

## ORIGINAL ARTICLE

**SYMPHONY, an information-theoretic method for gene–gene and gene–environment interaction analysis of disease syndromes**J Knights<sup>1</sup>, J Yang<sup>2</sup>, P Chanda<sup>2</sup>, A Zhang<sup>2</sup> and M Ramanathan<sup>1</sup>

We develop an information-theoretic method for gene–gene (GGI) and gene–environmental interactions (GEI) analysis of syndromes, defined as a phenotype vector comprising multiple quantitative traits (QTs). The K-way interaction information (KWII), an information-theoretic metric, was derived for multivariate normal distributed phenotype vectors. The utility of the method was challenged with three simulated data sets, the Genetic Association Workshop-15 (GAW15) rheumatoid arthritis data set, a high-density lipoprotein (HDL) and atherosclerosis data set from a mouse QT locus study, and the 1000 Genomes data. The dependence of the KWII on effect size, minor allele frequency, linkage disequilibrium, population stratification/admixture, as well as the power and computational time requirements of the novel method was systematically assessed in simulation studies. In these studies, phenotype vectors containing two and three constituent multivariate normally distributed QTs were used and the KWII was found to be effective at detecting GEI associated with the phenotype. High KWII values were observed for variables and variable combinations associated with the syndrome phenotype compared with uninformative variables not associated with the phenotype. The KWII values for the phenotype-associated combinations increased monotonically with increasing effect size values. The KWII also exhibited utility in simulations with non-linear dependence between the constituent QTs. Analysis of the HDL and atherosclerosis data set indicated that the simultaneous analysis of both phenotypes identified interactions not detected in the analysis of the individual traits. The information-theoretic approach may be useful for non-parametric analysis of GGI and GEI of complex syndromes. *Heredity* (2013) **110**, 548–559; doi:10.1038/hdy.2012.123; published online 20 February 2013

**Keywords:** gene–environment interactions; gene–gene interactions; K-way interaction information; syndromes; complex diseases

**INTRODUCTION**

A syndrome is the combination of signs and symptoms that are associated with a morbid process (Dirckx, 2001) and which co-occur frequently enough to indicate a common underlying disease mechanism; a prototypical example is metabolic syndrome, which is estimated to affect nearly 25% of the US population, and greatly increases the risk of cardiovascular disease (Ford *et al.*, 2002). The constellation of metabolic syndrome symptoms includes fasting hyperglycemia, high blood pressure, central obesity, increased triglycerides and decreased high-density lipoprotein (HDL) levels. The causes of metabolic syndrome are not known but aging, genetics, environmental and lifestyle factors such as physical activity and diet are considered to be important contributors. Delineating the gene–gene (GGI) and gene–environmental interactions (GEI) for syndromes may potentially enable identification of the causative pathobiological mechanisms.

There are some unique problems and methodological gaps in the GGI and GEI of syndromes. In particular, syndromes are a distinct class of clinical conditions that may not be appropriately characterized by any one (single) phenotype. Even when the constituent syndrome-affected phenotypes are individually amenable to genetic analysis, such an approach may miss loci involved in regulating the constellation of phenotypes. We hypothesize that an integrated and

unified analysis of the overall syndrome phenotype is preferable for identifying the master pathophysiological mechanism(s). The identification of a unifying pathophysiology for syndromes could lead to better and more targeted treatment strategies as the underlying causes could be addressed instead of separately treating each component. Effective analysis methods also need to address issues arising from linear, and potentially non-linear, dependencies or relationships between the constituent syndrome-affected phenotypes for particular genetic and environmental predictor combinations. Accounting for the confounding effects of dependencies among the genetic and environmental predictor variables caused by factors such as pairwise linkage disequilibrium (LD) and correlations among environmental variables is also a challenge common to GEI and GGI analysis. In this research, we develop a methodology for addressing these unique problems associated with the interaction analysis of syndromes.

