Published in final edited form as:
Genet Epidemiol. 2011 November ; 35(7): 706–721. doi:10.1002/gepi.20621.

Entropy-Based Information Gain Approaches to Detect and to Characterize Gene-Gene and Gene-Environment Interactions/ Correlations of Complex Diseases

R Fan[1], **M Zhong**[2], **S Wang**[1,3], **Y Zhang**[1], **A Andrew**[4], **M Karagas**[4], **H Chen**[5], **CI Amos**[6], **M Xiong**[7], and **J Moore**[8]

[1]Department of Statistics, Texas A&M University, College Station, TX 77843

[2]Global Pharmaceutical Research and Development, Abbott Laboratories 100 Abbott Park Rd, R436 AP9A-1, Abbott Park, IL 60064

[3]School of Information Science and Engineering, Yunnan University, Kunming 650091, P. R. China

[4]Department of Community and family Medicine, Dartmouth Medical School, Lebanon, NH 03756

[5]Surveillance Research Program, National Cancer Institute, 6116 Executive Blvd. #5016, Rockville, Maryland 20852

[6]Department of Epidemiology, MD Anderson Cancer Center, University of Texas, Houston, TX 77030

[7]Human Genetics Center, University of Texas, P. O. Box 20334, Houston, Texas 77225

[8]Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756

Abstract

For complex diseases, the relationship between genotypes, environment factors and phenotype is usually complex and nonlinear. Our understanding of the genetic architecture of diseases has considerably increased over the last years. However, both conceptually and methodologically, detecting gene-gene and gene-environment interactions remains a challenge, despite the existence of a number of efficient methods. One method that offers great promises but has not yet been widely applied to genomic data is the entropy-based approach of information theory. In this paper we first develop entropy-based test statistics to identify 2-way and higher order gene-gene and gene-environment interactions. We then apply these methods to a bladder cancer data set and thereby test their power and identify strengths and weaknesses. For two-way interactions, we propose an information-gain approach based on mutual information. For three-ways and higher order interactions, an interaction-information-gain approach is used. In both case we develop one-dimensional test statistics to analyze sparse data. Compared to the naive chi-square test, the test statistics we develop have similar or higher power and is robust. Applying it to the bladder cancer data set allowed to investigate the complex interactions between DNA repair gene SNPs, smoking status, and bladder cancer susceptibility. Although not yet widely applied, entropy-based approaches appear as a useful tool for detecting gene-gene and gene-environment interactions. The test statistics we develop add to a growing body methodologies that will gradually shed light on the complex architecture of common diseases.

Correspondence to: Dr. Ruzong Fan, Tel. 979-845-3152 or 3141 (main office), Fax 979-845-3144, rfan@stat.tamu.edu.

Keywords

gene-gene and gene-environment interactions; entropy; mutual information; interaction information; total correlation information

Introduction

Complex diseases result from mutual interactions between genetic variants and environmental factors and our understanding of their genetic architecture has considerably grown over the last decades. In recent years, there has been great enthusiasm to detect and to characterize gene-gene and gene-environment interactions of complex diseases using genome data [Mahdi et al., 2009; Moore and Williams, 2009; van-der-Woude et al., 2010; Wan et al., 2010; Zhang and Liu, 2007]. However, despite considerable effort, identifying and characterizing susceptibility genes of common complex human diseases and their network of interactions remains a great challenge. The challenge is both conceptual and technical: conceptually, it is not always clear how to define the interactions. There are two different arguments about gene-gene and gene-environment interactions: (1) statistical interaction, (2) biological interaction. Technically, traditional statistical approaches may not be useful because of the complexity and nonlinearity between complex traits and genetic, environment factors.

In traditional statistical models, i.e., linear models and generalized linear models such as logistic regressions, the genetic and environmental effects are decomposed into main and interaction effects [Fisher, 1918]. The statistical interactions are deviations from the main effects and don't make sense unless the main effect is significant. Moreover, the traditional statistical models may not work for high dimension sparse data. For instance, logistic regression models including interaction terms can fail to converge when some cells contain few individuals [Andrew et al., 2006]. Yet, one advantage of traditional statistical models is that the related theory is very mature and user-friendly softwares are available. For instance, variance partitioning and ANOVA are standard procedure in SAS for data analysis and model selection.

Biological interactions, on the other hand, happen at the cellular level and result from physical interactions between biomolecules such as DNA, RNA and proteins [Moore and Willams, 2009; Bateson, 1909; Bateson, 2002]. The biological gene-gene and gene-environment interaction is the interdependence between genetic and environmental factors and may cause complex diseases. In complex diseases, the relationship between genotypes, environmental factors and disease phenotypes is usually complex and nonlinear. Thus, biological interaction makes sense and it is valid in describing the complicated relation between genetic, environmental factors and disease phenotypes. In the absence of main effects, the biological gene-gene and gene-environment interactions may exist and can be important [Frankel and Schork, 1996]. However, the related theory to detect and to characterize the biological gene-gene and gene-environment interactions is not well-developed. There is a need to develop powerful methods and user-friendly softwares to identify and to interpret the complex genetic architecture of complex traits.

By using multiple genetic markers and environmental factors in analysis, it is usually a high-dimensional problem. For instance, assume we have two single nucleotide polymorphism (SNP) markers. Each of the two SNP has 3 genotypes, and then there are 9 genotype combinations if we consider the two SNPs simultaneously. If we add one environmental factor which has 2 categories, e.g., smoking vs non-smoking, there are $2 \times 9 = 18$ genotype-

environment combinations if we consider the two SNPs and the environmental factor simultaneously. Hence, one needs to handle the high-dimensional data.

In the multifactor dimensionality reduction (MDR) procedure, high dimensional genetic data are collapsed into a single dimensional variable allowing for such data to be analyzed [Hahn et al., 2003; Lou et al., 2007; Ritchie et al., 2001, 2003a, 2003b, 2004; Velez et al., 2007]. MDR is a non-parametric procedure that makes no assumption about the relationship between the phenotypes, the genetic, and the environmental factors. Since there was no alternative and powerful procedure, Andrew et al. [2006] ran logistic regression models to test three way interaction to replicate the findings of MDR. Unfortunately, the logistic regression models failed to converge due to the sparse nature of bladder cancer data. Thus, it is not only interesting but also necessary and important to develop novel statistical methods to detect and to characterize the complex biological gene-gene and gene-environment interactions of complex traits.

The traditional statistical models can not properly fit the nonlinear relationship between genotypes, environment factors and disease phenotypes in the absence of main effects. It may not be able and useful to model biological interactions. For the bladder cancer data of Andrew et al. [2006], the main effects of genetic polymorphisms were not observed and it is unclear if logistic regressions may fit the data well. The failure of convergence may be due to invalidness of the logistic regression model itself.

It is well-known that information theory based on entropy function is widely used to study nonlinear problems and complex system. The entropy function is a nonlinear transformation of interested variables. The entropy is commonly used in information theory to measure the uncertainty of random variables. The entropy-based approach is likely to be very useful to study the nonlinear relationship between genotypes, environment factors and phenotypes and to interpret the gene-gene and gene-environment interactions of complex diseases [Dong et al., 2008; Kang et al., 2008; Nothnagel et al., 2002]. In this article, we develop entropy-based approaches to detect and to characterize gene-gene and gene-environment interactions of complex diseases.

We start with the definition of entropy for genetic markers and environmental factors. Then, 2-way mutual information and information gain (IG) are introduced to describe gene-gene and gene-environment interactions. One idea of this article is to reduce high dimensional data to be a one-dimensional variable, and then to construct a χ^2-distribution statistic to test gene-gene interaction of complex diseases. We considered two di-allelic markers A and B in a case-control design. By using the information gain function, we reduce the 9-dimensional genotype combinations of the two markers to be a one-dimensional variable to construct the information gain based test T_{IG}. The method can be applied to test 2-way gene-environment interaction by treating the levels of environment factor as genotypes of a marker, i.e., one marker is replaced by the environment factor.

To generalize the 2-way methods to handle multiple K-way cases, $K \geq 3$, we need to distinguish two different concepts in information theory: interaction information and total correlation information (TCI). In 2-way case, the two concepts are the same. However, they are different in multiple K-way cases, $K \geq 3$. Roughly, the interaction information among multiple factors is the amount of information that is common to all the factors. The total correlation information, however, describes the total amount of dependence among all the factors. The K-way total correlation information can be decomposed into a summation of all lower and same order k-way interaction information, $2 \leq k \leq K$ [Jakulin, 2005].

We generalize the 2-way methods to detect and to characterize multiple K-way gene-gene and gene-environment interactions and correlations, $K \geq 3$. For multiple K-way interactions,

the *K*-way interaction information is proposed to extend the 2-way mutual information. For multiple *K*-way correlations, total correlation information is used. Correspondingly, the interaction information gain (IIG) and total correlation information gain (TCIG) can be defined as one-dimensional variables for case-control data. Thus, high-dimensional genetic data are collapsed as one-dimension variables via the information gains. Using the one-dimensional variables, test statistics are constructed which are χ_1^2-distributed. Compared with the naive χ^2 test statistics which usually have high degrees of freedom, the proposed information gain tests are easy to implement. In addition, the naive χ^2 tests are not always implementable due to sparse nature of high dimension genetic data.

Simulation study is performed to evaluate the robustness of the proposed test statistics by type I error rate calculations. Power analysis is carried out to show the usefulness of the proposed methods. The method is applied to bladder cancer data to explore gene-gene and gene-environment interactions and correlations of SNPs and smoking status with the disease. We use the bladder cancer data to show a forward selection procedure for the final model selection, and the procedure can be applied to the study of other complex traits.

Methods

In information theory, entropy measures the uncertainty associated with a random variable or a random system [Shannon, 1948]. The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\mathrm{E}[\log P(X)] = -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{1}$$

where $p(x) = P(X = x)$, $x \in \mathcal{X}$, is the probability mass function of the random variable X, and $\mathcal{X}$ is a finite set (e.g., {1, 2, ⋯, n}) or an enumerable infinite set (e.g., {1, 2, ⋯}) [Cover and Thomas, 2006]. The log is to the base 2. By definition, 0 log 0 = 0. The higher the entropy $H(X)$, the the higher the uncertainty we may predict the outcome about the variable X. The concept of the Shannon entropy has been used to select interesting combinations of polymorphisms for evaluating and for visualizing the information gain, which in turn allows for the detection of gene-gene and gene-environment interactions [Jakulin, 2005; Jakulin and Bratko, 2003, 2004; Jakulin et al., 2003; Moore et al., 2006; Wu et al., 2009].

Genotype and Environment-Based Entropy

For a case-control study design, we denote the disease status of an individual by D, and attribute the value $D = 0$ to healthy individuals (control) and the value $D = 1$ to affected ones (case). For explanation purposes, let us consider two di-allelic markers A and B (e.g., SNPs) and an environmental exposure E. Let us denote the alleles of markers A and B by A, a and B, b, respectively, and code the environmental factor E as $E = 0, 1, 2$ (e.g., non smoking, < 35 pack years, ≥ 35 pack years). There are three genotypes at marker A (AA, Aa, aa) and three genotypes at marker B (BB, Bb, bb). For convenience, we call each of the genotypes G_A at marker A and G_B at marker B by the number of A and B alleles present. That is,

$$G_A = \begin{cases} 2 & AA \\ 1 & Aa \\ 0 & aa \end{cases} \quad \text{and} \quad G_B = \begin{cases} 2 & BB \\ 1 & Bb \\ 0 & bb \end{cases}. \tag{2}$$

In the literature, genetic markers and environmental factors are treated as attributes [Jakulin and Bratko, 2003; Jakulin et al., 2003]. Using the entropy definition (1), we can define the

entropy $H(A)$ of marker A in the general population and the conditional entropy $H(A|D)$ in the affected population as

$$\begin{aligned} H(A) &= -\sum_{i=0}^{2} P(G_A=i)\log P(G_A=i), \\ H(A|D) &= -\sum_{i=0}^{2} P(G_A=i|D=1)\log P(G_A=i|D=1). \end{aligned} \tag{3}$$

Similarly, we may define the entropy $H(B)$ of marker B in the general population and the conditional entropy $H(B|D)$ in the affected population. For the environmental factor E, its entropy $H(E)$ in the general population and its conditional entropy $H(E|D)$ in the affected population can be defined, accordingly. When the markers A and B are combined, the entropy $H(A, B)$ in the general population and the conditional entropy $H(A, B|D)$ in the affected population are:

$$\begin{aligned} H(A,B) &= -\sum_{i=0}^{2}\sum_{j=0}^{2} P(G_A=i, G_B=j)\log P(G_A=i, G_B=j), \\ H(A,B|D) &= -\sum_{i=0}^{2}\sum_{j=0}^{2} P(G_A=i, G_B=j|D=1)\log P(G_A=i, G_B=j|D=1). \end{aligned} \tag{4}$$

