European Journal of Human Genetics (2009) 17, 1274 – 1286
© 2009 Macmillan Publishers Limited All rights reserved 1018-4813/09 $32.00
www.nature.com/ejhg

ARTICLE

The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors

Pritam Chanda[1], Lara Sucheston[2], Aidong Zhang[1] and Murali Ramanathan*[,3]

[1] *Department of Computer Science and Engineering, State University of New York, Buffalo, NY, USA;* [2] *Department of Biostatistics, State University of New York, Buffalo, NY, USA;* [3] *Department of Pharmaceutical Sciences, State University of New York, Buffalo, NY, USA*

We developed an information-theoretic metric called the Interaction Index for prioritizing genetic variations and environmental variables for follow-up in detailed sequencing studies. The Interaction Index was found to be effective for prioritizing the genetic and environmental variables involved in GEI for a diverse range of simulated data sets. The metric was also evaluated for a 103-SNP Crohn's disease dataset and a simulated data set containing 9187 SNPs and multiple covariates that was modeled on a rheumatoid arthritis data set. Our results demonstrate that the Interaction Index algorithm is effective and efficient for prioritizing interacting variables for a diverse range of epidemiologic data sets containing complex combinations of direct effects, multiple GGI and GEI.

European Journal of Human Genetics (2009) **17,** 1274–1286; doi:10.1038/ejhg.2009.38; published online 18 March 2009

Keywords: gene–environment interactions; gene–gene interactions; K-way interaction information

Introduction

With the development, validation and implementation of survey instruments, geographical information systems and approaches to identify genetic variations, such as single-nucleotide polymorphisms (SNPs), deletions, duplications and inversions across the genome, we now have powerful methods in hand to evaluate the role of genes and environment exposures in disease etiology.[1–4] However, the association hits from these genotyping studies may require large follow-on studies to comprehensively sequence the disease-associated regions to enable the discovery of less common genetic variations that may be contributing to disease. Comprehensive follow up studies for characterizing sequence variation in disease-associated regions of the human genome; however, are resource intensive and require large sample sizes. It is therefore essential to leverage the available information from existing genotyping studies to identify the most promising disease-associated regions, the possible environmental factors, the best study design and the appropriate study populations.

In this context, effective analysis tools for detecting gene–gene (GGI) and gene–environmental interactions (GEI) are critical for enabling efficient, well designed follow up sequencing studies. The GGI analysis can highlight important interactions among genetic variations in different regions of the genome and can be used to identify and prioritize regions for sequencing whereas GEI analysis can be employed in study design to ensure that the relevant informative environmental variables are collected.

Prioritizing genetic regions involved in GGI or GEI for sequencing studies can be difficult because the number of interactions, the order of interactions and their magnitudes can vary considerably making it difficult to make decisions regarding the relative importance of, for

*Correspondence: Dr M Ramanathan, Department of Pharmaceutical Sciences, State University of New York, 427 Cooke Hall, Buffalo, NY 14260, USA. Tel: +1 716 645 2842 ext. 242; Fax: +1 716 645 3693; E-mail: murali@buffalo.edu

Received 30 October 2008; revised 22 January 2009; accepted 12 February 2009; published online 18 March 2009

example, a few large magnitude interactions vis-à-vis numerous interactions of moderate magnitude.

We have developed methods for detecting disease-associated genetic variants, environmental variables, GGI and GEI using information theoretic metrics. We demonstrated the utility of the *k*-way interaction information (KWII), which is a multivariate extension of the KLD, for GEI analysis of discrete phenotypes.[5] Subsequently, we enhanced our initial approach by defining a novel metric, Phenotype-Associated Information (PAI) that accounts for the confounding effects dependencies among genetic and environmental variables caused by factors such as linkage disequilibrium.[6] The computational properties of the PAI metric were used as the basis for an efficient search algorithm, AMBIENCE, which identifies variable combinations involved in the strongest interactions.[6] Our methods were found to be remarkably effective for analyzing a diverse range of epidemiologic data sets containing complex combinations of direct effects, multiple GGI and GEI. In this report, our goal is to extend our information-theoretic method to identify the most promising genetic and environmental variables for detailed inspection. We identified an information theoretic metric, the Interaction Index, to effectively visualize and rank the genetic and environmental variables involved in interactions.

Materials and methods

Terminology and representation

Definition of interaction KWII is a parsimonious, multivariate measure of information gain.[7,8] In our information theoretic framework, we use the KWII as the measure of interaction information for each variable combination. In accordance with our earlier report,[6] we operationally define '*A positive KWII value for a variable combination indicates the presence of an interaction, negative values of KWII indicates the presence of redundancy and a KWII value of zero denotes the absence of K-way interactions*'.

k*-way interaction information* For the 3-variable case, the KWII is defined in terms of entropies of the individual variables, *H(A)*, *H(B)* and *H(C)* and the entropies, *H(AB)*, *H(AC)*, *H(BC)* and *H(ABC),* of the combinations of the variables:

$$\begin{aligned}\mathrm{KWII}(A;B;C) = &-H(A)-H(B)-H(C)+H(AB)\\ &+H(AC)+H(BC)-H(ABC)\end{aligned}$$

For the *k*-variable case on the set $v=\{X_1, X_2, \ldots, X_k\}$, the KWII can be written succinctly as an alternating sum over all possible subsets T of v using the difference operator notation of Han:[9]

$$\mathrm{KWII}(v) \equiv -\sum_{T\subseteq v}(-1)^{|v|-|T|}H(T)$$

The KWII represents the gain or loss of information due to the inclusion of additional variables in the model. It quantitates interactions by representing the information that cannot be obtained without observing all *k* variables at the same time.[7,8,10,11] In the bivariate case, the KWII is always positive but in the multivariate case, KWII can be positive or negative. The interpretation of KWII values is intuitive because positive values indicate synergy between variables, negative values indicate redundancy between variables and a value of zero indicates the absence of *k*-way interactions.

Total correlation information For the 3-variable case, the TCI[12] is defined in terms of entropies of the individual variables *H(A)*, *H(B)* and *H(C)* and the entropy of the joint distribution *H(ABC)*:

$$\mathrm{TCI}(A,B,C) = H(A)+H(B)+H(C)-H(ABC)$$

For the *k*-variable case on the set $v=\{X_1, X_2, \ldots, X_k\}$, the TCI, can be expressed as the difference between the entropies of the individual variables $H(X_i)$ and the entropy of the joint distribution $H(X_1X_2 \ldots X_k)$.

$$\mathrm{TCI}(X_1,X_2,\ldots,X_k) = \sum_{i=1}^{k}H(X_i)-H(X_1X_2,\ldots X_k)$$

The TCI is the amount of information shared among the variables in the set; equivalently, it can be viewed as a general measure of dependency. A TCI value that is zero indicates that the variables are independent. The maximal value of TCI occurs when one variable is completely redundant with the others; that is, knowing one variable provides complete knowledge regarding all the others.

Phenotype-associated interaction information

PAI is obtained from the TCI, which represents the overall dependency among the genetic and environmental variables and the phenotype variable by removing the TCI contributions representing the interdependencies among the genetic and environmental variables. The interdependencies among variables can be caused by factors such as LD or by a common source for multiple pollutant exposures. Accordingly, PAI is defined by:

$$\begin{aligned}\mathrm{PAI}(X_1,X_2,\ldots,X_k,P) =&\mathrm{TCI}(X_1,X_2,\ldots,X_k,P)\\ &-\mathrm{TCI}(X_1,X_2,\ldots,X_k)\end{aligned}$$

In the above equation, the genetic and environmental variables are denoted by the X_1, X_2, …, X_K, and the phenotype variable is denoted by *P*. In the PAI definition, the TCI$(X_1, X_2, \ldots, X_K, P)$ term represents the overall dependency among the genetic and environmental variables and the phenotype whereas the TCI$(X_1, X_2, \ldots, X_K)$ term represents the interdependencies among the genetic and environmental variables in the absence of the phenotype variable.