Previous reports from our group (Chanda *et al.*, 2008, 2009) have demonstrated the usefulness of the K-way interaction information (KWII) and the phenotype-associated information (PAI) for GGI and GEI analysis of discrete phenotypes and quantitative traits (QTs). This paper extends our methodology and focuses on the development and critical assessment of SYMPHONY, a unique information-theoretic method for GGI and GEI analysis of syndromes.

<sup>1</sup>Department of Pharmaceutical Sciences, State University of New York, Buffalo, NY, USA and <sup>2</sup>Department of Computer Science and Engineering, State University of New York, Buffalo, NY, USA

Correspondence: Professor M Ramanathan, Department of Pharmaceutical Sciences, State University of New York, 427 Cooke Hall, Buffalo, NY 14260, USA.

E-mail murali@buffalo.edu

Received 10 November 2012; revised 12 December 2012; accepted 17 December 2012; published online 20 February 2013

## MATERIALS AND METHODS

### Definitions, terminology and representation

**GGI and GEI.** The methods in this paper are applicable to both GEI and GGI analyses and henceforth, we will simply use the term GEI to refer to both. The term GGI will be used only when environmental variables are not present in a data set.

**Entropy.** The entropy (Shannon, 1997),  $H(X)$ , of a discrete random variable  $X$  is obtained from the probabilities  $p(X=x)$  of each event,  $x$ , in  $X$ , using the formula:

$$H(X) = - \sum_x p(X=x) \log p(X=x)$$

***K*-way interaction information.** The KWII and PAI were developed in our earlier publications (Chanda *et al*, 2007) and are recapitulated here. For the three-variable case, the KWII is defined in terms of entropies for the individual variables,  $H(A)$ ,  $H(B)$  and  $H(P)$  and the joint entropies for the combinations of the variables,  $H(A, B)$ ,  $H(A, P)$ ,  $H(B, P)$  and  $H(A, B, P)$ :

$$KWII(A, B, P) = -H(A) - H(B) - H(P) + H(A, B) + H(A, P) + H(B, P) - H(A, B, P)$$

For the  $K$ -variable case on the set  $v = \{X_1, X_2, \dots, X_K\}$ , the KWII can be written succinctly as an alternating sum over all possible subsets  $T$  of  $\{v, P\}$  using the difference operator notation of Han (1980):

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

The number of variables  $K$  in a combination is called the order of the combination;  $P$  is the phenotype variable. The KWII quantifies interactions by representing the information that cannot be obtained without observing all  $K$  variables at the same time (McGill, 1954; Fano, 1961; Jakulin and Bratko, 2004; Jakulin, 2005). The KWII of a given combination of variables is a parsimonious interaction metric because it does not contain contributions arising from the KWII of other lower order combinations (subsets) of these variables.

In the bivariate case, the KWII is always non-negative but in the multivariate case, the KWII can be positive or negative.

**Interaction definition.** Operationally, we define positive KWII values to indicate interactions (or net synergy) between the variables and negative KWII values to indicate net redundancy between variables. A value of zero indicates the net absence of  $K$ -way interactions.

**Phenotype-associated information.** The PAI is obtained from the total correlation information (TCI) (Watanabe, 1960), which is the difference between the entropies of the individual variables  $H(X_i)$  and the entropy of the joint distribution  $H(X_1 X_2 \dots X_K)$ .

$$TCI(X_1, X_2, \dots, X_K) = \sum_{i=1}^K H(X_i) - H(X_1 X_2, \dots, X_K)$$

The TCI can be viewed as a general measure of dependency.

The PAI represents the overall dependency among the set of genetic and environmental variables with the phenotype by removing the interdependencies among the genetic and environmental variables. The interdependencies among variables can be caused by factors such as LD or by correlated source patterns of pollutant exposures. Accordingly, the PAI is defined by:

$$PAI(X_1, X_2, \dots, X_K, P) = TCI(X_1, X_2, \dots, X_K, P) - TCI(X_1, X_2, \dots, X_K)$$

In the above equation, the genetic and environmental variables (predictors) are denoted by the  $X_1, X_2, \dots, X_K$ , and the quantitative trait is denoted by  $P$ .