Similarly, when one marker (e.g., A) and the environmental factor E are combined, the entropy $H(A, E)$ in the general population and the conditional entropy $H(A, E|D)$ in the affected population are:

$$\begin{aligned} H(A,E) &= -\sum_{i=0}^{2}\sum_{e=0}^{2} P(G_A=i, E=e)\log P(G_A=i, E=e), \\ H(A,E|D) &= -\sum_{i=0}^{2}\sum_{e=0}^{2} P(G_A=i, E=e|D=1)\log P(G_A=i, E=e|D=1). \end{aligned} \tag{5}$$

The entropy $H(B, E)$ in the general population and the conditional entropy $H(B, E|D)$ in the affected population can be defined in a similar manner. When both markers and the environmental factor are combined, the entropy $H(A, B, E)$ in the general population and the conditional entropy $H(A, B, E|D)$ in the affected population are:

$$\begin{aligned} H(A,B,E) &= -\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2} P(G_A=i, G_B=j, E=e)\log P(G_A=i, G_B=j, E=e), \\ H(A,B,E|D) &= -\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2} P(G_A=i, G_B=j, E=e|D=1)\log P(G_A=i, G_B=j, E=e|D=1). \end{aligned} \tag{6}$$

2-Way Mutual Information and Information Gain

The mutual information measures the interaction between two markers. In the general population, the mutual information of markers A and B, $I(A, B)$, is defined as [Shannon, 1948; Cover and Thomas, 2006]

$$\begin{aligned} I(A,B) &= H(A)+H(B)-H(A,B) \\ &= \sum_{i=0}^{2}\sum_{j=0}^{2} P(G_A=i, G_B=j)\log\frac{P(G_A=i, G_B=j)}{P(G_A=i)P(G_B=j)}. \end{aligned} \tag{7}$$

For the two markers A and B, $I(A, B) \geq 0$, and $I(A, B) = 0$ if and only if G_A and G_B are independent (i.e., $P(G_A = i, G_B = j) = P(G_A = i)P(G_B = j)$, p28 of Cover and Thomas [2006]).

Figure 1a) shows an I-diagram of $H(A)$, $H(B)$ and $I(A, B)$, and it is equivalent to a Venn diagram in set theory [Chanda et al., 2007; McGill, 1954; Yeung, 1991]. The left and right rectangles illustrate the magnitude of $H(A)$ and $H(B)$, respectively, and their overlap, colored in black, corresponds to the magnitude of the $I(A, B)$. In the affected population, the mutual information of markers A and B is defined as

$$I(A, B|D) = H(A|D) + H(B|D) - H(A, B|D) = \sum_{i=0}^{2}\sum_{j=0}^{2} P(G_A = i, G_B = j|D = 1)\log\frac{P(G_A = i, G_B = j|D = 1)}{P(G_A = i|D = 1)P(G_B = j|D = 1)}. \quad (8)$$

$I(A, B|D)$ measures the interaction between markers A and B given the disease. For the two markers A and B, $I(A, B|D) \geq 0$, and $I(A, B|D) = 0$ if and only if G_A and G_B are conditionally independent given the disease (i.e., $P(G_A = i, G_B = j|D = 1) = P(G_A = i|D = 1)P(G_B = j|D = 1)$).

The information gain of markers A and B in the presence of a disease can be defined as the difference between the mutual information in the affected population and that in the general population [Jakulin and Bratko, 2003, 2004; Jakulin et al., 2003; McGill, 1954; Moore et al., 2006]

$$IG(AB|D) = I(A, B|D) - I(A, B). \quad (9)$$

If the disease and the two markers are independent (i.e., $P(G_A = i, G_B = j|D = 1) = P(G_A = i, G_B = j)$), then $I(A, B|D) = I(A, B)$ and the information gain $IG(AB|D)$ is equal to 0. In that case, the interaction between markers A and B does not contribute to predicting disease risk. Hence, we can determine whether the gene-gene interaction between markers A and B predicts disease status by testing if the difference between estimates of mutual information is zero. Based on this rationale, we can build test statistics for practical applications.

For marker A or B and environmental factor E, the mutual information and the conditional mutual information can be defined as above. The information gain of marker A and environmental factor E in the presence of a disease can be defined as the difference $IG(AE|D) = I(A, E|D) - I(A, E)$. If the information gain is null, i.e., $IG(AE|D = 0)$, the marker A and the environmental factor E are independent of the disease status and the interaction between A and E does not predict disease status. Likewise, if the marker B and the environmental factor E are independent of the disease status D, then there is no information gain (i.e., $IG(BE|D) = I(B, E|D) - I(B, E) = 0$) and the interaction between B and E does not predict disease status.

3-Way Interaction Information and Total Correlation Information

The information gains $IG(AB|D)$, $IG(AE|D)$, and $IG(BE|D)$ represent 2-way interaction gains of two attributes given a disease. If we consider the three attributes A, B and E simultaneously, we can define the 3-way interaction information gain and the total correlation information as follows [Chanda et al., 2007; Han, 1980; McGill, 1954; Watanabe, 1960; Yeung, 1991]. In the general population, the 3-way interaction information of markers A and B and environmental factor E is defined as [Cover and Thomas, 2006, p49].

$$I(A,B,E)=-H(A)-H(B)-H(E)+H(A,B)+H(A,E)+H(B,E)-H(A,B,E).$$

The 3-way interaction information $I(A, B, E)$ contains interactions that can not be explained by the 2-way mutual information $I(A, B)$, $I(A, E)$, and $I(B, E)$. It represents the gain or loss of information by adding one attribute to a pair of attributes. Hence, the 3-way interaction information among attributes A, B and E can be understood as the amount of information that is common to all the attributes, but not present in any subset. The interaction information can be negative or positive.

Figure 1b) shows an I-diagram of $H(A)$, $H(B)$, $H(E)$, and $I(A, B, E)$ for two markers A and B and an environmental factor E [Chanda et al., 2007; McGill, 1954; Yeung-1991]. In Figure 1b), the black region corresponds to the magnitude of the $I(A, B, E)$. Compared to Figure 1a), Figure 1b) includes an additional rectangle corresponding to the magnitude of $H(E)$. If one attribute (e.g., the environmental factor E) is independent of two dependent attributes (e.g., the markers A and B), the interaction information $I(A, B, E)$ will be 0. This is because $P(G_A = i, G_B = j, E = e) = P(G_A = i, G_B = j)P(E = e)$ implies $P(G_A = i, E = e) = P(G_A = i)P(E = e)$ and $P(G_B = j, E = e) = P(G_B = j)P(E = e)$, and thus $I(A, B, E) = 0$. If all three attributes are independent, the interaction information $I(A, B, E)$ is of course equal to 0. Hence, $I(A, B, E)$ is an interaction among all three attributes.

The total correlation information is defined as the difference between the joint entropy $H(A, B, E)$ and the three individual entropies $H(A)$, $H(B)$ and $H(E)$, i.e.,

$$\begin{aligned} TCI(A,B,E) &= H(A)+H(B)+H(E)-H(A,B,E) \\ &= I(A,B)+I(A,E)+I(B,E)+I(A,B,E). \end{aligned} \tag{10}$$

The total correlation information describes the total amount of dependence among the three attributes A, B and E. It is always positive, or zero if and only if all the three attributes are independent, i.e., $P(G_A = i, G_B = j, E = e) = P(G_A = i)P(G_B = j)P(E = e)$. It will be different from zero even if only one pair of attributes are dependent. For instance, it is non-zero if the genetic markers A and B are independent of the environmental factor E but A and B are dependent or in linkage disequilibrium. Figure 1c) shows an I-diagram of $H(A)$, $H(B)$, $H(E)$, and $TCI(A, B, E)$ for two markers A and B and an environmental factor E [Chanda et al., 2007; McGill, 1954; Yeung-1991].

The second equality of relation (10) shows that the total correlation information $TCI(A, B, E)$ is equal to the summation of all 2-way mutual information $I(A, B)$, $I(A, E)$, $I(B, E)$, and 3-way interaction information $I(A, B, E)$. Thus, the 2-way mutual information and the 3-way interaction information can be seen as a decomposition of a 3-way dependency into a sum of 2-way and 3-way interactions [Jakulin, 2005]. The existence of 3-way correlations (i.e., $TCI(A, B, E) \neq 0$)) indicates the existence of some 2-way or 3-way interactions. On the other hand, the existence of 2-way or 3-way interactions can lead to 3-way correlations.

In the disease population, the 3-way interaction information $I(A, B, E|D)$ and total correlation information $TCI(A, B, E|D)$ of markers A and B and environmental factor E are defined as

$$I(A,B,E|D)=-H(A|D)-H(B|D)-H(E|D)+H(A,B|D)+H(A,E|D)+H(B,E|D)-H(A,B,E|D),$$
$$TCI(A,B,E|D)=H(A|D)+H(B|D)+H(E|D)-H(A,B,E|D).$$

The interaction information gain $IIG(ABE|D)$ and the total correlation information gain $TCIG(ABE|D)$ of markers A and B and environmental factor E in the presence of a disease can be defined as the differences [Jakulin and Bratko, 2003, 2004; Jakulin et al., 2003; McGill, 1954; Moore et al., 2006]

$$\begin{aligned} IIG(ABE|D) &= I(A, B, E|D) - I(A, B, E), \\ TCIG(ABE|D) &= TCI(A, B, E|D) - TCI(A, B, E). \end{aligned} \tag{11}$$

If the disease is independent of the two markers and the environmental factor E, $I(A, B, E|D) = I(A, B, E)$ and the information gain $IIG(ABE|D)$ is equal to 0. Similarly, $TCI(A, B, E|D) = TCI(A, B, E)$ and the total correlation information gain $TCIG(ABE|D)$ is equal to 0. Hence, we can test for the existence of gene-gene and gene-environment interactions or correlations between the disease and two markers A and B and the environmental factor E by testing if $IIG(ABE|D)$ and $TCIG(ABE|D)$ are zero. Based on this rationale, we can build test statistics for practical applications.

K-Way Interaction Information and Total Correlation Information

Suppose that we are interested in interactions or correlations between the disease and an arbitrary number K of attributes $\mathcal{A} = (A_1, \cdots, A_K)$, which can be genetic markers or environmental factors. For simplicity, we assume that each A_i can take three values 0, 1, or 2. For a vector of realization $\vec{a} = (a_1, \cdots, a_K)$ of $\mathcal{A} = (A_1, \cdots, A_K)$, we denote the joint probabilities as $P_{\vec{a}} = P(A_1 = a_1, \cdots, A_K = a_K) = P_{a_1 \cdots a_K}$ in the general population and as $Q_{\vec{a}} = P(A_1 = a_1, \cdots, A_K = a_K|D = 1) = Q_{a_1 \cdots a_K}$ in the affected population. Based on the joint probabilities, we can define the entropies $H(\mathcal{A}) = H(A_1, \cdots, A_K) = -\Sigma_{\vec{a}} P_{\vec{a}} \log P_{\vec{a}}$ and $H(\mathcal{A} | D) = H(A_1, \cdots, A_K | D) = -\Sigma_{\vec{a}} Q_{\vec{a}} \log Q_{\vec{a}}$.

For a subset $\mathcal{S} = (A_{i_1}, A_{i_2}, \cdots, A_{i_n}) \subseteq \mathcal{A} = (A_1, A_2, \cdots, A_K)$, we can define the related entropies $H(\mathcal{S})$ and $H(\mathcal{S} | D)$ in a similar manner. Here $\subseteq$ means that $\mathcal{S}$ is a subset of $\mathcal{A}$ and it can be equal to $\mathcal{A} = (A_1, A_2, \cdots, A_K)$. For a realization $\vec{s}$ of $\mathcal{S} = (A_{i_1}, A_{i_2}, \cdots, A_{i_n})$, the marginal probabilities are denoted as $P_{\vec{s}}$ and $Q_{\vec{s}}$. For individual attributes $A_1, \cdots, A_K$, the marginal probabilities and entropies are denoted as $P_{a_1, \cdots, \cdot}, \cdots, P_{\cdot, \cdots, a_K}$, $H(A_1), \cdots, H(A_K)$, $Q_{a_1, \cdots, \cdot}, \cdots, Q_{\cdot, \cdots, a_K}$, $H(A_1 | D), \cdots$, and $H(A_K | D)$. For a subset $\mathcal{S}$, let us denote $|\mathcal{S}| = |(A_{i_1}, A_{i_2}, \cdots, A_{i_n})| = n$, i.e., the number of attributes of $\mathcal{S}$. The K-way interaction information can be defined as [McGill, 1954; Han, 1980; Yeung, 1991]

$$\begin{aligned} I(\mathscr{A}) &= I(A_1, A_2, \cdots, A_K) = -\sum_{\mathscr{S} \subseteq \mathscr{A}} (-1)^{|\mathscr{A}| - |\mathscr{S}|} H(\mathscr{S}), \\ I(\mathscr{A}|D) &= I(A_1, A_2, \cdots, A_K|D) = -\sum_{\mathscr{S} \subseteq \mathscr{A}} (-1)^{|\mathscr{A}| - |\mathscr{S}|} H(\mathscr{S}|D). \end{aligned}$$

The K-way interaction information gain is defined as $IIG(\mathcal{A} | D) = I(\mathcal{A} | D) - I(\mathcal{A})$.

