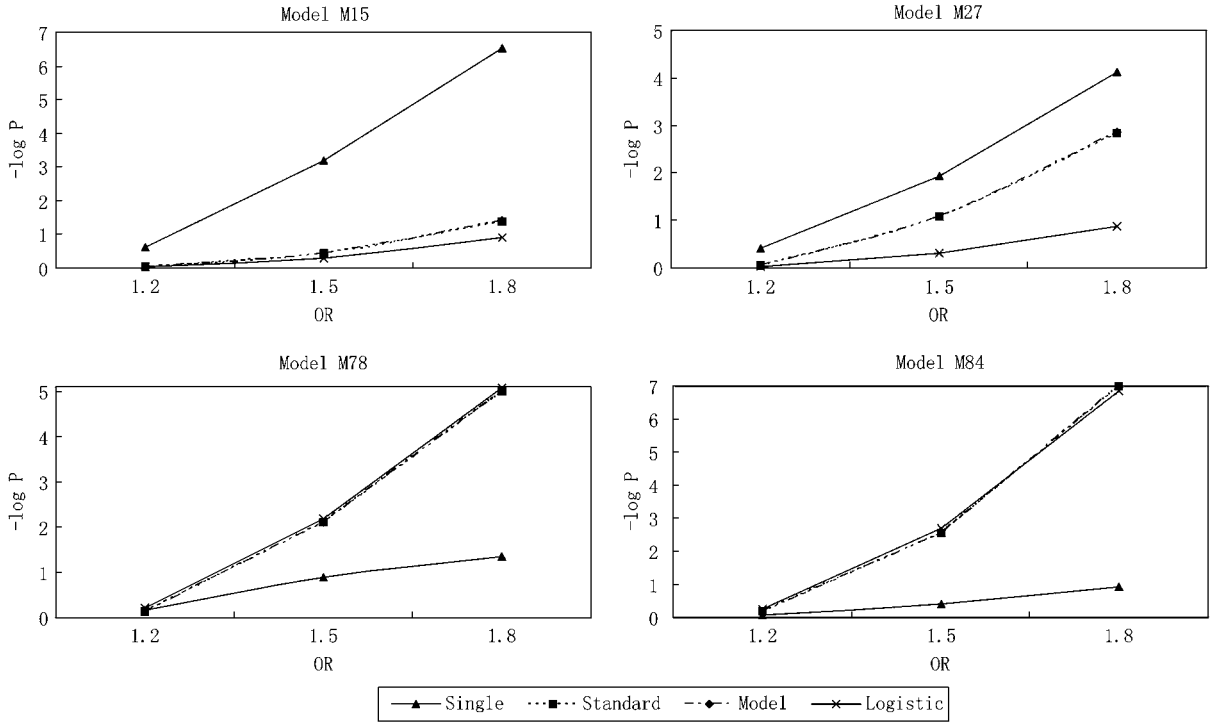

Figure 4 *P*-value curves generated by single association, ESNP2-S, ESNP2-Mx and logistic regression for simulated data (M15, M27, M78 and M84).

M1, M3, M11, M15 and M27, single-gene effects change more variously than interaction effects. Contrarily, although M78 and M84 are able to detect stronger interaction effects, yet only weak single-gene effects have been observed. Similarly, M1, M11 and M27 show moderate ability in identifying single-gene effects and can thus be considered to be standard and symmetric interaction models. On the other hand, whereas most of the genetic effects shown by M3 and M15 can be mainly interpreted by a single gene, only a few epistasis effects occur when OR is large. As shown in Figures 3 and 4, ESNP2-Mx using correct models gets approximately the same results with ESNP2-S because the model data are simulated perfectly. For each of the eight models, the corresponding two curves overlap entirely with each other. In the case of real data analysis, results will deviate from the proposed models. So ESNP2-Mx will also deviate from ESNP2-S more or less. As shown in Appendix Table 1 (Supplementary Table 1), which presents with incorrect models, ESNP2-Mx has much larger *P*-values than ESNP2-S.