In the PAI definition, the  $TCI(X_1, X_2, \dots, X_K, P)$  term represents the overall dependency between the predictors and the phenotype, whereas the  $TCI(X_1, X_2, \dots, X_K)$  term represents the interdependencies among the predictors in the absence of the phenotype.

Our approach utilizes the KWII as the principal measure of the GEI. However, we employ the PAI to facilitate efficient searching of the

combinatorial space (Chanda *et al*, 2008, 2009). KWII computations require the entropies of all subsets and are computationally intensive. PAI calculations are computationally more tractable because only individual and joint entropies are needed. The PAI contains useful information regarding the KWII and also increases monotonically with increased combination size, which makes it a suitable search metric for hill climbing algorithms (Chanda *et al*, 2008, 2009).

### Extension to syndrome multiple QT vector phenotypes

We have previously demonstrated that the general expressions for the KWII and PAI can be used for single QTs and categorical phenotypes (Chanda *et al*, 2008, 2009), as well as count/rate data (Knights and Ramanathan, 2012). However, specific expressions are necessary for the entropy of the multiple QT (MQT) vector phenotype and the subsequent joint entropies of the MQT vector phenotype with discrete variables resulting from the genetic variants, environmental variables and their combinations.

We assume that the MQT vector phenotype,  $\mathbf{P}$ , is multivariate normally distributed. The entropy of a multivariate normal distribution containing  $M$  variables,  $\mathbf{N}_M(\mu, \Sigma)$ , with mean vector  $\mu$  and covariance matrix  $\Sigma$  is (Gokhale *et al*, 1989):

$$H(\mathbf{P}) = \ln(\sqrt{(2\pi e)^M |\Sigma|})$$

For GGI and GEI analysis in this report, we are interested in  $H(X, \mathbf{P})$ , the entropy of the joint distribution of the MQT vector,  $\mathbf{P}$ , and a discrete variable  $X$ , representing, for example, genetic variants or environment variables of interest or their combinations.

$$H(X, \mathbf{P}) = - \sum_x \int_{\mathbf{P}} p(X=x, \mathbf{P}) \ln p(X=x, \mathbf{P}) d\mathbf{P}$$

Therefore:

$$H(X, \mathbf{P}) = - \sum_x \int_{\mathbf{P}} p(\mathbf{P}|X=x) p(X=x) \ln(p(\mathbf{P}|X=x) p(X=x)) d\mathbf{P}$$

We assume that the MQT vector,  $\mathbf{P}$ , given  $X=x$ , is also multivariate normally distributed  $\mathbf{N}_M(\mu_x, \Sigma_x)$ . By expanding, simplifying and substitution, we obtain:

$$H(X, \mathbf{P}) = H(X) + \sum_x p(X=x) \ln(\sqrt{(2\pi e)^M \Sigma_x})$$

The  $H(X)$  term contains only discrete variables and as a result, this entropy can be computed using the usual Shannon entropy formula:

$$H(X) = - \sum_x p(X=x) \ln p(X=x)$$

The  $p(X=x)$  are the probabilities of the event  $x$  (for example, the frequency of a particular genotype). These equations for the entropy of the MQT phenotype, the entropy of discrete variable combinations, and the joint entropy of the combinations composed of the MQT phenotype with discrete variables enable computation of the KWII and the PAI for GGI and GEI analysis of the syndrome phenotype.

### The SYMPHONY algorithm

We designed the SYMPHONY algorithm to mine GEI interactions involving an MQT vector based on our previous work on discrete and single QT. The input of the algorithm is a quadruple  $(\mathbf{X}, \mathbf{P}, \theta, \tau)$ :  $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$  is the set of genetic and environmental variables,  $\mathbf{P}$  is the syndrome MQT vector,  $\theta$  is the number of combinations to be picked up in each iteration and  $\tau$  is the highest order of interactions to be assessed.