The interaction index

The Interaction Index definition is derived from the interaction contributions (ICs) of each interaction wherein a variable X_i is present. The interaction contribution of a *k*th-order combination *v* involving X_i is denoted by $\mathrm{IC}_v^{(k)}(X_i)$. The order of a combination is the number of genetic or environmental variables in the combination.

Let *v* denote any subset of the genetic and environmental variables $Q=\{X_1, X_2, \ldots, X_n\}$. Let *P* denote the phenotype variable; all combinations in the following definitions include *P*. Let $S^k(X_i,v)$ denote the set of *k*th-order combinations such that each member contains X_i and is a subset of *v*.

The only first-order combination containing one genetic or environmental variable in which X_i participates is $v=\{X_i, P\}$. Therefore, the first-order interaction contribution of X_i, denoted by $\mathrm{IC}^{(1)}_{\{X_i,P\}}(X_i)$, is:

$$\mathrm{IC}^{(1)}_{\{X_i,P\}}(X_i) = \mathrm{PAI}(X_i,P)$$

The interaction contribution of any given combination of two genetic or environmental variables $v=\{X_i, X_j, P\}$ involving X_i, denoted by $\mathrm{IC}^{(2)}_{\{X_i,X_j,P\}}(X_i)$, is:

$$\begin{aligned}\mathrm{IC}^{(2)}_{\{X_i,X_j,P\}}(X_i) =&\mathrm{PAI}(X_i,X_j,P) - \mathrm{PAI}(X_j,P)\\ &-\mathrm{IC}^{(1)}_{\{X_i,P\}}(X_i)\end{aligned}$$

Note that the first-order interaction contribution $\mathrm{IC}^{(1)}_{\{X_i,P\}} \times (X_i)$ is removed in the definition of $\mathrm{IC}^{(2)}_{\{X_i,X_j,P\}}(X_i)$.

Likewise, the interaction contribution of any given combination of three genetic or environmental variables $v=\{X_i, X_j, X_k, P\}$ involving X_i is:

$$\begin{aligned}\mathrm{IC}^{(3)}_{\{X_i,X_j,X_k,P\}}(X_i) =&\mathrm{PAI}(X_i,X_j,X_k,P) - \mathrm{PAI}(X_j,X_k,P) -\\ &\sum_{\omega\in S^{(2)}(X_i,\{X_i,X_j,X_k,P\})} \mathrm{IC}^{(2)}_{\omega}(X_i) - \mathrm{IC}^{(1)}_{\{X_i,P\}}(X_i)\end{aligned}$$

Generalizing, the interaction contribution of any given combination of *n* genetic or environmental variables $v=\{X_i, X_j, X_k, \ldots, X_n, P\}$ involving X_i is defined by:

$$\begin{aligned}\mathrm{IC}^{n}_{\{X_i,X_j,X_k,\ldots,X_n,P\}}(X_i) =&\mathrm{PAI}(X_i,X_j,X_k\ldots X_n,P) - PAI(X_j,X_k\ldots X_n,P) -\\ &\sum_{\omega\in S^{(n-1)}(X_i,\{X_i,X_j,X_k,\ldots,X_n,P\})} \mathrm{IC}^{(n-1)}_{\omega}(X_i) -\\ &\sum_{\omega\in S^{(n-2)}(X_i,\{X_i,X_j,X_k,\ldots,X_{n-1},P\})} \mathrm{IC}^{(n-2)}_{\omega}(X_i) \quad -\ldots - \mathrm{IC}^{(1)}_{\{X_i,P\}}(X_i)\end{aligned}$$

The definitions subtract all lower order interaction contributions for X_i from the difference between *PAI* with X_i and *PAI* without X_i because this difference summarizes all the first order, second order, …., (n−1)th order interaction contributions for X_i.

The Interaction Index*, IID(*X_i*)*, for each variable X_i is defined as the sum of the average interaction contribution (IC) of each evaluated *K*-variable interaction wherein the variable is involved:

$$\mathrm{IID}(X_i) = \mathrm{IC}^{(1)}_{\{X_i,P\}} + \left\langle\left|\mathrm{IC}^{(2)}_{\{X_i,X_j,P\}}\right|\right\rangle + \ldots + \left\langle\left|\mathrm{IC}^{(n)}_{\{X_i,X_j,X_k,\ldots,X_n,P\}}\right|\right\rangle$$

The braces denote averages of the interaction contributions over all the combinations containing X_i of a particular size; the bars represent the absolute values. In the implementation, the average is taken over all sampled combinations.

Based on our KWII-based definition of interactions, a variable was considered to be informative if its Interaction Index value was greater than zero. In simulated data sets, we used replicates to obtain confidence intervals and with real data sets, we used permutations to obtain *P*-values to assess the significance.

Visualization of interaction index

The Interaction Index values of the gene and environmental variables were visualized as stacked bar graphs comprised of the 1-variable containing combinations, 2-variable containing combinations and 3-variable containing combinations.

Simulations for case studies

Simulated data sets were used to critically assess the effectiveness of Interaction Index metric to correctly identify and prioritize the interacting variables. We selected the interaction models for Case Studies 1 and 2 from our earlier paper[5] because it had necessary levels of complexity and also contained nuanced GEI patterns that could provide a challenging test for evaluating the Interaction Index. The model for Case Study 3 was constructed to be more complex and was motivated by genetic, environmental and biomarker variables implicated in congestive heart disease.

A population of 50 000 individuals with randomly varying genotypes and environmental exposures consistent with the underlying GEI models was generated for each of the case studies. The case–control study design was assumed. From the population of 50 000 individual genotypes, a sample of 500 cases and 500 controls was randomly selected. The value 1 was used to represent cases and 0 was used for controls. The SD due to sampling were calculated from 100 independent repetitions of this procedure.

The relative risk was defined as incidence of the disease phenotype in the group exposed to the disease-associated gene–environmental variable combination relative to the incidence in the group without the exposure.[13] We investigated relative risk values of 1.2–2.7 in intervals of 0.3.

Case studies 1A and 1B The underlying GEI model for case studies 1A and 1B is summarized in Figure 1a.

The simulated data for case studies 1A and 1B consisted of four environmental variables, *E1* through *E4*.[5] The environmental variables, *E1* and *E2*, were assumed associated with the disease phenotype whereas *E3* and *E4* were assumed to be uninformative. The environmental variables *E1* and *E3* were assumed to have two states, low exposure (assigned value = *L*) and high exposure (assigned value = *H*) that were treated as categorical variables. The environmental variable *E2* and *E4* were assumed to have 3 states, low exposure (assigned value = *L*), medium exposure (assigned value = *M*) and high exposure (assigned value = *H*) that were also treated as categorical variables. The percentage of subjects in low and high exposure groups of *E1* and *E3* were each 50%; the percentage of subjects in low, intermediate and high exposure groups of *E2* and *E4* were 33.33% each, respectively. The disease was modeled to occur for various combinations of exposure to the environmental variables *E1* and *E2* through interactions with alleles for two SNPs, *SNP 1* and *SNP 2*. The more common and less common (disease) alleles of *SNP 1* and *SNP 2* were assigned allele frequencies of 0.9 and 0.1, respectively. The other SNP variables were *SNP 3* through *SNP 6* were uninformative and had allele frequencies of 0.5. All SNPs were assumed to be diallelic with the three possible genotypes in Hardy–Weinberg equilibrium. A binary phenotype variable, *C*, representing case (assigned value = 1) or control (assigned value = 0) was used.