In the general population, the K-way total correlation information is defined as the difference between the summation of the individual entropies $H(A_1), \cdots, H(A_K)$ and the joint entropy $H(\mathcal{A})$ [Jakulin, 2005; Watanabe, 1960; Chanda et al., 2007], i.e.,

$$\begin{aligned} TCI(\mathscr{A}) &= H(A_1) + \cdots + H(A_K) - H(A_1, \cdots, A_K) \\ &= \sum_{\mathscr{S} \subseteq \mathscr{A}, |\mathscr{S}| \geq 2} I(\mathscr{S}). \end{aligned} \tag{12}$$

The total correlation information $TCI(\mathcal{A})$ is the total amount of dependence among all the attributes $\mathcal{A} = (A_1, \cdots, A_K)$. As the second equality in relation (12) shows, the total correlation information $TCI(\mathcal{A})$ is equal to the summation of all interaction information $I(\mathcal{S})$ including 2-way mutual information, $\mathcal{S} \subseteq \mathcal{A}$. Thus, the interaction information can be seen as a decomposition of a K-way dependency into a sum of k-way interactions, $k \leq K$ [Jakulin, 2005]. The existence of K-way correlations indicates the existence of some k-way interactions, $k \leq K$. On the other hand, the existence of low order k-way interactions can lead to high order K-way correlations.

In the affected population, the K-way total correlation information is defined as the difference between the summation of the individual entropies $H(A_1 \mid D), \cdots, H(A_K \mid D))$ and the joint entropy $H(\mathcal{A} \mid D))$, i.e.,

$$TCI(\mathcal{A}|D)=H(A_1|D)+\cdots+H(A_K|D) - H(A_1,\cdots,A_K|D).$$

The K-way total correlation information gain is defined as $TCIG(\mathcal{A} \mid D) = TCI(\mathcal{A} \mid D) - TCI(\mathcal{A})$. If the disease is independent of the attributes, the interaction information gain $IIG(\mathcal{A} \mid D)$ is equal to 0 and similarly, the total correlation information gain $TCIG(\mathcal{A} \mid D)$ is equal to 0. The test statistics can be built accordingly to test the interaction or correlation between the disease and the attributes $\mathcal{A} = (A_1, A_2, \cdots, A_K)$.

Test Statistics Based on the 2-Way Mutual Information Gain

Based on the above discussion about 2-way mutual information and information gain, we can construct test statistics to detect gene-gene and gene-environment interactions. In what follows, we discuss only the construction of a test statistic to detect a gene-gene interaction between markers A and B. The same procedure can be applied to construct a test statistic to detect a gene-environment interaction between marker A (or B) and environmental factor E.

Consider a case-control design with M controls from an unaffected population and N cases from an affected population. Assume that each individual in the sample is typed at both markers A and B. Let us denote by X_{ij} the count of controls whose genotypes are $(G_A = i, G_B = j)$, and by Y_{ij} he count of cases whose genotypes are $(G_A = i, G_B = j)$, $i, j = 0, 1, 2$. The test statistics can be built based on the column vectors $X = (X_{00}, X_{01}, X_{02}, X_{10}, X_{11}, X_{12}, X_{20}, X_{21})^\tau$ and $Y = (Y_{00}, Y_{01}, Y_{02}, Y_{10}, Y_{11}, Y_{12}, Y_{20}, Y_{21})^\tau$. Hereafter, the superscript τ denotes the transpose of a vector or a matrix. To remove redundancies, X_{22} is not included in X, and Y_{22} is not included in Y. Before defining our test statistics, let us introduce some notations.

In the general population, we denote the joint genotype probabilities for markers A and B by $P_{ij} = P(G_A = i, G_B = j)$. For the affected population, we denote the joint conditional genotype probabilities for markers A and B by $Q_{ij} = P(G_A = i, G_B = j|D = 1)$. One can see that P_{ij} and Q_{ij} both sum to 1 and that some parameters are redundant. Let us denote $P = (P_{00}, P_{01}, P_{02}, P_{10}, P_{11}, P_{12}, P_{20}, P_{21})^\tau$ and $Q = (Q_{00}, Q_{01}, Q_{02}, Q_{10}, Q_{11}, Q_{12}, Q_{20}, Q_{21})^\tau$. Both the column counting vector $\begin{pmatrix} X \\ X_{22} \end{pmatrix}$ and the column counting vector $\begin{pmatrix} Y \\ Y_{22} \end{pmatrix}$ follow a multinomial distribution. The mean vectors of X and Y are MP and NQ, respectively, and the variance-covariance matrix of X and Y are $M[\text{diag}(P) - PP^\tau]$ and $N[\text{diag}(Q) - QQ^\tau]$, respectively. In the following, let us denote $\Sigma = \text{diag}(P) - PP^\tau$ and $\Sigma_D = \text{diag}(Q) - QQ^\tau$.

The sample mean $\bar{X} = X/M$ serves as the estimate of P and the sample mean $\bar{Y} = Y/N$ serves as the estimate of Q. Assume that the sample sizes M and N are large enough that the large

sample theory applies. By the multivariate central limit theorem of large sample theory, $\sqrt{M}\left[\overline{X}-P\right]$ can be approximated by a multivariate normal distribution with a zero mean vector and a variance-covariance matrix Σ [Lehmann, 1983, Theorem 5.1.8, p343]. Similarly, $\sqrt{N}\left[\overline{Y}-Q\right]$ can be approximated by a multivariate normal distribution with a zero mean vector and a variance-covariance matrix Σ_D.

Now, let us define

$$f_{ij}=P_{ij}\log\frac{P_{ij}}{P_{i\cdot}P_{\cdot j}},\quad g_{ij}=Q_{ij}\log\frac{Q_{ij}}{Q_{i\cdot}Q_{\cdot j}},$$

where $P_{i\cdot}=\sum_{j=0}^{2}P_{ij}$, $P_{\cdot j}=\sum_{i=0}^{2}P_{ij}$, $Q_{i\cdot}=\sum_{j=0}^{2}Q_{ij}$ and $Q_{\cdot j}=\sum_{i=0}^{2}Q_{ij}$. Let $f=I(A,B)=\sum_{i=0}^{2}\sum_{j=0}^{2}f_{ij}$ and $g=I(A,B|D)=\sum_{i=0}^{2}\sum_{j=0}^{2}g_{ij}$. Then, the information gain can be expressed as

$$IG(AB|D)=g-f=\sum_{i=0}^{2}\sum_{j=0}^{2}g_{ij}-\sum_{i=0}^{2}\sum_{j=0}^{2}f_{ij}.$$

We denote the partial derivatives of functions f and g as $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$, which are column vectors. The elements of $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ are given in the Appendix A as

$$\begin{aligned}\frac{\partial f}{\partial P_{ij}}&=\log\frac{P_{ij}}{P_{i\cdot}P_{\cdot j}}-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}},\\ \frac{\partial g}{\partial Q_{ij}}&=\log\frac{Q_{ij}}{Q_{i\cdot}Q_{\cdot j}}-\log\frac{Q_{22}}{Q_{2\cdot}Q_{\cdot 2}}.\end{aligned}\qquad(13)$$

We further denote $\Lambda=\left[\frac{\partial f}{\partial P}\right]^{\tau}\sum\frac{\partial f}{\partial P}/M+\left[\frac{\partial g}{\partial Q}\right]^{\tau}\sum_{D}\frac{\partial g}{\partial Q}/N$.

We denote the estimate of P_{ij} as $\hat{P}_{ij}=X_{ij}/M$ and the estimate of Q_{ij} as $\hat{Q}_{ij}=Y_{ij}/N$. Similarly, we denote the estimates of other parameters as $\widehat{P}_{i\cdot}=\sum_{j=0}^{2}\widehat{P}_{ij}$, etc. Then, the estimates $\hat{f}$, $\hat{\Lambda}$, and $\hat{g}$ of f, Λ, and g can be calculated by replacing P_{ij} and Q_{ij} using $\hat{P}_{ij}$ and $\hat{Q}_{ij}$.

Based on the large sample theory, $\sqrt{M}(\widehat{f}-f)$ tends to a normal distribution with a zero mean and a variance $\left[\frac{\partial f}{\partial P}\right]^{\tau}\sum\frac{\partial f}{\partial P}$. Similarly, $\sqrt{N}(\widehat{g}-g)$ tends to a normal distribution with a zero mean and a variance $\left[\frac{\partial g}{\partial Q}\right]^{\tau}\sum_{D}\frac{\partial g}{\partial Q}$ [Lehmann, 1983, Theorem 2.5.3, p112]. Note that $f=I(A,B)=I(A,B|D)=g$ under the null hypothesis of independence between the disease and the two markers A and B, and so $\hat{g}-\hat{f}=(\hat{g}-g)-(\hat{f}-f)$ tends to a normal distribution with a zero mean and a variance Λ. With these discussions in mind, the statistical tests can be constructed as

$$T_{IG}=(\widehat{g}-\widehat{f})^2/\widehat{\Lambda},$$
$$T=(\widehat{P}-\widehat{Q})^{\tau}\left(\frac{\widehat{\Sigma}}{M}+\frac{\widehat{\Sigma}_D}{N}\right)^{-1}(\widehat{P}-\widehat{Q}). \quad (14)$$

The test T_{IG} is based on the information gain $IG(AB \mid D)$. The test T is a naive χ^2-distributed statistic, which is based on the 2 by 8 contingency table to compare the counts of case and controls for genotype combinations of markers A and B. Under the null hypothesis that the two markers are independent of the disease, the test T_{IG} is centrally χ^2_1-distributed with 1 degree of freedom and the test T is centrally χ^2_8-distributed with 8 degrees of freedom. Under the alternative hypothesis that the disease and the two markers are not independent, the test T_{IG} is non-centrally χ^2_1-distributed with a non-centrality parameter $\lambda_{IG} = (g - f)^2/\Lambda$ and the test T is non-centrally χ^2_8-distributed with a non-centrality parameter $\lambda_T=(P-Q)^{\tau}\left(\frac{\Sigma}{M}+\frac{\Sigma_D}{N}\right)^{-1}(P-Q)$.

The statistics T_{IG} and T are overall test statistics to test the association between the markers A and B and the disease. If the markers are associated with the disease (i.e., the markers are not independent of the disease), we need to know which genotypes are associated with the disease. For genotype ($G_A = i$, $G_B = j$), we can test if it is associated with the disease using one of the two following tests

$$T_{E,ij}=\frac{(\widehat{g}_{ij}-\widehat{f}_{ij})^2}{\widehat{\Lambda}_{ij}},$$
$$T_{ij}=\frac{(\widehat{P}_{ij}-\widehat{Q}_{ij})^2}{\widehat{\mathrm{Var}}(\widehat{P}_{ij}-\widehat{Q}_{ij})}, \quad (15)$$

where $\hat{\Lambda}_{ij}$ is the estimate of Λ_{ij} and $\widehat{\mathrm{Var}}(\widehat{P}_{ij} - \widehat{Q}_{ij})$ is the estimate of $\mathrm{Var}(\hat{P}_{ij} - \hat{Q}_{ij})$. The estimates $\hat{\Lambda}_{ij}$ and $\widehat{\mathrm{Var}}(\widehat{P}_{ij} - \widehat{Q}_{ij})$ are given by

$$\Lambda_{ij}=\frac{\left[\frac{\partial f_{ij}}{\partial P}\right]^{\tau}\Sigma\frac{\partial f_{ij}}{\partial P}}{M}+\frac{\left[\frac{\partial g_{ij}}{\partial Q}\right]^{\tau}\Sigma_D\frac{\partial g_{ij}}{\partial Q}}{N},$$
$$\mathrm{Var}(\widehat{P}_{ij} - \widehat{Q}_{ij})=\frac{P_{ij}(1-P_{ij})}{M}+\frac{Q_{ij}(1-Q_{ij})}{N}.$$

The test T_{ij} compares the difference $\hat{P}_{ij} - \hat{Q}_{ij}$ of the proportions of cases and controls with genotype ($G_A = i$, $G_B = j$), and the test $T_{E,ij}$ is based on the difference $\hat{f}_{ij} - \hat{g}_{ij}$. If genotype ($G_A = i$, $G_B = j$) is strongly associated with the disease, the differences $\hat{P}_{ij} - \hat{Q}_{ij}$ and $\hat{f}_{ij} - \hat{g}_{ij}$ tend to be different from 0 and significant results are likely to be found using the tests T_{ij} and/or $T_{E,ij}$.