The SYMPHONY algorithm employs the PAI iteratively to search for the set of candidate combinations for genetic and environmental variables that have high interactions involving the MQT vector. In the  $i$ th ( $i \leftarrow 1$  to  $\tau$ ) iteration,  $\theta$  combinations which contain  $i$  genetic/environmental variables with the top-ranked PAI values are retained in  $X_i$  and are passed to the  $(i+1)$ th iteration for the PAI computation of combinations with  $i+1$  variables. After  $\tau$  iterations, we can get the promising combinations  $\{X_1, X_2, \dots, X_\tau\}$  which contain up to  $\tau$  variables. Finally, for the combination  $c$ , which is the subset of one of the

combinations in  $X_j$ , calculate  $KWII(c, \mathbf{P})$  to measure interaction for combination  $(c, \mathbf{P})$ .

The pseudo code for SYMPHONY is shown in Supplementary data.

### Methods for the case studies

**Case Study 1, two single-nucleotide polymorphism interactions with a bivariate phenotype.** The model used for this prototypical example of GGI is shown in Figure 1a. The allele frequencies for all bi-allelic single-nucleotide polymorphisms (SNPs) were 0.5.

There were two traits  $P1$  and  $P2$  comprising the MQT vector, whose bivariate normal joint distribution  $N_2(\mu, \Sigma)$  had mean vector  $\mu$  and covariance matrix  $\Sigma$ . The means for the  $P1$  and  $P2$  traits were zero for all genotype combinations and standard deviations of the  $P1$  and  $P2$  traits were both unity for all genotype combinations. However, the genotype combinations differed in the extent to which the  $P1$  and  $P2$  traits were correlated. The Pearson correlation ( $R^2$ ) value for the correlation between the traits  $P1$  and  $P2$  was 0.5 for all combinations containing the *aa* genotype at SNP 1, 0.9 for the  $\{A^*, bb\}$  genotype combinations and 0.1 for all  $\{A^*, B^*\}$  genotype combinations (the asterisk in a genotype combination is a 'wild card' that denotes that any allele is suitable). In the nomenclature of Figure 1a, the covariance matrices are:

$$\Sigma_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$$

A total of 100 independent simulated data sets, each of sample size  $n=2000$ , were used for power calculations. The Pearson correlation ( $R^2$ ) value for the correlation between the traits  $P1$  and  $P2$  was 0.5 for all genotype combinations containing the *aa* genotype at SNP 1,  $(0.5+E)$  for the  $\{A^*, bb\}$  genotype combinations and  $(0.5-E)$  for all  $\{A^*, B^*\}$  genotype combinations. In the nomenclature of Figure 1a, the covariance matrices are

$$\Sigma_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 1 & 0.5+E \\ 0.5+E & 1 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 1 & 0.5-E \\ 0.5-E & 1 \end{pmatrix}$$

The parameter  $E$  was treated as the effect size for this model; the range of  $E$  was  $0 < E < 0.5$ . The distribution of the KWII for an effect size of zero was obtained and its 95th percentile value was computed. Power was defined as the fraction of the simulations whose KWII value exceeded the 95th percentile value of the KWII distribution for the zero effect size.

Permutation-based KWII  $P$ -values were obtained by randomly permuting the phenotype MQT vectors; 10 000 independent permutations of the phenotype were conducted for each combination of interest.

**Case Study 2, two-SNP interactions with a three-variate phenotype.** The model used for this Case Study is shown in Figure 2a. The allele frequencies for all bi-allelic SNPs were 0.5.