As shown in Figures 3 and 4, ESNP2 gets similar power with logistic regression, especially when interaction effects weigh much heavier and single-gene effects behave inconspicuously (such as M78, M84). For one thing, our ESNP2-S can detect interaction effects as sensitively as logistic regression; for another thing, ESNP2-Mx bears the power of selecting the best-fit model for the effects by analyzing and integrating different two-locus models with ESNP2-S.

Analysis of real data

We then incorporated a real data set of malaria cohort study (Tables 1 and 2) performed by Williams *et al*[34] in Kenya to further evaluate the availability of ESNP2-S and ESNP2-Mx. Previous studies have shown that an important causative protein in malaria is Hemoglobin (Hb), which has two variants – heterozygote HbAS and homozygote HbSS that can cause sickle cell anemia. While HbSS is a lethal mutation leading to premature death, individuals with HbAS are protective against malaria. Additionally, there exist other mutations that can protect against severe and fatal malaria – heterozygote $-\alpha/\alpha\alpha$ and homozygote $-\alpha/-\alpha$, which cause α^+-thalassemia.

Table 1 shows data of genotype combinations for malaria admission, formatted as case/control. χ^2-tests show that sickle cell anemia has strong association (3×10^{-9}) with malaria resistance, while α^+-thalassemia has small association (0.063). We first applied ESNP2-S on this data set to detect the potential interaction effects. The resulting *P*-value of this step, as equaling to 6.8×10^{-5}, provides supporting evidence for interaction. Then we performed ESNP2-Mx using negative epistasis model (Table 3) to calculate the interaction model effects and got a *P*-value of 0.008. As Table 1 is a cohort data, we corrected it with surveying time (child years of follow-up) as shown in Table 2. The corrected data have also been analyzed by our models, resulting in the *P*-value being $<10^{-6}$ of the interaction model effects. We thus concluded that the

Table 1 Malaria admission data of sickle cell anemia and α^+-thalassemia combination effects

		α^+-*Thalassemia*		
		$\alpha\alpha/\alpha\alpha$	$-\alpha/\alpha\alpha$	$-\alpha/-\alpha$
Hb	HbAA	168:458[a]	187:680	56:246
	HbAS	6:107	9:141	10:36

[a]Data are from Williams's cohort data,[34] shown as case/control format.

Table 2 Malaria admission data of sickle cell anemia and α^+-thalassemia combination effects with adjustment for child years of follow-up

		α^+-*Thalassemia*		
		$\alpha\alpha/\alpha\alpha$	$-\alpha/\alpha\alpha$	$-\alpha/-\alpha$
Hb	HbAA	69.14:188.49[a]	55.69:202.52	47.22:207.42
	HbAS	13.77:245.63	15.28:239.39	54.67:196.72

[a]Data are from Williams's cohort data[34] with adjustment for child years of follow-up, shown as case/control format.

Table 3 A negative epistasis model protecting against malaria

		α^+-*Thalassemia*		
		$\alpha\alpha/\alpha\alpha$	$-\alpha/\alpha\alpha$	$-\alpha/-\alpha$
Hb	HbAA	1[a]	1	1
	HbAS	0	0	1

[a]1, high-risk genotype combinations; 0, low-risk genotype combinations.

negative epistasis model fits the malaria data satisfactorily. While the logistic regression model coupled with Wald test can only tell whether interactions between sickle cell anemia and α^+-thalassemia exist and affect malaria resistance, our method goes a step further by not only confirming the results but also finding that the negative epistasis model is proper for the relationship.