Under the null hypothesis that the disease and the genotype ($G_A = i$, $G_B = j$) are independent, the test $T_{E,ij}$ and T_{ij} are centrally χ^2_1-distributed. Under the alternative hypothesis that the disease and the genotype ($G_A = i$, $G_B = j$) are not independent, the test $T_{E,ij}$ and T_{ij} are non-centrally χ^2_1-distributed with non-centrality parameters $\lambda_{E,ij} = (g_{ij} - f_{ij})^2/\Lambda_{ij}$ and $\lambda_{ij} = (P_{ij} - Q_{ij})^2/\mathrm{Var}(\hat{P}_{ij} - \hat{Q}_{ij})$, respectively.

Test Statistics Based on the 3-Way Interaction Information Gain and Total Correlation Information Gain

Again, consider a case-control design with M controls from an unaffected population and N cases from an affected population. Let us denote by X_{ije} the count of controls whose genotypes are ($G_A = i$, $G_B = j$, $E = e$) and by Y_{ije} the count of cases whose genotypes are ($G_A = i$, $G_B = j$, $E = e$), $i, j, e = 0, 1, 2$. The test statistics can be built based on two column vectors X and Y, where X includes all X_{ije} except X_{222}, and Y includes all Y_{ije} except Y_{222} to remove the redundancy.

In the general population, we denote the joint genotype probabilities for markers A and B and environmental factor E by $P_{ije} = P(G_A = i, G_B = j, E = e)$. In the affected population, we denote the joint conditional genotype probabilities by $Q_{ije} = P(G_A = i, G_B = j, E = e|D = 1)$. Let us denote a column vector P which includes all P_{ije} except P_{222} and we denote a column vector Q which includes all Q_{ije} except Q_{222}. The column counting vectors $\begin{pmatrix} X \\ X_{222} \end{pmatrix}$ and $\begin{pmatrix} Y \\ Y_{222} \end{pmatrix}$ follow the multinomial distributions $\left(M, \begin{pmatrix} P \\ P_{222} \end{pmatrix}\right)$ and $\left(N, \begin{pmatrix} Q \\ Q_{222} \end{pmatrix}\right)$, respectively. The mean vector of $\bar{X} = X/M$ is P and that of $\bar{Y} = Y/N$ is Q. The variance-covariance matrices of X and Y are $M\Sigma$ and $N\Sigma_D$, respectively, where $\Sigma = \text{diag}(P) - PP^{\tau}$ and $\Sigma_D = \text{diag}(Q) - QQ^{\tau}$. As before, we assume that the sample sizes M and N are large enough for the large sample theory to apply. Based on the multivariate central limit theorem, $\sqrt{M}\left[\bar{X} - P\right]$ and $\sqrt{N}\left[\bar{Y} - Q\right]$ tend to a multivariate normal distribution with a zero mean vector and variance-covariance matrices Σ and Σ_D, respectively [Lehmann 1983, Theorem 5.1.8, p343].

Denote $P_{i\cdot\cdot} = \sum_{j=0}^{2}\sum_{e=0}^{2} P_{ije}$, $P_{\cdot j\cdot} = \sum_{i=0}^{2}\sum_{e=0}^{2} P_{ije}$, $P_{\cdot\cdot e} = \sum_{i=0}^{2}\sum_{j=0}^{2} P_{ije}$. Similarly, $Q_{i\cdot\cdot}$, $Q_{\cdot j\cdot}$, and $Q_{\cdot\cdot e}$ can be defined in a similar manner. Now, let us define

$$f_{ije} = P_{ije}\log\frac{P_{ije}}{P_{i\cdot\cdot}P_{\cdot j\cdot}P_{\cdot\cdot e}}, \quad g_{ije} = Q_{ije}\log\frac{Q_{ije}}{Q_{i\cdot\cdot}Q_{\cdot j\cdot}Q_{\cdot\cdot e}},$$

If $f = TCI(A, B, E) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2} f_{ije}$ and $g = TCI(A, B, E|D) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2} g_{ije}$, the total correlation information gain can be expressed as

$$TCIG(ABE|D) = g - f = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{0}^{2} g_{ije} - \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2} f_{ije}.$$

We denote the partial derivatives of functions f and g as $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$, which are column vectors. The elements of $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ are given in the Supplementary Materials Appendix A as

$$\begin{aligned} \frac{\partial f}{\partial P_{ije}} &= \log\frac{P_{ije}}{P_{i\cdot\cdot}P_{\cdot j\cdot}P_{\cdot\cdot e}} - \log\frac{P_{222}}{P_{2\cdot\cdot}P_{\cdot 2\cdot}P_{\cdot\cdot 2}}, \\ \frac{\partial g}{\partial Q_{ije}} &= \log\frac{Q_{ije}}{Q_{i\cdot\cdot}Q_{\cdot j\cdot}Q_{\cdot\cdot e}} - \log\frac{Q_{222}}{Q_{2\cdot\cdot}Q_{\cdot 2\cdot}Q_{\cdot\cdot 2}}. \end{aligned} \tag{16}$$

Here, $\Lambda = \left[\frac{\partial f}{\partial P}\right]^{\tau}\sum\frac{\partial f}{\partial P}/M + \left[\frac{\partial g}{\partial Q}\right]^{\tau}\sum_{D}\frac{\partial g}{\partial Q}/N$.

To build a 3-way interaction information gain based test statistic, we denote $P_{ij\cdot}=\sum_{e=0}^{2}P_{ije}$, $P_{\cdot je}=\sum_{i=0}^{2}P_{ije}$, $P_{i\cdot e}=\sum_{j=0}^{2}P_{ije}$; and we define $Q_{ij\cdot}$, $Q_{\cdot je}$ and $Q_{i\cdot e}$ in a similar manner. Let us define

$$h_{ije}=P_{ije}\log\frac{P_{ije}P_{i\cdot\cdot}P_{\cdot j\cdot}P_{\cdot\cdot e}}{P_{ij\cdot}P_{\cdot je}P_{i\cdot e}},\qquad \ell_{ije}=Q_{ije}\log\frac{Q_{ije}Q_{i\cdot\cdot}Q_{\cdot j\cdot}Q_{\cdot\cdot e}}{Q_{ij\cdot}Q_{\cdot je}Q_{i\cdot e}},$$

If $h=I(A,B,E)=\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2}h_{ije}$ and $\ell=I(A,B,E|D)=\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2}\ell_{ije}$, the 3-way interaction information gain of markers A and B and environmental factor E can be expressed as

$$IIG(ABE|D)=\ell-h=\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{0}^{2}\ell_{ije}-\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{e=0}^{2}h_{ije}.$$

We denote the partial derivatives of functions h and ℓ as $\frac{\partial h}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$, which are column vectors. The elements of $\frac{\partial h}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$ are given in the Supplementary Materials Appendix B as

$$\begin{aligned}\frac{\partial h}{\partial P_{ije}}&=\log\frac{P_{ije}P_{i\cdot\cdot}P_{\cdot j\cdot}P_{\cdot\cdot e}}{P_{ij\cdot}P_{i\cdot e}P_{\cdot je}}-\log\frac{P_{222}P_{2\cdot\cdot}P_{\cdot 2\cdot}P_{\cdot\cdot 2}}{P_{22\cdot}P_{2\cdot 2}P_{\cdot 22}},\\ \frac{\partial \ell}{\partial Q_{ije}}&=\log\frac{Q_{ije}Q_{i\cdot\cdot}Q_{\cdot j\cdot}Q_{\cdot\cdot e}}{Q_{ij\cdot}Q_{i\cdot e}Q_{\cdot je}}-\log\frac{Q_{222}Q_{2\cdot\cdot}Q_{\cdot 2\cdot}Q_{\cdot\cdot 2}}{Q_{22\cdot}Q_{2\cdot 2}Q_{\cdot 22}}.\end{aligned}\tag{17}$$

Here, $\Gamma=\left[\frac{\partial h}{\partial P}\right]^{\tau}\sum\frac{\partial h}{\partial P}/M+\left[\frac{\partial \ell}{\partial Q}\right]^{\tau}\sum_{D}\frac{\partial \ell}{\partial Q}/N$.

We denote the estimate of P_{ije} and Q_{ije} as $\hat{P}_{ije}=X_{ije}/M$ and $\hat{Q}_{ije}=Y_{ije}/N$, respectively. Similarly, we denote the estimates of other parameters as $\widehat{P}_{i\cdot\cdot}=\sum_{j=0}^{2}\sum_{e=0}^{2}\widehat{P}_{ije}$, etc. Then, the estimates $\hat{f}$, $\hat{g}$, $\hat{h}$, $\hat{\ell}$, $\hat{\Lambda}$, and $\hat{\Gamma}$ of f, g, h, ℓ, Λ, and Γ can be calculated by replacing P_{ije} and Q_{ije} using $\hat{P}_{ije}$ and $\hat{Q}_{ije}$. The statistical tests to test the correlations and interactions between markers A and B, environmental factor E and the disease can be constructed by

$$\begin{aligned}T_{TCIG}&=(\widehat{g}-\widehat{f})^{2}/\widehat{\Lambda},\\ T_{IIG}&=(\widehat{h}-\widehat{\ell})^{2}/\widehat{\Gamma}.\end{aligned}\tag{18}$$

The test T_{TCIG} is based on the total correlation information gain $TCIG(ABE|D)$ and can be used to test for the existence of 3-way correlations. The test T_{IIG}, on the other hand, is based on the 3-way interaction information gain $IIG(ABE|D)$ and can be used to test for the existence of 3-way interactions. Under the null hypothesis that the two markers A and B, and the environmental factor E are independent of the disease, the test statistics T_{TCIG} and T_{IIG} are centrally χ_1^2-distributed. Under the alternative hypothesis that the two markers and the environmental factor are not independent of the disease, the test statistics T_{TCIG} and T_{IIG} are non-centrally χ_1^2-distributed with non-centrality parameters $\lambda_{TCIG}=(g-f)^2/\Lambda$ and $\lambda_{IIG}=(h-\ell)^2/\Gamma$, respectively.

Test Statistics Based on the *K*-Way Interaction Information Gain and Total Correlation Information Gain

Given a case-control sample with M controls and N cases, we are going to construct test statistics to test K-way interactions between K attributes $A_1, \cdots, A_K$. Hereafter, $|\vec{s}|$ is the number of elements in a vector $\vec{s}$. The approach is similar to the one we used for lower-order interactions. Let us denote

$$TCIG(\mathscr{A}|D)=g-f=\sum_{a_1=0}^{2}\cdots\sum_{a_K=0}^{2}[g_{a_1\cdots a_K}-f_{a_1\cdots a_K}],$$

$$g_{a_1\cdots a_K}=Q_{a_1\cdots a_K}\log\frac{Q_{a_1\cdots a_K}}{\prod_{\vec{s}\subset(a_1,\cdots,a_K)\atop|\vec{s}|=1}Q_{\vec{s}}},$$

$$f_{a_1\cdots a_K}=P_{a_1\cdots a_K}\log\frac{P_{a_1\cdots a_K}}{\prod_{\vec{s}\subset(a_1,\cdots,a_K)\atop|\vec{s}|=1}P_{\vec{s}}},$$

$$\prod_{\vec{s}\subset(a_1,\cdots,a_K)\atop|\vec{s}|=1}Q_{\vec{s}}=Q_{a_1,\cdots,}\cdots Q_{\cdots,a_K}$$

where is the product of all individual marginal probabilities of $A_1, \cdots, A_K$ in the affected population and

$$\prod_{\vec{s}\subset(a_1,\cdots,a_K)\atop|\vec{s}|=1}P_{\vec{s}}=P_{a_1,\cdots,}\cdots P_{\cdots,a_K}$$

is the product of all individual marginal probabilities of $A_1, \cdots, A_K$ in the general population. We denote the partial derivatives of functions f and g as $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$, which are column vectors. The elements of $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ are given as

$$\frac{\partial f}{\partial P_{a_1\cdots a_K}}=\log\frac{P_{a_1\cdots a_K}}{P_{a_1,\cdots,}\cdots P_{\cdots,a_K}}-\log\frac{P_{2\cdots2}}{P_{2,\cdots,}\cdots P_{\cdots,2}},$$

$$\frac{\partial g}{\partial Q_{a_1\cdots a_K}}=\log\frac{Q_{a_1\cdots a_K}}{Q_{a_1,\cdots,}\cdots Q_{\cdots,a_K}}-\log\frac{Q_{2\cdots2}}{Q_{2,\cdots,}\cdots Q_{\cdots,2}},\tag{19}$$

which can be proven along the vein of relation (16) in Supplementary Materials Appendix C. We define $\Lambda=\left[\frac{\partial f}{\partial P}\right]^{\tau}\sum\frac{\partial f}{\partial P}/M+\left[\frac{\partial g}{\partial Q}\right]^{\tau}\sum_{D}\frac{\partial g}{\partial Q}/N$, where P is a column vector including all $P_{a1\cdots aK}$ except $P_{2\cdots2}$, $\Sigma = \text{diag}\,(P)-PP^{\tau}$, Q is a column vector including all $Q_{a1\cdots aK}$ except $Q_{2\cdots2}$, and $\Sigma_D = \text{diag}\,(Q)-QQ^{\tau}$.