In both case studies 1A and 1B, the *E1* and *E2* variables were assumed to act independently of each other and the case phenotype value was assigned when combinations of

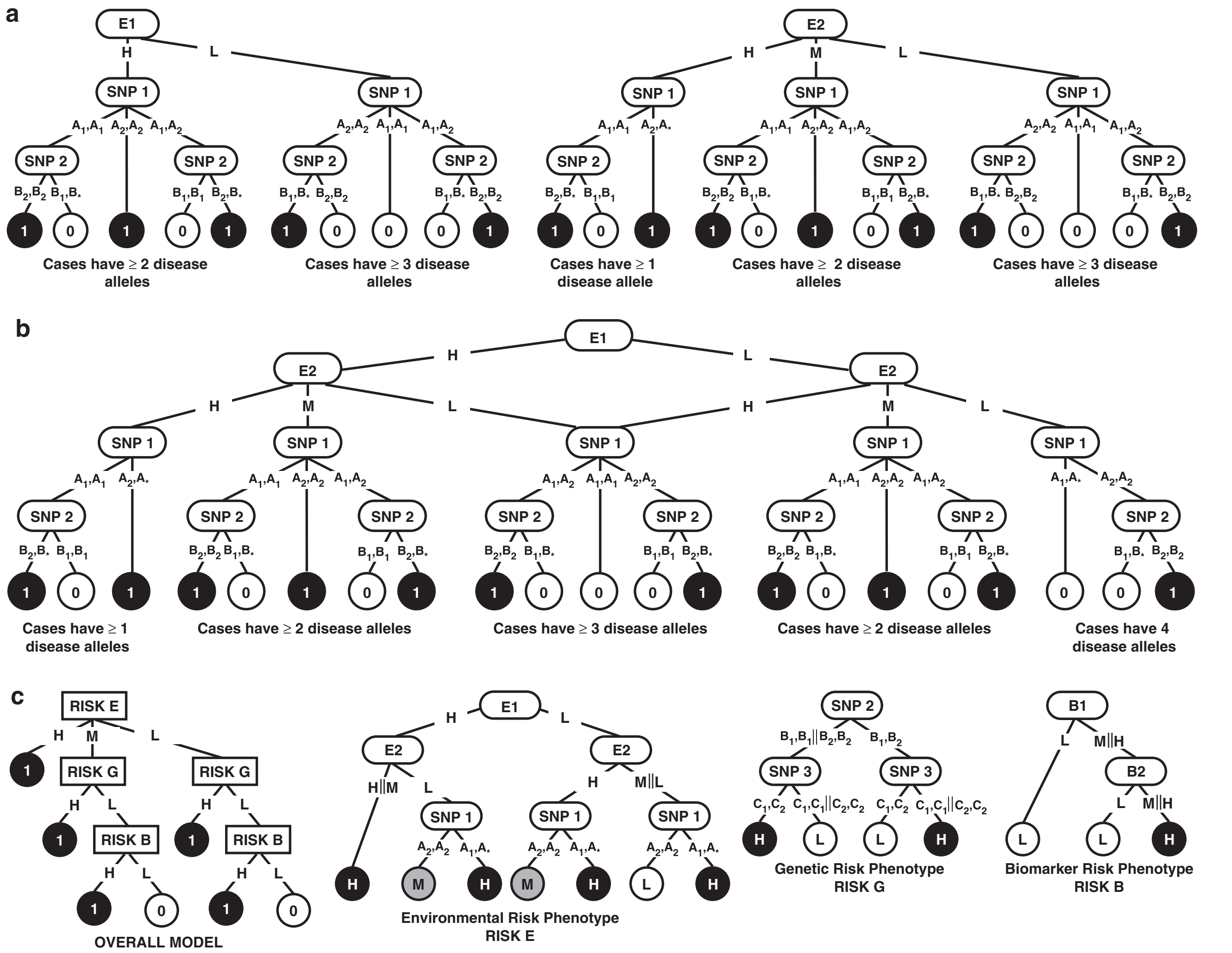
a
E1
H
L
SNP 1
SNP 2
Cases have ≥ 2 disease alleles
Cases have ≥ 3 disease alleles
E2
M
Cases have ≥ 1 disease allele
Cases have ≥ 2 disease alleles
Cases have ≥ 3 disease alleles
b
Cases have ≥ 1 disease alleles
Cases have ≥ 2 disease alleles
Cases have ≥ 3 disease alleles
Cases have ≥ 2 disease alleles
Cases have 4 disease alleles
c
RISK E
RISK G
RISK B
OVERALL MODEL
Environmental Risk Phenotype RISK E
SNP 3
Genetic Risk Phenotype RISK G
B1
B2
Biomarker Risk Phenotype RISK B

Figure 1 (**a**) Shows the interaction model used to generate the data for case study 1A and case study 1B. (**b** and **c**) Shows the interaction model used to generate the data for case study 2 and case study 3, respectively. In (**a** and **b**), the environmental variables *E1* (with states *H*, *L*) and *E2* (with states *H*, *M* and *L*) independently interact with two SNP variables, *SNP 1* (with alleles A_1 and A_2) and *SNP 2* (with alleles B_1 and B_2) to determine the disease status (controls are indicated by 0 and cases are indicated by 1). The asterisk in a genotype represents a 'wild card' indicating that either allele is allowable.

the SNP genotypes and either environmental variable resulted in a case.

The difference between case study 1A and B was that in case study 1B, the SNP variables *SNP 3* and *SNP 4* are assumed to be in linkage disequilibrium with $R^2=0.9$. The SNP variables *SNP 3* and *SNP 4* were assumed to be independent in case study 1A.

Case study 2 This case study differs from case study 1A in that an interaction between environmental variables *E1* and *E2* is incorporated (Figure 1b).[5]

Case study 3 This case study is summarized in Figure 1c and contains a complex combination of environmental, SNP variables and biomarker variables that determine the disease phenotype.

The model for case study 3 consisted of four environmental variables, *E1* through *E4*, four SNP variables, *SNP 1* through *SNP 4* and two biomarker variables *B1* and *B2*. The overall risk of developing the disease phenotype was determined by contributions from three components termed: (i) environmental risk component (Risk E), (ii) the genetic risk component (Risk G) and (iii) the biomarker risk component (Risk B). The Risk E component was assumed to have three states (High *H*, Medium *M*, and Low *L*) whereas the Risk G and Risk B were assumed to have two states (High *H* and Low *L*). The *E1* and *E2* environmental variables interacted with *SNP 1* to determine the environmental risk component (Risk E) of disease risk in Figure 2c. The gene–gene interactions between *SNP 2* and *SNP 3* variables determined the genetic risk component (Risk G) of disease risk whereas interactions between the two biomarkers *B1* and *B2* variables determined Risk B.

The environmental variables, *E1* and *E2*, were disease-associated whereas *E3* and *E4* were assumed to be uninformative. The environmental variables *E1* and *E3* were each assumed to have two states, low exposure (assigned value = *L*) and high exposure (assigned value = *H*); the remaining environmental variables *E2* and *E4* each had an additional state of intermediate exposure (assigned value = *I*). The percentage of subjects in low and high exposure groups of *E1* and *E3* were each 50%; the percentage of subjects in low, intermediate and high exposure groups of *E2* and *E4* were each 33.33%, respectively.