The MQT vector phenotype  $\mathbf{P}$  comprising three traits  $P1$ ,  $P2$  and  $P3$ , whose three-variate, multinormal joint distribution was  $N_3(\mu, \Sigma)$  and had mean vector  $\mu$  and covariance matrix  $\Sigma$ . The means for the  $P1$ ,  $P2$  and  $P3$  traits were zero for all genotype combinations, and standard deviations of the  $P1$ ,  $P2$  and  $P3$  traits were each unity for all genotype combinations. However, the genotype combinations differed in the covariance matrices to which the  $P1$ ,  $P2$  and  $P3$  traits were correlated. In the nomenclature of Figure 2a, the covariance matrices are

$$\Sigma_0 = \begin{pmatrix} 1 & 0.5 & 0.9 \\ 0.5 & 1 & 0.5 \\ 0.9 & 0.5 & 1 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 1 & 0.9 & 0.5 \\ 0.9 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.9 \\ 0.5 & 0.9 & 1 \end{pmatrix}$$

As in Case Study 1, a total of 100 independent simulated data sets each of sample size  $n=2000$  were used for power calculations. Power was defined as the fraction of the simulations whose KWII value exceeded the 95th percentile value of the KWII distribution for the zero effect size. The  $P$ -values were calculated using the permutation-based approach in Case Study 1.

**Case Study 3, bivariate phenotype with non-linear dependencies.** The model used for this Case Study is shown in Figure 3a. The allele frequencies for all bi-allelic SNPs were 0.5.

The phenotype  $\mathbf{P}$  comprising two inter-dependent QT,  $P1$  and  $P2$ . The dependence between  $P1$  and  $P2$  was modeled by using a non-linear function that differed between the genotype combinations as shown in Figure 3a. For the genotype combinations containing the *aa* genotype at SNP 1,  $P2$  was independent of  $P1$  and defined by:

$$P2 = N_1(0, 1)$$

For the  $\{A^*, bb\}$  genotype combination, the relationship between  $P1$  and  $P2$  was

$$P2 = P1^2 + P1 + N1(0, 1)$$

For all  $\{A^*, B^*\}$  genotype combinations, the relationship between  $P1$  and  $P2$  was

$$P2 = -P1^2 - P1 + N1(0, 1)$$

For analysis, the relationship between  $P1$  and  $P2$  for all subjects was fitted to a second-order polynomial with linear regression. The linear regression equations were used to calculate the residuals  $P2_R$  between  $P2$  and the regression equation. For KWII calculations, the phenotype  $\mathbf{P}$  comprising  $P1$  and  $P2_R$  was assumed to be a bivariate normal distribution.

The effect size  $E$  was varied by altering the standard deviation of the normal random variate as  $N_1(0, 1/E)$ . The  $P$ -values were calculated using the permutation-based approach in Case Study 1.

### Analysis of public domain data sets

**Hybrid simulations with Genetic Analysis Workshop 15 Genotypes.** The data corresponding to Problem 3 of Genetic Analysis Workshop 15 (GAW15) were obtained from the GAW site (<http://www.gaworkshop.org/>) and used with permission.

These data consist of 100 replicates of simulated data that are modeled after the rheumatoid arthritis data and were generated by Miller *et al.* (2007). Each replicate includes 1500 nuclear families, each with 2 parents and an affected sib pair, and 2000 unrelated controls.

For evaluating the effectiveness of SYMPHONY on a data set with a larger number predictors, we used a subset of data from the larger GAW15 Problem 3 containing 9187 SNPs distributed across the genome to mimic a 10-K SNP chip set. Additionally, sex, age and smoking status were included as covariates; the age variable was binned into three intervals of equal width. The analyses were conducted with unphased genotypes. We pooled sequential pairs of replicates from the GAW15 replicates to generate 50 replicate data sets containing 4000 subjects. We refer to this data set as the '10K GAW15 Data set.'