Similarly, let us denote

$$IIG(\mathscr{A}|D)=\ell - h=\sum_{a_1=0}^{2}\cdots\sum_{a_K=0}^{2}[\ell_{a_1\cdots a_K}-h_{a_1\cdots a_K}],$$

$$\ell_{a_1\cdots a_K}=Q_{a_1\cdots a_K}\log\frac{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}Q_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=0}}{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}Q_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=1}},$$

$$h_{a_1\cdots a_K}=P_{a_1\cdots a_K}\log\frac{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}P_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=0}}{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}P_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=1}},$$

where $|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=0$ means that the subset $(a_1,\cdots,a_K)\setminus\vec{s}$ contains an even number of elements, and $|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=1$ means that the subset $(a_1,\cdots,a_K)\setminus\vec{s}$ contains an odd number of elements. Moreover, the product $\prod_{\vec{s}\subset(a_1,\cdots,a_K)}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=0}Q_{\vec{s}}$ does not contain $Q_{a_1,\cdots,a_K}$ since $\vec{s}\subset(a_1,\cdots,a_K)$ means that $\vec{s}$ is a real subset of $(a_1,\cdots,a_K)$ [i.e., $\vec{s}\neq(a_1,\cdots,a_K)$]. The same logic applies to the other products. We denote the partial derivatives of functions ℓ and h as $\frac{\partial f}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$, which are column vectors. The elements of $\frac{\partial h}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$ are given as

$$\frac{\partial h}{\partial P_{a_1\cdots a_K}}=\log\frac{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}P_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=0}}{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}P_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=1}}-\log\frac{\prod_{\vec{s}\subset(2,\cdots,2)}P_{\vec{s}}^{|(2,\cdots,2)\setminus\vec{s}|_{mod(2)}=0}}{\prod_{\vec{s}\subset(2,\cdots,2)}P_{\vec{s}}^{|(2,\cdots,2)\setminus\vec{s}|_{mod(2)}=1}},$$

$$\frac{\partial \ell}{\partial Q_{a_1\cdots a_K}}=\log\frac{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}Q_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=0}}{\prod_{\vec{s}\subset(a_1,\cdots,a_K)}Q_{\vec{s}}^{|(a_1,\cdots,a_K)\setminus\vec{s}|_{mod(2)}=1}}-\log\frac{\prod_{\vec{s}\subset(2,\cdots,2)}Q_{\vec{s}}^{|(2,\cdots,2)\setminus\vec{s}|_{mod(2)}=0}}{\prod_{\vec{s}\subset(2,\cdots,2)}Q_{\vec{s}}^{|(2,\cdots,2)\setminus\vec{s}|_{mod(2)}=1}},\quad(20)$$

which can be proven along the vein of relation (17) in Supplementary Materials Appendix C. Here, $\Gamma=[\frac{\partial h}{\partial P}]^{\tau}\sum\frac{\partial h}{\partial P}/M+\left[\frac{\partial \ell}{\partial Q}\right]^{\tau}\sum_{D}\frac{\partial \ell}{\partial Q}/N$. To test for the existence of K-way interactions and correlations between the disease and the attributes $A_1,\cdots,A_K$, the χ_1^2-distributed test statistics can be constructed as $T_{TCIG}=(\hat{g}-\hat{f})^2/\hat{\Lambda}$ and $T_{IIG}=(\hat{h}-\hat{\ell})^2/\hat{\Gamma}$, respectively.

Association Test Statistics based on 1-way Entropy Loss

Suppose that we are interested in testing for the existence of an association between an attribute and a disease in a case-control study. The entropy of the attribute can be used as the basis to construct test statistics. The attribute here can be a single marker or an environment factor. In addition, if two or more markers are in strong linkage disequilibrium and their haplotype data are available, the haplotype data can be treated as an attribute. Here we use marker A as the attribute. It is well-known that the entropy is maximized when a system reaches its equilibrium state. In the one locus case, the equilibrium state refers to the Hardy-

Weinberg equilibrium. Since the assumption of Hardy-Weinberg equilibrium is likely to be true in the general population, the entropy $H(A)$ can reach the maximum. In the affected population, the assumption of Hardy-Weinberg equilibrium may not be true and the conditional entropy $H(A|D)$ may decrease. The entropy loss of marker A in the presence of a disease can be defined as follows:

$$EL(A|D)=H(A)-H(A|D). \tag{21}$$

If the disease and marker A are independent, $H(A|D) = H(A)$. Then the entropy loss $EL(A|D)$ is equal to 0. Hence, we can test for the existence of an association between marker A and a disease by testing if the entropy loss is zero. Based on this rationale, we can build test statistics for practical applications.

In the general population, we denote the genotype probabilities for marker A by $P_i = P(G_A = i)$. In the affected population, we denote the conditional genotype probabilities for marker A by $Q_i = P(G_A = i|D = 1)$. Here $P = (P_0, P_1)^\tau$, and $Q = (Q_0, Q_1)^\tau$. The entropy loss can be expressed as

$$EL(A|D)=H(A)-H(A|D)=\sum_{i=0}^{2}P_i\log P_i-\sum_{i=0}^{2}Q_i\log Q_i.$$

We denote the partial derivatives of the entropy functions $H(A)$ and $H(A|D)$ as the column vectors $\frac{\partial H(A)}{\partial P}$ and $\frac{\partial H(A|D)}{\partial Q}$. We can show that the elements of $\frac{\partial H(A)}{\partial P}$ and $\frac{\partial H(A|D)}{\partial Q}$ are given by

$$\begin{aligned}\frac{\partial H(A)}{\partial P_i}&=\log P_i-\log P_2,\\ \frac{\partial H(A|D)}{\partial Q_i}&=\log Q_i-\log Q_2.\end{aligned} \tag{22}$$

For a case-control design with M controls and N cases, let us denote by X_i the count of controls whose genotypes are $(G_A = i)$ and by Y_i the count of cases whose genotypes are $(G_A = i)$, i, = 0, 1, 2. We denote the estimates of P_i and Q_i as $\hat{P}_i = X_i/M$ and $\hat{Q}_i = Y_i/N$. Then, the estimates $\hat{H}(A)$, $\hat{\Omega}$, and $\hat{H}(A|D)$ of $H(A)$, Ω, and $H(A|D)$ can be calculated by replacing P_i and Q_i using $\hat{P}_i$ and $\hat{Q}_i$. Here $\Omega=\left[\frac{\partial H(A)}{\partial P}\right]^\tau\sum\frac{\partial H(A)}{\partial P}/M+\left[\frac{\partial H(A|D)}{\partial Q}\right]^\tau\sum_D\frac{\partial H(A|D)}{\partial Q}/N$, where $\Sigma = \text{diag}(P) - PP^\tau$ and $\Sigma_D = \text{diag}(Q) - QQ^\tau$. The entropy loss-based statistics to test for the existence of an association between marker A and a disease can be constructed by

$$T_{EL}=\left(\widehat{H}(A)-\widehat{H}(A|D)\right)^2/\widehat{\Omega}.$$

Under the null hypothesis, T_{EL} is centrally χ_1^2-distributed. Under the alternative hypothesis, T_{EL} is non-centrally χ_1^2-distributed with a non-centrality parameter $\lambda_{EL} = (H(A) - H(A|D))^2/\Omega$.

Results

In this section, we apply the proposed methods to the bladder cancer data of Andrew et al. [2006] to search for interactions between the disease and the genetic variants and smoking

factor. Then, we investigate the robustness of the proposed test statistics by type I error rate calculation, using the joint genotype frequencies of the bladder cancer data. We perform power analysis using the analytical non-centrality parameters of 2-way tests under a few interaction models taken from the literature [Moore et al., 2002; Ritchie et al., 2003a].

Application to Bladder Cancer Data

The bladder cancer data of Andrew et al. [2006] consists of 355 cases and 559 controls. The genotype data of 7 SNPs are available, i.e., three SNPs (APE1 148, XRCC1 399, and XRCC1 194) belong to the BER pathway, one (XRCC3 241) belongs to the DSB pathway, and the remaining three (XPC PAT, XPD 751, and XPD 312) belong to the NER pathway. In addition to the bladder cancer status, the following information about each individual is also available: gender, age, and smoking status given in pack years (e.g., non smoking, < 35 pack years, ≥ 35 pack years).

In the MDR analysis of Andrew et al. [2006], the combination of XPD 751 and XPD 312 was the best two-factor model, which was confirmed by the interaction dendrogram and logistic regression analysis. The three-factor model added Pack-years of smoking to XPD 751 and XPD 312 was the most accurate model, which however was not confirmed by the interaction dendrogram or logistic regression analysis (the logistic regression model failed to converge).

We applied the proposed methods to the bladder cancer data of Andrew et al. [2006], and the results are presented in Table 1. For 2-way interaction, we confirmed the result of Andrew et al. [2006]. The combination of XPD 751 and XPD 312 was the only significant SNP combination detected by our 2-way mutual information gain test statistic (T_{IG} = 51.62, p-value = 6.75e-13), and none of the rest two-factor combinations provided significant result by T_{IG}. By adding each of the remaining 5 SNPs and Pack years, the 3-way total correlation information gain test statistic T_{TCIG} provided a significant result (p-value ≤ 2.67e-9). However, the 3-way interaction information gain test statistic T_{IIG} provided no significant result for any of the three-factor combinations of XPD 751, XPD 312, and one for the remaining 5 SNPs and Pack years (p-value ≥ 0.23).

The only significant result provided by the 3-way interaction information gain test statistic T_{IIG} at 5% significance level came from the combination of XRCC1 399, XRCC1 194, and XRCC3 241 (T_{IIG} = 4.25, p-value = 0.04). However, the result was hardly significant when we adjusted for multiple comparisons, using the Bonferroni procedure for example, which suggested that there is no 3-way interaction combination based on our analysis. The very significant results provided by the 3-way total correlation information gain test statistic T_{TCIG} (Table 1) were most likely due to the 2-way combination of XPD 751 and XPD 312.

Type I Error Rates

Using the joint SNP genotype frequencies of bladder cancer data, we performed simulations to evaluate the type I error rates of 2-way information gain based test statistic T_{IG} and 3-way test statistics T_{IIG} and T_{TCIG}, and the results are presented in Table 2. Each empirical type I error rate in Table 2 was calculated based on 100,000 simulations. That is, we simulated 100,000 random samples of $N = M$ = 100, 150, 200, 250, 300, 400, 500, 600, 700 cases and controls, respectively. These sample sizes were used consistently for all error rate calculations. In each sample, M and N were generated according to the multinomial distributions (M, P) and (N, Q), respectively. Here $P = Q$ are the joint genotype frequencies estimated from the bladder cancer data. For instance, the joint genotype frequencies of the combination of Xpd 751 and Xpd 312 is $P = Q$ = (156, 42, 15, 60, 193, 19, 5, 33, 36)^T/(156 + 42 + 15 + 60 + 193 + 19 + 5 + 33 + 36) = (156, 42, 15, 60, 193, 19, 5, 33, 36)^T/559, which

was used to generate simulation data to calculate the empirical type I error rates of the 2-way test statistic T_{IG}. In Table D.1 of the Supplementary Materials Appendix D, we present the SNP combinations and their joint genotype frequencies used in the calculations of empirical type I error rates of T_{IG}, T_{IIG}, and T_{TCIG}.

We assumed $P = Q$ in our simulation to calculate the type I error rates, i.e., the disease status D is independent of genetic/environmental factors. We calculated an empirical test value for each sample. The empirical type I error rates at nominal levels $\alpha = 0.01$ and $\alpha = 0.05$ are reported in Table 2 and represent the proportions of the test values calculated for the 100,000 samples, that exceed the 99-th and 95-th percentiles of the χ^2_1-distribution. Because the disease status D is independent of genetic or environmental factors, the empirical type I error rates reported in Table 2 can be thought as false positives.