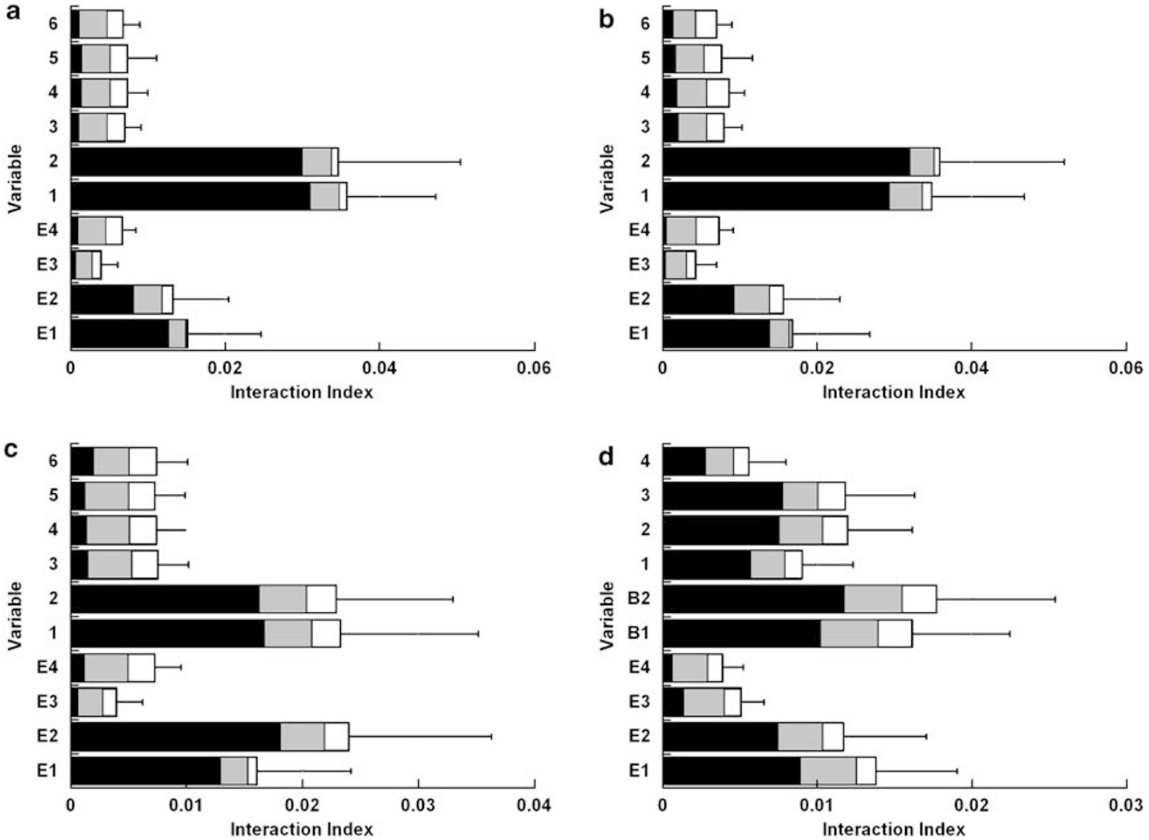

Figure 2 (**a**–**d**) Shows the Interaction Index for case studies 1A, 1B, 2 and 3, respectively, for a relative risk value of 1.8. The stacked bars show the overall interaction index for each SNP or environmental variable; the black regions correspond to the 1-variable contribution, the gray and white regions of the bars correspond to the contributions of combinations of 2-variable and 3-variables, respectively.

Both biomarker variables, *B1* and *B2* were assumed to be associated with the disease phenotype and were each assumed to have three states, low exposure (assigned value = *L*), medium exposure (assigned value = *M*) and high exposure (assigned value = *H*). The percentage of subjects in the low, medium and high exposure groups of *B1* and *B2* were 33.33% each, respectively.

All four SNP variables were assumed to be diallelic with the three possible genotypes in Hardy–Weinberg equilibrium. The more common and less common ('disease') alleles of *SNP 1*, *SNP 2* and *SNP 3* were assigned allele frequencies of 0.9 and 0.1, respectively. The remaining SNP variable *SNP 4* was uninformative and had allele frequencies of 0.5.

A binary phenotype variable, *C*, representing case (assigned value = 1) or control (assigned value = 0) was used. The disease was modeled to occur for various combinations of exposure to the environmental variables *E1* and *E2* through interactions with the biomarker variables *B1* and *B2* and alleles for three SNPs, *SNP 1*, *SNP 2* and *SNP 3*. Variables *E1*, *E2* and *SNP 1* interact to affect the intermediate risk *R1* of the disease.

Prototypical examples of typical environmental variables in congestive heart disease are inflammation and smoking. Biomarkers that are predictive of the risk, congestive heart disease, include factors such as C-reactive peptide and blood cholesterol levels in serum.

Power calculations Power was obtained from 1000 independent repetitions of the simulation procedure for each case study. The calculations were based on a sample size of 500 per group for relative risk values of 1.2 through 2.7 in intervals of 0.3. The distribution of the Interaction Index for a relative risk value of 1 was obtained and its 95th percentile value was computed. Positive values of Interaction Index indicate the presence of significant interactions for the variable and accordingly, power at the relative risk values greater than 1 was defined as the fraction of the simulations whose Interaction Index values exceeded the 95th percentile value of the Interaction Index distribution for the relative risk of 1.

Analysis of public domain data sets

GEI analysis of genetic analysis workshop 15 data The data corresponding to Problem 3 of Genetic Analysis Workshop 15 (GAW15) were obtained from the GAW site (http://www.gaworkshop.org/gaw15data.htm) and used with permission.

These data consist of 100 replicates of simulated data that are modeled after the rheumatoid arthritis (RA) data. Miller *et al*[14] generated the data and the following data description was obtained from the web site: http://genetsim.org/gaw15/answers/. Each replicate includes 1500 nuclear families each with two parents and an affected sib pair and 2000 unrelated controls. The data contains three types of autosomal markers: (i) 730 microsatellite markers with an average spacing of 5 cM; (ii) 9187 SNPs distributed on the genome to mimic a 10 K SNP chip set, and (iii) 17 820 SNPs on chromosome 6. The data include map information, with lists of markers and their locations, and simulated family, marker, and phenotype data. The HLA DR genotype was also available and the phenotype/covariate data included rheumatoid arthritis affection status, age at ascertainment, lifetime smoking, anti-CCP, immunoglobulin M (IgM), severity, age at onset and age at death.

This simulated data set mimics the epidemiology and familial pattern of RA, a complex genetic disease with several loci contributing to disease susceptibility. As summarized in Table 1, the data set models interaction of nine loci: C, DR and D on chromosome 6, A on chromosome 16, B on chromosome 8, E on chromosome 18, F on chromosome 11, G and H on chromosome 9. In addition, sex, age, smoking status, Anti-CCP measure, IgM measure, severity, DR allele from father, DR allele from mother, age at onset, age at death are included as covariates. The biomarkers, anti-cyclic citrullinated peptide antibody (Anti-CCP) and IgM measures are defined for the

Table 1 Effects of major trait loci and covariates in the GAW15 data set

Locus	*Chr*	*SNP no.*	*Phenotype*	*Effects*
DR	6	152–155	RA	Affects risk of RA
A	16	30–31	RA	Controls effect of DR on RA risk
B	8	442	RA	Controls effect of smoking on RA risk
C	6	152–155	RA	Increases RA risk only in women
D	6	161–162	RA	Rare allele increases RA risk 5-fold
E	18	268–269	RA, Anti-CCP	Affects of DR on anti-CCP and increases RA risk
F	11	387–389	IgM	QTL for IgM
G	9	185–186	Severity	25% QTL for severity
H	9	192–193	Severity	25% QTL for severity
Age	—		RA	Affects RA risk through smoking and sex ratio
Sex	—		RA	Affects RA risk locus C
Smoking	—		RA, IgM	Affects RA risk with locus B and through IgM.

cases only. All SNP loci are diallelic and alleles are coded as 1 and 2.

For our analysis, which aimed to evaluate the effectiveness of the Interaction Index, we have used the set of 9187 SNPs along with sex, age and smoking status as covariates. We used the first of the 100 replicates in our analysis. We refer to this data set as the '10K GAW15 Dataset.' The Age, Anti-CCP and IgM variables, which are continuous measures, were discretized by binning into five intervals of equal width. Although haplotype-phase information was provided, we chose to not include it and treat the data as genotype data. We conducted separate analyses with RA affection status, Anti-CCP and IgM as phenotypes of interest. The IgM variable was included as a covariate in the analysis of Anti-CCP as phenotype and vice versa. All the GAW analyses were performed by computing the PAI values for combinations containing up to two variables (excluding the phenotype variable) using AMBIENCE.[6]

GGI analysis of interactions in chromosome 5 We assessed the effectiveness of the Interaction Index metric for identifying key interactions in a genotype data set from Daly *et al*[15] containing 103 SNPs spanning a 616 kb region of chromosome 5q31 that has been linked to Crohn's disease.[16,17] The data set contains genotypes for 129 parent–child trios comprised of 144 cases and 243 controls.[15] For our analysis, subjects and SNPs with missing genotypes were eliminated resulting in 40 SNPs and 150 subjects.