Discussion

The epistasis models have been developed and enhanced to be more complex and perfect since it was first proposed by Bateson[8] one hundred years ago, resulting in the appearance of numerous published models.[15] While these models are effective and powerful in practice application, they demand different number of disease alleles and variously show interaction effects. The classical epistasis model or the so-called modifying-effect model (Mod, M15) is similar to the single-locus recessive model when regardless of Aabb combination. The threshold model (T, M11) demands that at least three disease alleles are affected; the jointly recessive–recessive model (RR, M1) requires the presence of two copies of disease alleles.[15] Our simulation results in Figures 3 and 4 show marginal and interaction effects of various two-locus models, which have also been supported by marginal effects computed by Li and Reich.[15] Identifying fitting models for the interacting SNPs will contribute to our knowledge about both single gene and interaction effects. Advances in researches on human diseases[35] and mouse models[36] have proposed some gene–gene interaction models that contribute to discovering the genetic mechanisms of common diseases. Moreover, the models are widely used to simulate data for testing the power of novel methods.[20–22]

In the situation of single SNP association, the routine strategy includes performing allele/genotype tests to ascertain susceptible SNP, followed by searching for the best-fitting method to get proper genetic models (additive, dominant, recessive, etc.).[37] The existing interaction approaches consider only the first step and calculate interaction effects, while our methods implement the second step: combining genetic models and gene–gene interaction measurements with the information concept, entropy. The advantage of our method has been exhibited fully in the real data analysis where we got a negative epistatic model. The key statistical parameter is called gain ratio, which describes normalized information gain. A similar work has been carried out by Moore *et al*.[22] The authors used attribute interaction[23] to select SNPs, which serve as a coordinate format of information gain. Gain ratio has the further advantage of eliminating potential effects of the initial data set (the ratio between case and control) and computing interaction effects whenever marginal effects exist. By integrating genetic models with gain ratio, genotype combinations can be distinguished by the level of risks, and different genotype combinations can be grouped together if they have the same risk. In the algorithm, we only categorize the risk into two classes according to its value, namely high risk and low risk; however, more levels can be used if needed.

As gain ratio of ESNP2 does not belong to any classical statistic models, we performed bootstrapping to compute the *P*-values. The power of bootstrapping depends on bootstrapping times as well as the intensity of interaction effects. Longer bootstrapping times can be consumed if higher precise results are needed. Simulation results showed that bootstrapping could gain similar power as that of logistic regression under most circumstances.

On the whole, we developed an entropy-based method called ESNP2 to explore gene–gene interaction in SNPs aimed at discovering their potential biological or genetic meaning. Two options are provided: the model-free option ESNP2-S detects gene–gene interaction and the model-based ESNP2-Mx gets the best-fitting model by fitting various two-locus genetic models on the potential interaction. Both simulation data and real data of malaria show that this method is effective in detecting gene–gene interaction and revealing potential genetic models, which

may contribute to the understanding of genetic mechanisms of most common diseases. Further researches should try to develop a more powerful statistic strategy that can get the *P*-values directly instead of bootstrapping. Another direction should be to focus on extending the methods and rationale to multilocus.

Acknowledgements

We thank Dr. Momiao Xiong for critical reviews of this manuscript and Ms Yi Wang for writing a modification. We thank Peilin Jia, Guohui Ding, Ziliang Qian and Hong Li for their valuable discussions. We thank two anonymous reviewers for their critical comments. WH is supported by the grants from Chinese National Science Fund for distinguished young scholars (30625019), the National Basic Research Program (2004CB518605), Chinese High-Tech Program (2006AA020706), and Shanghai Science and Technology Committee (06XD14015). YL is supported by the grants from National Basic Research Program (2001CB510209, 2003CB715901, 2004CB518606).