We then calculated 9 empirical type I error rates for each combination of genotype frequencies, i.e., As the combination of SNPs Xpd 751 and Xpd 312 was the only one to provide very significant T_{IG} value (Table 1), we calculated the type I error rate only for this combination. To calculate the empirical type I error rate for T_{TCIG}, we added one of the remaining five SNPs or pack year to Xpd 751 and Xpd 312 and calculated the joint SNP genotype frequencies. This resulted in six combinations of three factors or attributes, i.e., Xpd 751 and Xpd 312 plus one SNP or Pack years. These six combinations provided very significant results of total correlation between the bladder cancer and the three attributes (Table 1). The results of Table 2 show that the empirical type I error rates of 2-way test statistic T_{IG} and 3-way test statistic T_{TCIG} are around the nominal level $\alpha = 0.01$ or $\alpha = 0.05$ when the sample sizes $M = N \geq 300$. Therefore, the test statistics T_{IG} and T_{TCIG} are reasonably conservative and robust. The very significant results of T_{IG} and T_{TCIG} in Table 1 were most likely from the strong interaction between the bladder cancer and the two SNPs Xpd 751 and Xpd 312.

In our simulation to calculate the entries of Table 2, the null hypothesis of T_{IG} was that the disease status D is independent of genetic markers A = Xpd 751 and B = Xpd 312, i.e., $Q_{ij} = P(G_A = i, G_B = j|D = 1) = P(G_A = i, G_B = j) = P_{ij}$, but that the genotypes of SNPs A and B are not independent from each other. The null hypothesis of T_{TCIG} in turn was that $Q_{ije} = P(G_A = i, G_B = j, E = e|D = 1) = P(G_A = i, G_B = j, E = e) = P_{ije}$, i.e., the disease status D is independent of both genetic and environmental factors, but pair-wise and three-way dependences between genetic and environmental factors are allowed. Actually, the genotypes of SNPs Xpd 751 and Xpd 312 are strongly dependent of each other (Pearson $\chi^2 = 256.83$, p-value < 0.00005). In addition, Xpd 751 and Xpd 312 are in strong linkage disequilibrium. Therefore, the simulated data were generated under the null hypothesis of either T_{IG} or T_{TCIG} since the two SNPs, Xpd 751 and Xpd 312, are correlated to each other. The empirical type I error rates of tests T_{IG} and T_{TCIG} reported in Table 2 were around the nominal levels, and the two tests were reasonably robust.

To calculate the empirical type I error rates of the interaction information gain-based test statistic T_{IIG}, we chose the three SNP combinations which were significantly correlated to each other. In the case of three SNPs A, B, and C, significantly correlated means that the four null hypotheses

$$\begin{aligned} H_1&: P(G_A, G_B, G_C) = P(G_A, G_B)P(G_C), \\ H_2&: P(G_A, G_B, G_C) = P(G_A)P(G_B, G_C), \\ H_3&: P(G_A, G_B, G_C) = P(G_B)P(G_A, G_C), \\ H_4&: P(G_A, G_B, G_C) = P(G_A)P(G_B)P(G_C), \end{aligned}$$

are all unlikely to be true. We used the Pearson χ^2 test to choose the three SNP combinations. In Table D.2 of the Supplementary Materials Appendix D, we present the three attribute combinations of theses SNPs. Utilizing the joint genotype frequencies of the three SNP combinations in Table D.2 of the Supplementary Materials, we performed simulations to calculate the empirical type I error rates for the 3-way test statistic T_{IIG}. Since each of the three SNP combinations was selected based on the existence of significant correlations between these SNPs, the simulated data were likely to be generated under the null hypothesis of T_{IIG}, i.e., $I(A, B, C|D) = I(A, B, C)$. The empirical type I error rates of the 3–way test statistic T_{IIG} reported in Table 2 were generally slightly lower than the nominal levels, which suggests that the test T_{IIG} is conservative and robust. The test T_{IIG} was more conservative than the test T_{TCIG} since the the former was constructed to detect the 3-way or higher order interactions and the the latter was constructed to detect the correlations. The existence of 2-way or 3-way interactions implies 3-way correlations, but 3-way correlations are not necessarily due to 3-way interactions.

In Table E.1 of the Supplementary Materials Appendix E, we present the type I error rates of 1-way entropy loss test statistic T_{EL}. The test statistic T_{EL} is reasonably robust and conservative.

Power Comparison

After evaluating the robustness of the test statistic T_{IG} by type I error rate calculation, we performed power calculations for the information gain based test T_{IG} and the naive test T. We were mainly concerned with the performance of the test statistic T_{IG} for nonlinear interactions and in the absence of main effect. To achieve the goal, six models of two-locus penetrance functions and allele frequencies were taken from Moore et al. [2002], Figures 5–10, and the penetrance functions and allele frequencies are presented in Table 3. Similarly, four models were taken from Ritchie et al. [2003a], Figure 2, and the related parameters are presented in Table 4.

To make a comparison, we calculated the theoretical power curves of both test statistics T_{IG} and T based on their non-centrality parameters $\lambda_{T_{IG}}$ and λ_T, and the results are plotted in Figures 2 and 3, respectively. Generally, the power of the information gain-based test T_{IG} was similar to or higher than that of a naive test T. For models 2 and 4–6 in Table 3 and model 1–2 in Table 4, the power curves of T_{IG} were higher than those of T. For the other models, the power was similar. Hence, in terms of power, the information gain-based test T_{IG} performed equally well or better.

By construction, high dimension data are collapsed to build χ^2_1-distributed test T_{IG} which is based on a one-dimension variable. However, the test T is based on genotype frequency comparison of high dimension data and it is χ^2_8-distributed. The information of high dimension data is condensed in T_{IG}. The degrees of freedom of T is 8 and so it is less powerful than T_{IG} which has only 1 degree of freedom. Intuitively, the reduction of degrees of freedom leads to high power for the test statistic T_{IG}.

Discussion

In this paper, we propose information gain based test statistics to detect and to characterize gene-gene and gene-environment interactions of complex diseases. For 2-way interaction, an information gain based approach is proposed using mutual information. The information gain in the presence of disease is defined as a one-dimensional variable through mutual information and entropy function of genetic markers, i.e., $IG(AB \mid D) = I(A, B|D) - I(A, B)$. Based on the one-dimensional information gain, a test statistic T_{IG} is constructed and is

showed to be χ_1^2-distributed. As equation (14) shows, the information gain based test T_{IG} does not involve matrix inverse calculation which facilitates the implementation in practical applications because it is based on the normalization of a one-dimensional random variable $\hat{g} - \hat{f} = \hat{I}(A, B|D) - \hat{I}(A, B)$. However, the calculation of the naive test T involves matrix inverse calculation and it is almost impossible to use it for sparse data as in our simulation calculation of empirical type I error rates. One can calculate the generalized inverse of matrix to implement the naive test T, and then its degrees of freedom changes from dataset to dataset. By power comparison, we clearly showed that the naive test T does not have an advantage over the information gain test T_{IG}.

In Wu et al. [2009], a mutual information based approach was proposed to construct a statistic to test 2-way gene-environment interaction by using a multi-dimensional vector. Under the null hypothesis of independence of the genetic marker and the environmental factor, the test statistic was showed to be a χ_2^2-distributed variable with 2 degrees of freedom [Wu et al., 2009]. Some of the theoretical justification in our discussion such as mutual information is similar to that of Wu et al. [2009]. However, our way to construct the test statistic T_{IG} is different. In addition, our test statistic T_{IG} is χ_1^2-distributed no matter under the null hypothesis of independence of disease status and genetic and environmental factors or under the alternative hypothesis. Under the null hypothesis, T_{IG} is centrally χ_1^2-distributed. Under the alternative hypothesis, the T_{IG} is non-centrally χ_1^2-distributed.

The methods are generalized to test high order K-way interactions and correlations of genetic markers and environmental factors, $K \geq 3$. Two approaches are proposed: (1) an interaction information gain based approach, and (2) a total correlation information gain based approach. Such as the 2-way case, the interaction information gain and total correlation information gain are defined as one-dimensional variables. The related χ_1^2-distributed test statistics T_{IIG} and T_{TCIG} are constructed to test higher order interactions and total correlations, respectively. The test statistic T_{IIG} is based on interaction information gain and it can test K-way interactions, $K \geq 3$. The test statistic T_{TCIG}, however, is based on total correlation information gain and it can test K-way correlations, $K \geq 3$. One may want to notice that correlation can be treated as the interaction in 2-way case, but they are not the same for high order K-way cases, $K \geq 3$.

The power analysis of high order K-way cases, $K \geq 3$, is not carried out in this article. Our problem is that we can not find appropriate models of high order K-way interaction such as those of 2-way cases. It would be interesting to explore some high order K-way interaction models first. Then, it will make more sense to calculate the theoretical power of the high order interaction models and make comparison with the simulated results. To our understanding, the area is still very new and a lot of work still need to be done to understand the high order interactions. The current paper is just a starting point. We will continue our research and report our results to scientific community in the future.

The proposed method was applied to bladder cancer data of Andrew et al. [2006]. We confirmed the significant result of 2-way interaction combination of XPD 751 and XPD 312 in Andrew et al. [2006]. However, we found that there was no significant result of 3-way interaction combinations for the bladder cancer data after adjusting for multiple tests. In the meantime, significant 3-way correlations were found which were basically from the 2-way interaction combination of XPD 751 and XPD 312.

In practice, one can use forward procedure to detect the interactions using test statistics T_{IG} and T_{IIG}. As the first step, one can test 2-way interactions by T_{IG} first. In the presence of 2-way interactions, one can search for evidence of 3-way and higher order interactions by

T_{IIG}. If there are multiple genetic markers, the proposed method can be used to construct genet network to interpret the relation among the markers and environmental factors with the disease. Our analysis of the bladder cancer data provides an example of the procedure. Similarly, one can use test statistics T_{IG} and T_{TCIG} to detect the correlations, but the high order correlations may be actually from low order interactions.

One advantage of the proposed method is that it collapses high-dimensional genetic and environment data into a single dimension, and this makes it possible to build test statistic for high-dimensional sparse data to detect and to characterize gene-gene and gene-environment interactions and correlations. For instance, there are 27 genotype combinations if we consider 3 di-allelic markers. By using 3-way interaction information gain and total correlation information gain of the three markers, we may reduce the 27-dimensional data to be one-dimensional variables to construct the three-way information gain based test statistics. The principle applies to high order K-way interactions and correlations.

To our knowledge, there is no much research about gene-gene and gene-environment interactions using entropy-based approaches, although investigators are paying more and more attention to the research [Dong et al., 2008; Kang et a., 2008; Moore et al., 2006; Wu et al., 2009; Chanda et al., 2007]. It is a new and an interesting area which deserves more attention and investigation. In this article, we make no assumption about population history. It is unclear which kind of impact would appear in the presence of population structure, genotyping error, missing genotypes, phenocopy, and genetic heterogeneity. It would be interesting and important to systematically investigate the issues in the future study. So far, we focus on qualitative trait of complex trait, i.e., either with disease or no disease. It would be interesting to extend the method for quantitative traits. Besides, new methods and models need to be developed to analyze other data type such as sibling and nuclear family [Lou et al., 2008; Martin et al., 2006]. These can be exciting areas for future investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was supported by a Research and Travel Support from the Intergovernmental Personnel Act (IPA), National Cancer Institute, NIH for Fan R., the National Cancer Institute grant R01-CA133996 for Amos C., and NIH grant LM009012 for Moore J. H. We thank Ms. Davnah R. Urbach a lot for helping us in the writings of the paper to remove numerous typographical, grammatical, and bibliographical errors.

References

Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. Carcinogenesis. 2006; 27:1030–1037. [PubMed: 16311243]

Bateson B. William Bateson: A biologist ahead of his time. Am J Hum Genet. 2002; 81:49–58.

Bateson, W. Mendel’s Principles of Heredity. Cambridge: Cambridge University Press; 1909.

Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M. Information-theoretic metrics for visualizing gene environment interactions. Am J Hum Genet. 2007; 81:939–863. [PubMed: 17924337]

Cover, TM.; Thomas, JA. Elements of Information Theory. 2. Wiley-Interscience; 2006.

Dong C, Chu X, Wang Y, Wang Y, Jin L, Shi T, Huang W, Li Y. Exploration of gene-gene interaction effects using entropy-based methods. Eur J of Human Genetics. 2008; 16:229–235. [PubMed: 17971837]

Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. Trans Royal Soc Edinburgh. 1918; 52:399–433.

Frankel WN, Schork NJ. Who's afraid of epistasis. Nature Genetics. 1996; 14:371–373. [PubMed: 8944011]

Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003; 19:376–382. [PubMed: 12584123]

Han TS. Multiple mutual informations and multiple interactions in frequency data. Information and Control. 1980; 46:26–45.

Jakulin, A. PhD thesis. 2005. Machine Learning Based on Attribute Interactions.