Results

Evaluation of the Interaction Index

We propose the Interaction Index as a PAI-derived measure capable of prioritizing genetic variations for detailed follow up sequencing studies. The Interaction Index is a criterion that summarizes the relative contributions of the variables to the disease associations and we evaluated its ability to rank disease-associated SNPs for case studies 1A, 1B, 2 and 3. Furthermore, we compared the results from: (i) our analysis of the Daly data set[15] to those obtained by Rioux *et al*[17] and (ii) our analysis of the '10K GAW15 Dataset' and compared to the answers provided by Miller *et al*.[14]

Case studies The results from an Interaction Index analysis of case studies 1, 2 and 3 are summarized in Figure 2a–c for a relative risk value of 1.8. Figure 2 summarizes the Interaction Index and its components as a stacked bar graph: the black, gray and white regions indicate the relative contributions of 1-, 2- and 3-variable contributions, respectively. The Interaction Index value of the variables in each case study correctly identifies the disease-associated role of the variable in the underlying interaction model for that case study. Variables *E1*, *E2*, *SNP 1* and *SNP 2* have higher values of Interaction Index than the remaining variables in case study 1A (Figure 2a), case study 1B (Figure 2b) and case study 2 (Figure 2c). For case study 3 (Figure 2d), the variables *E1*, *E2*, *B1*, *B2*, *SNP 1*, *SNP 2* and *SNP 3* have high Interaction Index peaks.

Figure 3 shows the dependence of Interaction Index values of the key causative variables on relative risk for case studies 1A, 2 and 3. The results show that the interaction index increases monotonically with increasing relative risk. At larger values of relative risk, the Interaction Index exhibits a plateau. Likewise, Figure 4 shows the dependence of power of the Interaction Index values on relative risk for case studies 1A, 2 and 3. As expected, the power increases with increasing relative risk. In case study 1A, for a relative risk of 1.5, the power for variables *E1*, *E2*, *SNP 1* and *SNP 2* were 0.77, 0.52, 0.98 and 0.96, respectively. For case study 2, for a relative risk of 1.5, the power for variables *E1*, *E2*, *SNP 1* and *SNP 2* were 0.69, 0.87, 0.88 and 0.81, respectively. For case study 3 for all values of relative risk, *SNP 1* had lower values power than *SNP 2* or *SNP 3*; the biomarker *B1* had lower value of power than *B2*. These differences in power may be because *SNP 1* and *B1* are more distal to the phenotype.

GGI analysis of interactions in chromosome 5 Rioux *et al*[17] found 11 SNPs (IGR2055a_1, IGR2060a_1, IGR2063b_1, IGR2078a_1, IGR2096a_1, IGR2198a_1,I-GR2230a_1, IGR2277a_1, IGR3081a_1, IGR3096a_1, and IGR3236a_1) with alleles that were associated with the risk of Crohn's disease. Nine of 11 significant SNPs were present in the data set we analyzed; SNPs IGR2078a_1 and IGR2277a_1 were missing. From the Interaction Index analysis of the Daly *et al*[15] data set (Figure 5a), all of the nine SNPs identified by Rioux *et al*[17] as significantly associated with Crohn's disease and present in the data set are found to be significant at a significance level of 0.05 (Table 2). Figure 5a demonstrates that SNPs identified by Rioux *et al*[17] (eg, SNPs 34, 20, 30, 32 and so on.) as significant contained strong 1- and 2-variable containing interaction contributions and are identified by our Interaction Index approach. However, there are two SNPs, for example, 28 and 33, involved in interactions that are identified by the Interaction Index method but were not identified by Rioux *et al*[17] These SNPs are more easily prioritized with the Interaction Index because it accounts for higher-order interactions but their relatively low 1-variable associations with phenotype caused these to be missed in the Rioux *et al*[17] analyses.

GEI analysis of genetic analysis workshop 15 data Figure 5b–d present the Interaction Index value for the variables involved in the analysis of '10 K GAW15 Dataset' with RA affection status, Anti-CCP and IgM as phenotypes of interest, respectively. The GAW15 data set contained 100 replicates from repetitions of the simulation procedure that

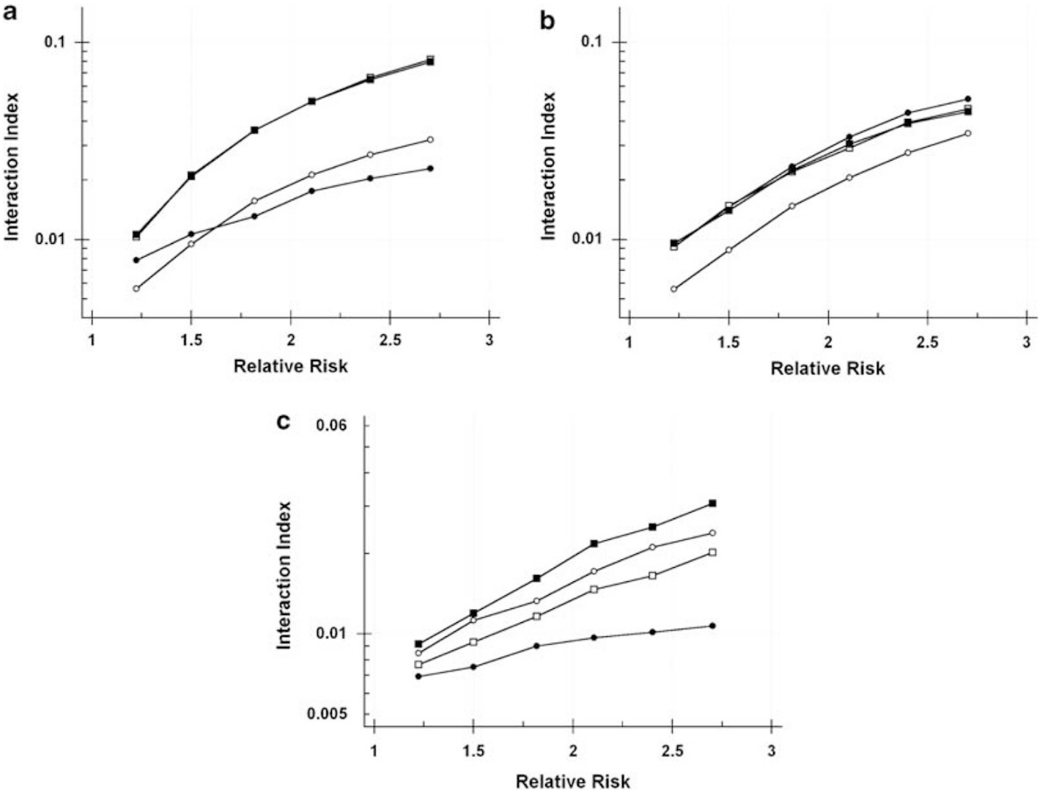

Figure 3 (**a** and **b**) Shows dependence of Interaction Index on relative risk for *E1* (open circles), *E2* (filled circles), *SNP 1* (open squares) and *SNP 2* (filled squares) in case studies 1A and 2, respectively. For case study 3, the *E1* (open circles), *SNP 1* (filled circles), *SNP 2* (open squares) and *B1* (filled squares) are shown in (**c**).

enabled us to compute the 95% confidence interval for the Interaction Index of each variable.

In the RA affection status analysis, the top 10 expected Interaction Index peaks included loci C, DR, F, D, and the covariates, smoking, age, sex (Figure 5b). Loci D and DR affect and increase risk of RA, respectively. Locus F, a quantitative trait locus (QTL) for IgM, is responsible for 30% of the phenotypic variance of IgM. IgM was included in the hazard model used to generate RA affection status. Locus C increases the female risk of RA. The Interaction Index appropriately identifies locus C as showing evidence of a higher-order interaction (the gray portion of the total length of bar) whereas loci D, DR and F show almost entirely 1-variable association with the phenotype. Although most of the sex effect is through locus C, the male-to-female sex ratio in the general population affects RA directly as do age and smoking.