References

1 Weiss KM, Terwilliger JD: How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000; **26**: 151–157.
2 Cardon LR, Bell JI: Association study designs for complex diseases. *Nat Rev Genet* 2001; **2**: 91–99.
3 Moore JH, Ritchie MD: The challenges of whole-genome approaches to common diseases. *JAMA* 2004; **291**: 1642–1643.
4 Wang W, Barratt BJ, Clayton DG, Todd JA: Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005; **6**: 109–118.
5 Cordell HJ, Clayton DG: A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* 2002; **70**: 124–141.
6 Ritchie MD, Hahn LW, Roodi N *et al*: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; **69**: 138–147.
7 Hunter DJ: Gene–environment interactions in human diseases. *Nat Rev Genet* 2005; **6**: 287–298.
8 Bateson W: *Mendel's principles of heredity*. United Kingdom: Cambridge, 1909.
9 Moore JH: A global view of epistasis. *Nat Genet* 2005; **37**: 13–14.
10 Malmberg RL, Held S, Waits A, Mauricio R: Epistasis for fitness-related quantitative traits in arabidopsis thaliana grown in the field and in the Greenhouse. *Genetics* 2005; **171**: 2013–2027.
11 Segrè D, Deluna A, Church GM, Kishony R: Modular epistasis in yeast metabolism. *Nat Genet* 2005; **37**: 77–83.
12 Culverhouse R, Suarez BK, Lin J, Reich T: A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 2002; **70**: 461–471.
13 Marchini J, Donnelly P, Cardon RC: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005; **37**: 413–447.
14 Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC: Routine discovery of complex genetic models using genetic algorithms. *Appl Soft Comput* 2004; **4**: 79–86.
15 Li W, Reich J: A complete enumeration and classification of two-locus disease models. *Hum Hered* 2000; **50**: 334–349.
16 Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002; **11**: 2463–2468.
17 Carlborg O, Haley CS: Epistasis: too often neglected in complex trait studies? *Nat Rev Gene* 2004; **5**: 618–625.
18 Zhao N: *Medical statistics*. China: Beijing, 2004.
19 Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 2006; **78**: 15–27.
20 Zhao J, Jin L, Xiong MM: Test for interaction between two unlinked loci. *Am J Hum Genet* 2006; **79**: 831–845.
21 Evans DM, Marchini J, Morris AP, Cardon LR: Two-stage two-locus models in genome-wide association. *PLoS Genet* 2006; **2**: e157.
22 Moore JH, Gilberta JC, Tsai C *et al*: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* 2006; **241**: 252–261.
23 Jakulin A, Bratko I: Analyzing attribute interactions. *Lect Notes Artif Intell* 2006; **2838**: 229–240.
24 Hoh J, Wille A, Ott J: Trimming, weighting and grouping SNPs in human case–control association studies. *Genome Res* 2001; **11**: 2115–2119.
25 Hahn LW, Ritchie MD, Moore JH: Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* 2003; **19**: 376–382.
26 Culverhouse R, Klein T, Shannon W: Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 2004; **27**: 141–152.
27 Nelson MR, Kardia SLR, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001; **11**: 458–470.
28 Xu Q, Jia YB, Zhang BY *et al*: Association study of an SNP combination pattern in the dopaminergic pathway in paranoid schizophrenia: a novel strategy for complex disorders. *Mol Psychiatry* 2004; **9**: 510–521.
29 Chen CH, Chang CJ, Yang WS, Chen CL, Fann CS: A genome-wide scan using tree-based association analysis for candidate loci related to fasting plasma glucose levels. *BMC Genet* 2003; **4** (Suppl 1): S65.
30 Bureau A, Dupuis J, Falls K *et al*: Identifying SNPs predictive of phenotype using Random Forests. *Genet Epidemiol* 2005; **28**: 171–182.
31 Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. *Philos Trans R Soc Edinb* 1918; **52**: 399–433.
32 Shannon CE: A mathematical theory of communication. *Bell Syst Tech J* 1948; **27**: 379–423, 623–656.
33 Efron B, Tibshirani RJ: *An introduction to the bootstrap*. London: Chapman & Hall, 1993.
34 Williams TN, Mwangi TW, Wambua S *et al*: Negative epistasis between the malaria-protective effects of α^+-thalassemia and the sickle cell trait. *Nat Genet* 2005; **37**: 1253–1257.
35 Merry A, Roger JH, Curnow RN: A two-locus model for the inheritance of a familial disease. *Ann Hum Genet* 1979; **43**: 71–80.
36 McCallion AS, Stames E, Conlon RA, Chakravarti A: Phenotype variation in two-locus mouse models of Hirschsprung disease: tissue-specific interaction between Ret and Ednrb. *Proc Natl Acad Sci USA* 2003; **100**: 1826–1831.
37 Arking DE, Pfeufer A, Post W: A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat Genet* 2006; **38**: 644–651.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)