Jakulin A, Bratko I. Analyzing attribute interact ions. Lecture Notes in Artificial Intelligence. 2003; 2838:229–240.

Jakulin, A.; Bratko, I. In: Greiner, R.; Schuurmans, D., editors. Testing the significance of attribute interactions; Proceedings of the 21st International Conference on Machine Learning; Banff, Canada. 2004. p. 409-416.

Jakulin A, Bratko I, Smrke D, Demsar J, Zupan B. Attribute interactions in medical data analysis. Lecture Notes in Artificial Intelligence. 2003; 2780:229–238.

Kang G, Yue W, Zhang J, Cui Y, Zuo Y, Zhang D. An entropy-based approach for testing genetic epistasis underlying complex diseases. Journal of Theoretical Biology. 2008; 250:362–374. [PubMed: 17996908]

Lehmann, EL. Theory of Point Estimation. John Wiley & Sons; 1983.

Lou XY, Chen GB, Yan L, Ma JZ, Mangold1 JE, Zhu J, Elston RC, Li MD. A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. Am J Hum Genet. 2008; 83:457–467. [PubMed: 18834969]

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet. 2007; 80:1125–1137. [PubMed: 17503330]

Mahdi H, Fisher BA, Källberg H, Plant D, Malmström V, Rönnelid J, Charles P, Ding B, Alfredsson L, Padyukov L, Symmons DPM, Venables PJ, Klareskog L, Lundberg K. Specific interaction between genotype, smoking and autoimmunity to citrullinated α-enolase in the etiology of rheumatoid arthritis. Nature Genetics. 2009; 41:1319–1324. [PubMed: 19898480]

Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH. A novel method to identify gene-gene effects in nuclear families: The MDR-PDT. Genet Epidemiol. 2006; 30:111–123. [PubMed: 16374833]

McGill WJ. Multivariate information transmission. Psychometrika. 1954; 19:97–116.

Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology. 2006; 241:252–261. [PubMed: 16457852]

Moore, JH.; Hahn, LW.; Ritchie, MD.; Thornton, TA.; White, BC. In: Langdon, WB.; Cantu-Paz, E.; Mathias, K.; Roy, R.; Davis, D.; Poli, R.; Balakrishnan, K.; Honavar, V.; Rudolph, G.; Wegener, J.; Bull, L.; Potter, MA.; Schultz, AC.; Miller, JF.; Burke, E.; Jonoska, N., editors. Applications of genetic algorithms to the discovery of complex models for simulation studies in human genetics; Proceedings of the Genetic and Evolutionary Computation Conference. Morgan Kaufmann; San Francisco. 2002. p. 1150-1155.

Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am J Hum Genet. 2009; 85:309–320. [PubMed: 19733727]

Nothnagel M, Furst R, Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered. 2002; 54:186–198. [PubMed: 12771551]

Ritchie MD, Coffey CS, Moore JH. Genetic programming neural networks as a bioinformatics tool in human genetics. Lect Notes Comput Sci. 2004; 3102:438–448.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001; 69:138–147. [PubMed: 11404819]

Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting genegene interactions in the presence of genotyping error, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003a; 24:150–157. [PubMed: 12548676]

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimization of neural network architecture using genetic programming improves the detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinform. 2003b; 4:28.

Shannon CE. A mathematical theory of communications. The Bell System Technical Journal. 1948; XXVII:379–423. 623–656.

van der Woude D, Alemayehu WD, Verduijn W, de Vries RRP, Houwing-Duistermaat JJ, Huizinga TWJ, Toes REM. Gene-environment interaction influences the reactivity of autoantibodies to citrullinated antigens in rheumatoid arthritis. Nature Genetics. 2010; 42:814–816. [PubMed: 20877316]

Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. A balanced accuracy metric for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol. 2007; 31:306–315. [PubMed: 17323372]

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, Yu Y. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet. 2010; 87:325–340. [PubMed: 20817139]

Watanabe S. Information theoretical analysis of multivariate correlation. IBM J Res Dev. 1960; 4:66–82.

Wu X, Jin L, Xiong MM. Mutual information for testing gene-environmental interaction. PLos One. 2009:e4578. [PubMed: 19238204]

Yeung RW. A new outlook on Shannons information measures. IEEE Transactions on Information Theory. 1991; 37:466–474.

Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nature Genetics. 2007; 39:1167–1173. [PubMed: 17721534]

Appendix A Proof of Relation (13)

A.1 The Subscripts *ij* Do Not Contain 2

Notice

$$\begin{aligned}\frac{\partial f_{00}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{00}\log\frac{P_{00}}{P_{0\cdot}P_{\cdot 0}}\right]=\log\frac{P_{00}}{P_{0\cdot}P_{\cdot 0}}+\left[1-\frac{P_{00}}{P_{0\cdot}}-\frac{P_{00}}{P_{\cdot 0}}\right]\log e,\\ \frac{\partial f_{01}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{01}\log\frac{P_{01}}{P_{0\cdot}P_{\cdot 1}}\right]=-\frac{P_{01}}{P_{0\cdot}}\log e,\\ \frac{\partial f_{10}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{10}\log\frac{P_{10}}{P_{1\cdot}P_{\cdot 0}}\right]=-\frac{P_{10}}{P_{\cdot 0}}\log e,\\ \frac{\partial f_{11}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{11}\log\frac{P_{00}}{P_{1\cdot}P_{\cdot 1}}\right]=0.\end{aligned}$$

In addition, we have

$$\begin{aligned}\frac{\partial f_{02}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{02}\log\frac{P_{02}}{P_{0\cdot}P_{\cdot 2}}\right]=\left[-\frac{P_{02}}{P_{0\cdot}}+\frac{P_{02}}{P_{\cdot 2}}\right]\log e,\\ \frac{\partial f_{12}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{12}\log\frac{P_{12}}{P_{1\cdot}P_{\cdot 2}}\right]=\frac{P_{12}}{P_{\cdot 2}}\log e,\\ \frac{\partial f_{20}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{20}\log\frac{P_{20}}{P_{2\cdot}P_{\cdot 0}}\right]=\left[\frac{P_{20}}{P_{2\cdot}}-\frac{P_{20}}{P_{\cdot 0}}\right]\log e,\\ \frac{\partial f_{21}}{\partial P_{00}}&=\frac{\partial}{\partial P_{00}}\left[P_{21}\log\frac{P_{21}}{P_{2\cdot}P_{\cdot 1}}\right]=\frac{P_{21}}{P_{2\cdot}}\log e.\end{aligned}$$

Moreover, we have

$$\frac{\partial f_{22}}{\partial P_{00}}=\frac{\partial}{\partial P_{00}}\left[P_{22}\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}\right]=-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}+\left[-1+\frac{P_{22}}{P_{2\cdot}}+\frac{P_{22}}{P_{\cdot 2}}\right]\log e.$$

Therefore, we have

$$\frac{\partial f}{\partial P_{00}}=\sum_{i=0}^{2}\sum_{j=0}^{2}\frac{\partial f_{ij}}{\partial P_{00}}=\log\frac{P_{00}}{P_{0\cdot}P_{\cdot 0}}-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}.$$

Similarly, we have for $i, j = 0, 1$

$$\frac{\partial f}{\partial P_{ij}}=\log\frac{P_{ij}}{P_{i\cdot}P_{\cdot j}}-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}.$$

A.2 The Subscripts *ij* Contain One 2

Notice

$$\begin{aligned}
\frac{\partial f_{00}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{00}\log\frac{P_{00}}{P_{0\cdot}P_{\cdot 0}}\right]=-\frac{P_{00}}{P_{\cdot 0}}\log e,\\
\frac{\partial f_{01}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{01}\log\frac{P_{01}}{P_{0\cdot}P_{\cdot 1}}\right]=0,\\
\frac{\partial f_{02}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{02}\log\frac{P_{02}}{P_{0\cdot}P_{\cdot 2}}\right]=\frac{P_{02}}{P_{\cdot 2}}\log e,\\
\frac{\partial f_{10}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{10}\log\frac{P_{10}}{P_{1\cdot}P_{\cdot 0}}\right]=-\frac{P_{10}}{P_{\cdot 0}}\log e,\\
\frac{\partial f_{11}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{11}\log\frac{P_{11}}{P_{1\cdot}P_{\cdot 1}}\right]=0,\\
\frac{\partial f_{12}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{12}\log\frac{P_{12}}{P_{1\cdot}P_{\cdot 2}}\right]=\frac{P_{12}}{P_{\cdot 2}}\log e,\\
\frac{\partial f_{20}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{20}\log\frac{P_{20}}{P_{2\cdot}P_{\cdot 0}}\right]=\log\frac{P_{20}}{P_{2\cdot}P_{\cdot 0}}+\left[1-\frac{P_{20}}{P_{\cdot 0}}\right]\log e,\\
\frac{\partial f_{21}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{21}\log\frac{P_{21}}{P_{2\cdot}P_{\cdot 1}}\right]=0,\\
\frac{\partial f_{22}}{\partial P_{20}}&=\frac{\partial}{\partial P_{20}}\left[P_{22}\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}\right]=-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}+\left[-1+\frac{P_{22}}{P_{\cdot 2}}\right]\log e.
\end{aligned}$$

Therefore, we have

$$\frac{\partial f}{\partial P_{20}}=\log\frac{P_{20}}{P_{2\cdot}P_{\cdot 0}}-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}.$$

Similarly, we have for $i, j = 0, 1$

$$\begin{aligned}
\frac{\partial f}{\partial P_{i2}}&=\log\frac{P_{i2}}{P_{i\cdot}P_{\cdot 2}}-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}},\\
\frac{\partial f}{\partial P_{2j}}&=\log\frac{P_{2j}}{P_{2\cdot}P_{\cdot j}}-\log\frac{P_{22}}{P_{2\cdot}P_{\cdot 2}}.
\end{aligned}$$

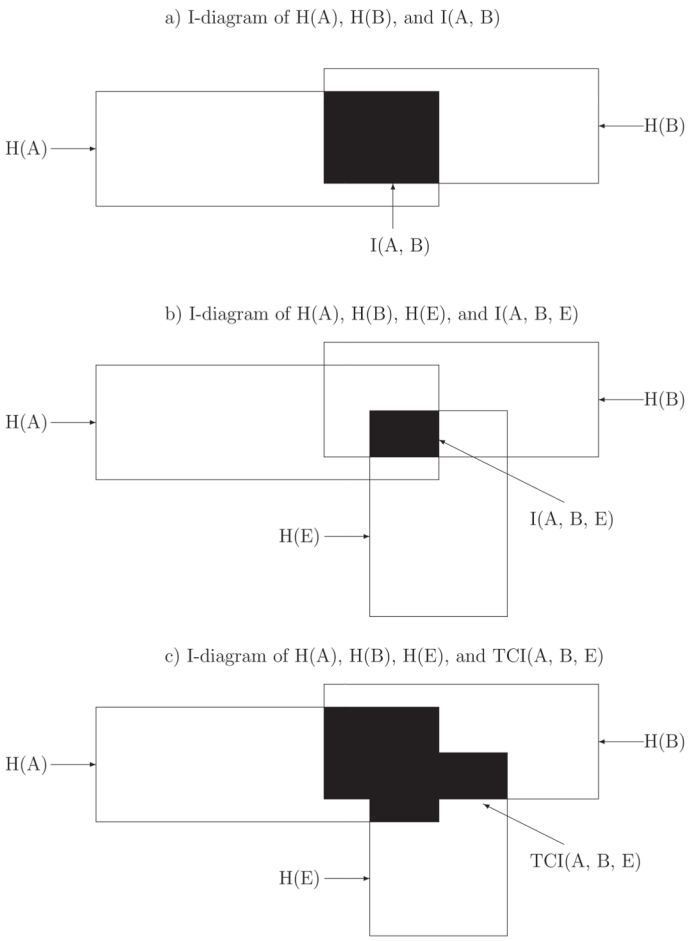

Figure 1.
a) The I-diagram of entropies $H(A)$, $H(B)$, and 2-way mutual information $I(A, B)$; b) The I-diagram of entropies $H(A)$, $H(B)$, $H(E)$, and 3-way interaction information $I(A, B, E)$; c) The I-diagram of entropies $H(A)$, $H(B)$, $H(E)$, and 3-way total correlation information $TCI(A, B, E)$.