With Anti-CCP as the phenotype, the important roles of loci DR and E were clearly evident as these were the first and second highest Interaction Index values, respectively. Locus E controls the effect of the DR locus on the Anti-CCP phenotype and increases the risk of RA. Again the Interaction Index correctly designates that both loci are involved in higher-order interactions (Figure 5c).

With IgM as the phenotype, the first variable identified by the index was the IgM QTL, Locus F (Figure 5d). Both the RA disease-associated loci and covariates as well as their respective roles (single variable or higher-order interaction) were consistently and accurately elucidated using the Interaction Index.

Computational complexity We assessed computational complexity of Interaction Index calculation using terminology from Corman *et al*.[18] The computation of Interaction Index of a variable of interest involves interaction contributions of combinations containing the variable. Let m be the sample size of the data and n be the number of variables (excluding the phenotype variable) and K be the maximum interaction order of interest. Each *PAI* consumes of the order $O(m^2)$ computations and each interaction contribution of order k contains $O(2^k)$ PAI terms. For $K=4$, the computational complexity of the Interaction Index is $O(m^2)+(n-1)\,O(m^2\,2)+{}^nC_2\,O(m^2\,2^2)+{}^nC_3\,O(m^2\,2^3)$, which is equivalent to $O(m^2\,n^3)$.

Taken together, these results indicate that the Interaction Index is a useful approach for prioritizing genetic regions for detailed sequencing and for identifying the critical environmental variables and covariates for follow

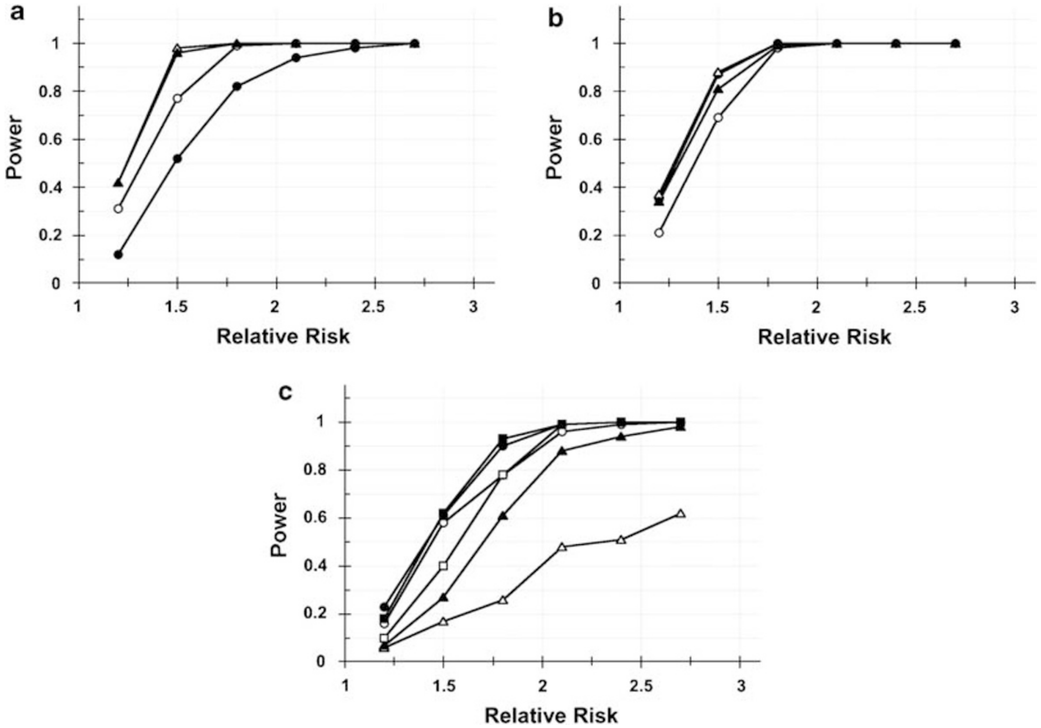

Figure 4 (**a**–**c**) Show the power of the interaction index on relative risk for *E1* (open circles), *E2* (filled circles), *SNP 1* (open triangles) and *SNP 2* (filled triangles) in case studies 1A, 2 and 3 respectively. For case study 3, the *B1* (open squares) and *B2* (filled squares) are additionally shown in (**c**).

Table 2 The *P*-values of the Interaction Index for the SNPs found to be significant from Daly *et al*'s[15] data

SNP name	P-*value*
IGR2096a_1	0.040
IGR2060a_1	0.024
IGR_2055a_1	0.024
IGR_2063b_1	0.020
IGR3029a_2	0.030
IGR3163a_1	0.005
IGR2230a_1	0.008
IGR3096a_1	0.005
IGR3081a_1	0.005
IGR2198a_1	0.002
IGR3236a_1	<0.001

SNPs identified by Rioux *et al*[17] are in bold.

up study design. The Interaction Index is a criterion that summarizes the relative contributions of the variables to the disease associations. Our approach may facilitate decisions regarding the relative importance of interactions.

Discussion

In this report, we developed and evaluated the Interaction Index, a PAI-based information theoretic metric that accounts for the role of genetic variants and environmental variables in GGI and GEI. The Interaction Index can be used to assess the role and contribution of individual SNPs to the disease phenotype.

We developed the Interaction Index as a metric for prioritizing genetic regions for follow up sequencing studies and to target critical environmental variables for acquisition in subsequent study designs. In contrast to the Interaction Index, which is designed to identify SNPs significantly associated with a phenotype, the methods employed by PRIORITIZER software aid in the selection of chromosomal areas for further sequencing. The candidate genes are prioritized using a Bayesian approach by combining information from sources such as Gene Ontology, KEGG, BIND, HPRD, Reactome.[19] However, the approach is more suitable with a candidate gene approach and is not as powerful for genome-wide data because the availability of functional information capable of discriminating individual SNPs may be limited. In such approaches, sequence conservation across species and other functional information can be considered. In the Interaction Index approach, the method utilizes high dimensional information derived from the individual SNP data and exposure profiles within the epidemiological data set to assess GGI and GEI.

To our knowledge, there are no other methods that address the prioritization problem for genetic and environmental