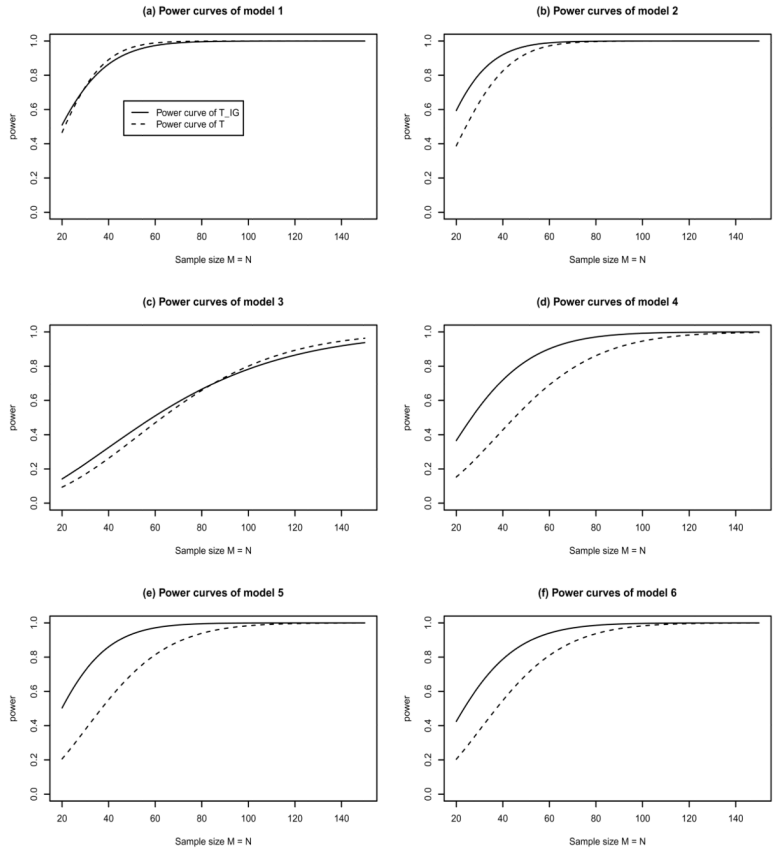

Figure 2.
The power curves of test statistics T_{IG} and T at a significance level α = 0.01 for the six models of Table 3.

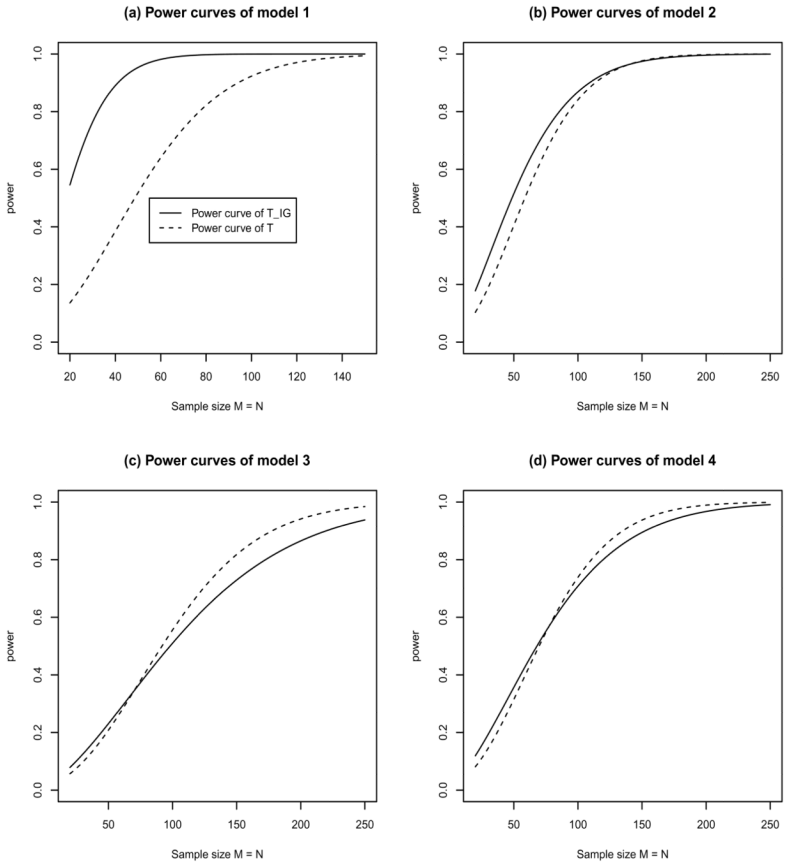

Figure 3.
The power curves of test statistics T_{IG} and T at a significance level α = 0.01 for the four models of Table 4.

Table 1

Results of the 2-way test statistic T_{IG} and the 3-way test statistics T_{IIG} and T_{TCIG} of the bladder cancer data of Andrew et al. [2006].

No. of Factors	SNPs	Test	P-value
2-way interaction	Xpd_751, Xpd_312 *	T_{IG} = 51.62	6.75e-13
	XRCC1_194, XPC PAT #	T_{IG} = 2.47	0.12
3-way interaction and 3-way correlation	Xpd_751, Xpd_312, APE1 148	T_{TCIG} = 44.54	2.49e-11
		T_{IIG} = 0.19	0.66
	Xpd_751, Xpd_312, XPC PAT	T_{TCIG} = 43.87	3.51e-11
		T_{IIG} = 0.12	0.73
	Xpd_751, Xpd_312, Pack_years	T_{TCIG} = 40.93	1.58e-10
		T_{IIG} = 0.11	0.74
	Xpd_751, Xpd_312, XRCC3_241	T_{TCIG} = 40.20	2.29e-10
		T_{IIG} = 0.12	0.73
	Xpd_751, Xpd_312, XRCC1_399	T_{TCIG} = 39.90	2.68e-10
		T_{IIG} = 1.46	0.23
	Xpd_751, Xpd_312, XRCC1_194	T_{TCIG} = 35.41	2.67e-9
		T_{IIG} = 0.30	0.58
	Xpd_751, XRCC1_194, XRCC3_241	T_{TCIG} = 5.67	0.02
		T_{IIG} = 1.80	0.18
	XRCC1_399, XRCC1_194, XRCC3_241	T_{TCIG} = 3.67	0.06
		T_{IIG} = 4.25	0.04 †
	XPC PAT, XRCC1_194, Pack_years	T_{TCIG} = 3.97	0.05
		T_{IIG} = 0.21	0.65

* - the most significant result of T_{IG};

- the second most significant result of T_{IG};

† - the only significant result of 3-way interaction information gain test statistic T_{IIG} at 5% significance level.

Table 2

Type I error rates of 2-way test statistic T_{IG} and 3-way test statistics T_{IIG} and T_{TCIG} based on the joint genotype frequencies of bladder cancer data at nominal levels α = 0.01 and α = 0.05. Each of the entries is based on 100, 000 simulations. Pack_years is used as a SNP in simulating its three category genotypes.

α	Test	SNPs used to generate Joint Genotype Frequency	Sample Sizes $M = N$								
			100	150	200	250	300	400	500	600	700
0.01	T_{IG}	Xpd_751, Xpd_312	0.01499	0.01357	0.01200	0.01184	0.01157	0.01070	0.01094	0.01130	0.01039
	T_{TCIG}	Xpd_751, Xpd_312, APE1 148	0.01701	0.01570	0.01434	0.01382	0.01297	0.01262	0.01159	0.01194	0.01122
		Xpd_751, Xpd_312, XPC PAT	0.01886	0.01568	0.01321	0.01280	0.01208	0.01154	0.01058	0.01051	0.01060
		Xpd_751, Xpd_312, Pack_years	0.01888	0.01572	0.01468	0.01366	0.01235	0.01249	0.01223	0.01112	0.01063
		Xpd_751, Xpd_312, XRCC3_241	0.01636	0.01484	0.01289	0.01244	0.01175	0.01176	0.01099	0.01074	0.01059
		Xpd_751, Xpd_312, XRCC1_399	0.01751	0.01454	0.01340	0.01212	0.01234	0.01153	0.01098	0.01168	0.01058
		Xpd_751, Xpd_312, XRCC1_194	0.01418	0.01259	0.01212	0.01203	0.01031	0.01041	0.01027	0.00971	0.01001
	T_{IIG}	APE1, XRCC1_399, XRCC1_194	0.00948	0.00976	0.00908	0.00857	0.0079	0.00743	0.00752	0.00795	0.00753
		Xpd_751, Xpd_312, XRCC1_194	0.00808	0.00961	0.01057	0.01133	0.01092	0.01053	0.00950	0.00924	0.00843
		XRCC3_241, APE1, XRCC1_194	0.00939	0.00733	0.00648	0.00665	0.00677	0.00663	0.00794	0.00848	0.00889
		XRCC3_241, APE1, XRCC1_399	0.00871	0.00797	0.00784	0.00744	0.00818	0.00735	0.00752	0.00782	0.00786
		XRCC3_241, XRCC1_399, XRCC1_194	0.0119	0.01005	0.00887	0.00848	0.00796	0.00762	0.00852	0.00820	0.00845
0.05	T_{IG}	Xpd_751, Xpd_312	0.06148	0.05813	0.05510	0.05575	0.05316	0.05288	0.05173	0.05097	0.05091
	T_{TCIG}	Xpd_751, Xpd_312, APE1 148	0.06852	0.06355	0.06171	0.05974	0.05776	0.0548	0.05584	0.05373	0.05356
		Xpd_751, Xpd_312, XPC PAT	0.06886	0.06400	0.05987	0.05697	0.05601	0.05478	0.05337	0.05043	0.05220
		Xpd_751, Xpd_312, Pack_years	0.07143	0.06579	0.06289	0.06016	0.05851	0.05541	0.05494	0.05439	0.05311
		Xpd_751, Xpd_312, XRCC3_241	0.06608	0.06200	0.05898	0.05620	0.05451	0.05326	0.05137	0.05281	0.05057
		Xpd_751, Xpd_312, XRCC1_399	0.06594	0.06303	0.05968	0.05701	0.05519	0.05466	0.05297	0.05234	0.05223
		Xpd_751, Xpd_312, XRCC1_194	0.05857	0.05689	0.05453	0.05403	0.05321	0.05059	0.05109	0.04992	0.05136
	T_{IIG}	APE1, XRCC1_399, XRCC1_194	0.0573	0.05309	0.05025	0.0463	0.04481	0.04352	0.04516	0.04511	0.04674
		Xpd_751, Xpd_312, XRCC1_194	0.04551	0.04972	0.05394	0.05424	0.05379	0.05187	0.04912	0.04636	0.04599
		XRCC3_241, APE1, XRCC1_194	0.05195	0.04382	0.04132	0.04125	0.04072	0.04327	0.04446	0.04600	0.04696
		XRCC3_241, APE1, XRCC1_399	0.04834	0.04498	0.04412	0.04328	0.04369	0.04312	0.04446	0.04379	0.04317
		XRCC3_241, XRCC1_399, XRCC1_194	0.06258	0.05403	0.04737	0.04526	0.04456	0.04447	0.04672	0.04736	0.04800

Table 3

Six models of two-locus penetrance functions and allele frequencies taken from Moore et al. [2002], Figures 5–10.

(a) Model 1, $P_A = P_B = 0.5$			
	BB	**Bb**	**bb**
AA	0.083	0.076	0.964
Aa	0.056	0.508	0.085
aa	0.977	0.098	0.062

(b) Model 2, $P_A = P_B = 0.5$			
	BB	**Bb**	**bb**
AA	0.094	0.905	0.097
Aa	0.967	0.097	0.937
aa	0.027	0.990	0.080

(c) Model 3, $P_A = P_B = 0.5$			
	BB	**Bb**	**bb**
AA	0.967	0.314	0.137
Aa	0.313	0.312	0.742
aa	0.129	0.779	0.075

(d) Model 4, $P_A = P_B = 0.5$			
	BB	**Bb**	**bb**
AA	0.967	0.139	0.799
Aa	0.057	0.655	0.627
aa	0.974	0.544	0.019

(e) Model 5, $P_A = P_B = 0.5$			
	BB	**Bb**	**bb**
AA	0.017	0.451	0.711
Aa	0.520	0.571	0.039
aa	0.640	0.053	0.949

(f) Model 6, $P_A = P_B = 0.5$			
	BB	**Bb**	**bb**
AA	0.954	0.256	0.360
Aa	0.010	0.731	0.300
aa	0.801	0.093	0.808

Table 4

Four models of two-locus penetrance functions and allele frequencies taken from Ritchie et al. [2003a], Figure 2.

(a) Model 1, $P_A = P_B = 0.25$			
	BB	**Bb**	**bb**
AA	.08	.07	.05
Aa	.10	0	.10
aa	.03	.10	.04

(b) Model 2, $P_A = P_B = 0.25$			
	BB	**Bb**	**bb**
AA	0	.01	.09
Aa	.04	.01	.08
aa	.07	.09	.03

(c) Model 3, $P_A = P_B = 0.1$			
	BB	**Bb**	**bb**
AA	.07	.05	.02
Aa	.05	.09	.01
aa	.02	.01	.03

(d) Model 4, $P_A = P_B = 0.1$			
	BB	**Bb**	**bb**
AA	.09	.001	.02
Aa	.08	.07	.005
aa	.003	.007	.02