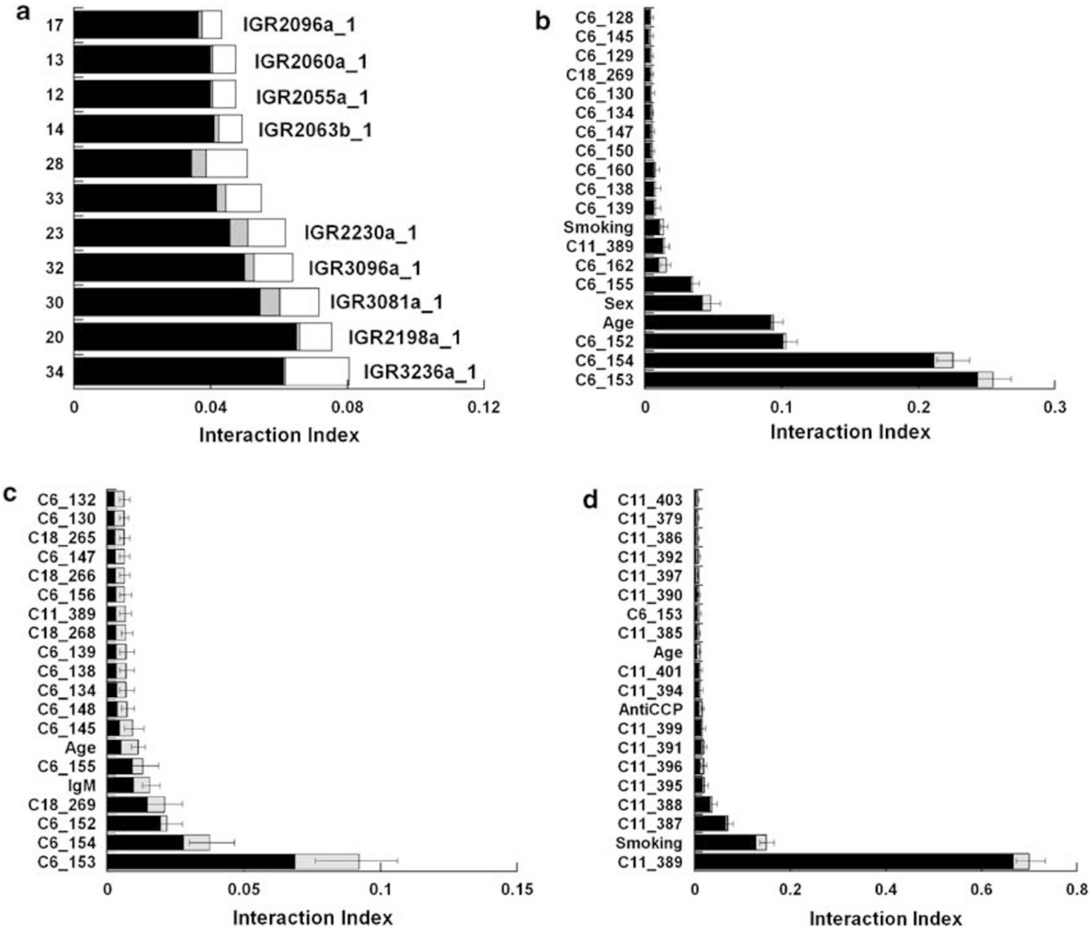

Figure 5 (**a**) Shows the interaction index for the Daly *et al* data set.[15] The nine SNPs present in the data and found to be significantly associated with disease phenotype by Rioux *et al* are highlighted in bold with IGR numbers from Rioux *et al*.[17] The stacked bars show the overall interaction index for each SNP; the black regions correspond to the 1-variable contribution, the gray and white regions of the bars correspond to the contributions of combinations of 2-variable and 3-variables, respectively. (**b**–**d**) Show the interaction index for the '10K GAW15 data set' with RA affection status, Anti-CCP and IgM as phenotypes, respectively. The stacked bars show the overall interaction index for each SNP or covariate; the black regions correspond to the 1-variable contribution, whereas the gray regions of the bars correspond to the contributions of combinations of 2-variable combinations. The error bars in (**b**–**d**) represent the 95th percentile and 5th percentile values obtained from 100 replicates of the data set.

variables based on the participation in GEI. One potential criticism of our approach is the apparent complexity of the underlying equations. Although the mathematical expressions for IC appear complex, their underlying framework is intuitive and utilizes inductive logic. The PAI represents the total phenotype-associated information for a set of variables and represents a general measure of phenotype association wherein the dependencies among variables has been subtracted out. We select only those PAI components of a specific order that contain the variable of interest. The information already obtained from the lower order interactions is removed. The Interaction Index is based on sound information theoretic foundation as it can be shown that the IC of a combination of variables converges to the KWII of the variables when all interactions of a given order or less that contain only the variables of the combination are considered (see Appendix).

However, we have formulated the underlying expressions in terms of the PAI rather than the KWII for reasons of computational efficiency: the PAI is more easily computed because it requires only the individual and joint entropies that are needed for the TCI calculations. KWII computations require the entropies of all subsets and impose computational burden. The PAI also has the additional advantage that the TCI for interdependencies among multiple variables such as those caused by LD are removed. Our previously published results have demonstrated that these PAI methods are effective at accounting for LD and can be used to analyze a diverse range of epidemiologic data sets containing complex combinations of direct effects, multiple GGI and GEI.[5,6]

The variable identification and prioritization problem that the Interaction Index addresses does not provide the best context for comparing information theoretic approaches with other alternatives for identifying genetic interactions. The effectiveness and power of the KWII, however, can be compared to other interaction analysis methods such as those based on probabilistic genetic models, dimensionality reduction or regression. Recall from Methods that positive KWII values indicate synergy between the variables, negative values indicate redundancy between variables and a value of zero indicates the absence of *K*-way interactions.[5,6] This makes the interpretation of the KWII values intuitive because of the qualitative similarity to the interpretation of the coefficient of product terms in logistic, polytomous or continuous regression modeling, wherein a positive product coefficient value identifies interaction terms and implies that synergistic responses are more frequent than antagonist and competitive responses; a negative coefficient value implies that antagonist and competitive responses are more frequent than a synergistic response.[20] We have compared the KWII[5,6] to logistic regression,[21] logic regression,[22] pedigree disequilibrium test,[23] multi-factor dimensionality reduction,[24] restricted partitioning method[25] and others. The power of the KWII for detecting GEI in these experiments was comparable to or better than the competing methods examined.

GEI analysis involves multiple testing, which is associated with increased Type I error and false-discovery rates. Furthermore, the tests in GEI analysis involves high level dependence because of LD among SNPs and because different combinations of genetic and environmental variables can share subsets of variables; for example, combinations {*W*, *X*, *Y*} and {*X*, *Y*, *Z*} both contain the variables *X*, and *Y*. The availability of *P*-values from permutation testing allows users to easily eliminate variables whose Interaction Index values do not meet uncorrected, nominal *P*-value thresholds of for example, $P \leq 0.05$. However, for the remaining variable combinations with lower *P*-values, corrections for multiple testing are also warranted. For multiple testing approaches, such as the method of Obreiter *et al*[26] as implemented in the program SDminP (http://www.dkfz.de/SDMinP/software.html) can be employed. SDminP calculates empirical and adjusted *P*-values for correlated and uncorrelated hypotheses using a Free Step-Down Resampling Method[27] for controlling the familywise error rate (FWER). It utilizes computationally efficient algorithms[28,29] that reduce the re-sampling effort. Other multiple testing options ranging from the conservative Bonferroni correction to the false-discovery rate based Benjamini–Hochberg method[30] can also be used.

At present the Interaction Index metric definition equally weights constituent interaction contributions. Although a weighting function could be implemented at this step and used to assign order or importance to genotypes and/or environmental variables, there are challenges to incorporating external information with the Interaction Index and with other potential methods for GEI analysis. To be useful, the weighting schemes should: (i) minimize extensive need for user input; these can be onerous given the large number of variables in typical GEI-studies. (ii) be generalizable to combinations. It is likely that majority of the external biological data will relate to individual SNPs and environmental variables and information on combinations will be limited. Automated extraction of data from genome databases may be necessary and Boolean rules for extending the information from individual SNP to combinations will need to be devised. (iii) provide interpretable results. To date, the mathematical and statistical properties of weighted entropy measures has not been studied in depth.

While other powerful information theory methods have been proposed for genome-wide data analysis, these metrics were not designed to capture the first and second-order interactions characteristic of complex diseases, but rather test for allelic association with a phenotype.[31–33] Dong *et al*[34] have proposed a method called ESNP2 based on information gain for analyzing two-SNP epistasis in case–control studies and for identifying appropriate two-SNP interaction models. The information-theoretic approach of this report addresses a different problem and is also more generalizable because higher-order interactions are encompassed. Furthermore, we use the PAI to account for LD. The approach is flexible and can be used when the genetic and environmental variables have different numbers of classes or when the phenotype has more than two classes. This means that SNP and microsatellite markers can be analyzed together if necessary. Another critical advantage with our approach is that it provides options for user interactions and visualization. The ability to interact with data enriches the user's experience and can enable detection of features that are otherwise difficult to find.

Our approach characterizes the relative roles of the informative genetic and environmental variables, identifies the subsets of genetic variations and environmental factors involved in the interactions that together could provide a framework for developing explanatory models for the observed patterns of disease associations.

Acknowledgements

This work was partly supported by the NSF and the NIH. Support from the National Multiple Sclerosis Society (RG3743) and fellowship support from Novartis are also gratefully acknowledged.

Conflict of Interest

Use of the information in this article for commercial, non-commercial, research, grant or any purposes other than peer review not permitted prior to publication without expressed written permission of the author.

References

1 Venter JC, Adams MD, Myers EW *et al*: The sequence of the human genome. *Science* 2001; **291**: 1304–1351.

2 Olivier M, Aggarwal A, Allen J *et al*: A high-resolution radiation hybrid map of the human genome draft sequence. *Science* 2001; **291**: 1298–1302.

3 McPherson JD, Marra M, Hillier L *et al*: A physical map of the human genome. *Nature* 2001; **409**: 934–941.

4 Sachidanandam R, Weissman D, Schmidt SC *et al*: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409**: 928–933.

5 Chanda P, Zhang A, Brazeau D *et al*: Information-theoretic metrics for visualizing gene-environment interactions. *Am J Hum Genet* 2007; **81**: 939–963.

6 Chanda P, Sucheston L, Zhang A *et al*: AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* 2008; **180**: 1191–1210.

7 Greiner R, Schuurmans D (eds): Testing the significance of attribute interactions. *Proceedings of the Twenty-first International Conference on Machine Learning (ICML-2004)*. Banff, Canada.

8 Jakulin A: *Machine Learning Based on Attribute Interactions*. Ph.D. thesis, Ljubljana, Slovenia: University of Ljubljana, 2005.

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

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

11 Fano RM: *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: MIT Press, 1961.

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

13 Zhang J, Yu KF: What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998; **280**: 1690–1691.

14 Miller MB, Lind GR, Li N, Jang S-Y: Genetic analysis workshop 15: simulation of a complex genetic model for rheumatoid arthritis in nuclear families including a dense SNP map with linkage disequilibrium between marker loci and trait loci. *BMC Genetics* 2007; **1** (Suppl 1): S4.

15 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.

16 Onnie C, Fisher SA, King K *et al*: Sequence variation, linkage disequilibrium and association with Crohn's disease on chromosome 5q31. *Genes Immun* 2006; **7**: 359–365.

17 Rioux JD, Daly MJ, Silverberg MS *et al*: Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 2001; **29**: 223–228.

18 Corman TH, Leiserson CE, Rivest RL: *Introduction to Algorithms*. Cambridge, MA: MIT Press, 2001.

19 Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006; **78**: 1011–1025.

20 Greenland S: Basic problems in interaction assessment. *Environ Health Perspect* 1993; **101** (Suppl 4): 59–66.

21 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.

22 Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L: Sequence analysis using logic regression. *Genet Epidemiol* 2001; **21** (Suppl 1): S626–S631.

23 Martin ER, Monks SA, Warren LL, Kaplan NL: A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 2000; **67**: 146–154.

24 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.

25 Culverhouse R: The use of the restricted partition method with case-control data. *Hum Hered* 2007; **63**: 93–100.

26 Obreiter M, Fischer C, Chang-Claude J, Beckmann L: SDMinP: a program to control the family wise error rate using step-down minP adjusted *P*-values. *Bioinformatics* 2005; **21**: 3183–3184.

27 Westfall PH, Young SS: *Resampling-based Multiple Testing*. New York, NY: Wiley, 1993.

28 Becker KG, Barnes KC, Bright TJ, Wang SA: The genetic association database. *Nat Genet* 2004; **36**: 431–432.

29 Bhasi K, Zhang L, Zhang A, Ramanathan M: Analysis of pharmacokinetics, pharmacodynamics, and pharmacogenomics data sets using VizStruct, a novel multidimensional visualization technique. *Pharm Res* 2004; **21**: 777–780.

30 Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc (Ser B: Methodol)* 1995; **57**: 289–300.

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

32 Zhao J, Boerwinkle E, Xiong M: An entropy-based genome-wide transmission/disequilibrium test. *Hum Genet* 2007; **121**: 357–367.

33 Li Y, Xiang Y, Deng H, Sun Z: An entropy-based index for fine-scale mapping of disease genes. *J Genet Genomics* 2007; **34**: 661–668.

34 Dong C, Chu X, Wang Y *et al*: Exploration of gene-gene interaction effects using entropy-based methods. *Eur J Hum Genet* 2008; **16**: 229–235.

Appendix

Relationship of interaction contribution to the PAI and KWII

The interaction information involving two variables A, B and phenotype variable P can be written as

$$\begin{aligned}\mathrm{KWII}(A,B,P) &= -\{H(A)+H(B)+H(P)\}+\{H(AB)\\ &\quad +H(AP)+H(BP)\}-H(ABP)=\{H(AB)\\ &\quad +H(P)-H(ABP)\}-\{H(A)+H(P)\\ &\quad -H(AP)\}-\{H(B)+H(P)-H(BP)\}\\ &= \mathrm{PAI}(A,B,P)-\mathrm{KWII}(A,P)-\mathrm{KWII}(B,P)\end{aligned}$$

Thus:

$$\mathrm{PAI}(A,B,P)=\mathrm{KWII}(A,B,P)+\mathrm{KWII}(A,P)+\mathrm{KWII}(B,P)$$

Similarly PAI for three variables A, B, C and phenotype variable P can be expressed as:

$$\begin{aligned}\mathrm{PAI}(A,B,C,P) =&\mathrm{KWII}(A,B,C,P)+\mathrm{KWII}(A,B,P)\\ &+\mathrm{KWII}(A,C,P)+\mathrm{KWII}(B,C,P)\\ &+\mathrm{KWII}(A,P)\mathrm{KWII}(B,P)+\mathrm{KWII}(C,P)\end{aligned}$$

Generalizing to *PAI* for K variables X_1, X_2, …, X_K:

$$\mathrm{PAI}(X_1,X_2,\ldots,X_k,P)=\sum_{\xi\subseteq\{X_1,X_2,\ldots,X_K\},|\xi|\geq 1}\mathrm{KWII}(\xi,P)$$

Now, we show that the IC of a combination of variables converges to the KWII of the variables when all interactions of a given order or less that contain only the variables of the combination are considered. For variable A, when the observed phenotype-associated interactions are {*A, P*}, {*B, P*} and {*A, B, P*}, the interaction contribution *IC*({*A, B, P*}) is given by:

$$\begin{aligned}\mathrm{IC}(\{A,B,P\}) =&\mathrm{PAI}(A,B,P)-\mathrm{PAI}(B,P)-\mathrm{IC}(\{A,P\})\\ =&((\mathrm{KWII}(A,B,P)+\mathrm{KWII}(A,P)\\ &+\mathrm{KWII}(B,P))-\mathrm{KWII}(B,P)-\mathrm{KWII}(A,P)\\ =&\mathrm{KWII}(A,B,P)\end{aligned}$$

Similarly, when the observed phenotype-associated interactions are {*A, P*}, {*B, P*}, {*C, P*}, {*A, B, P*}, {*B, C, P*}, {*A, C, P*} and {*A, B, C, P*}, the interaction contribution *IC*({*A, B, C P*}) is given by:

$$\begin{aligned}\mathrm{IC}(\{A,B,C,P\}) =&\mathrm{PAI}(A,B,C,P)-\mathrm{PAI}(B,C,P)\\ &-(\mathrm{IC}(\{A,B,P\})+\mathrm{IC}(\{A,C,P\})\\ &+\mathrm{IC}(\{A,P\}))=\mathrm{KWII}(A,B,C,P)\\ &+\mathrm{KWII}(A,C,P)+\mathrm{KWII}(B,C,P)\\ &+\mathrm{KWII}(A,B,P)+\mathrm{KWII}(A,P)\\ &+\mathrm{KWII}(B,P)+\mathrm{KWII}(C,P)-\mathrm{KWII}(B,C,P)\\ &-\mathrm{KWII}(B,P)-\mathrm{KWII}(C,P)-\mathrm{KWII}(A,B,P)\\ &-\mathrm{KWII}(A,C,P)-\mathrm{KWII}(A,P)=\mathrm{KWII}(A,B,C,P)\end{aligned}$$

Our results follow as a result of the generalization of this approach.