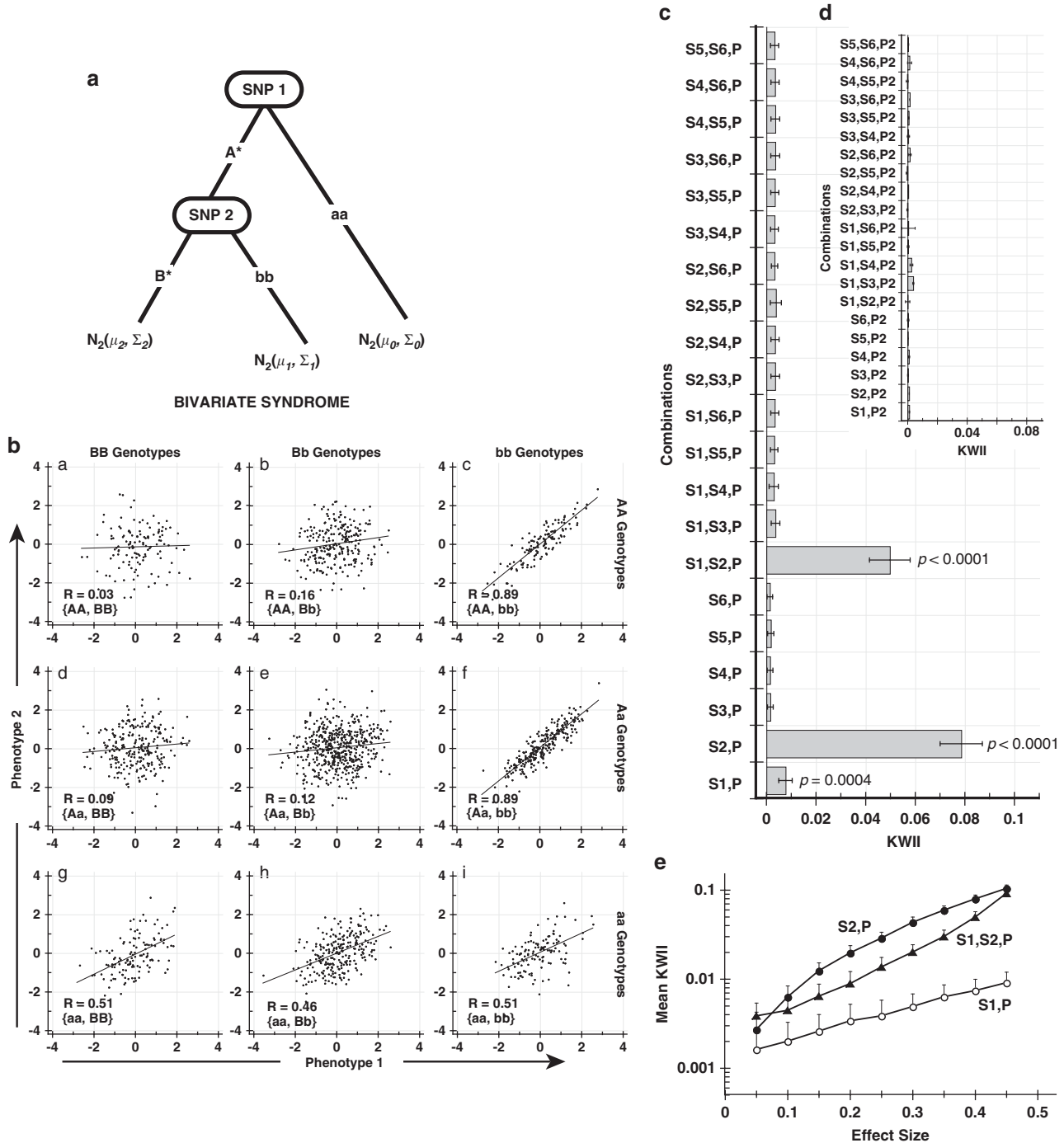
The original GAW15 data set was not simulated to include MQT vector phenotypes. To enable structured evaluation of SYMPHONY, we simulated an MQT vector phenotype random variable against the background of GAW15 genotypes and covariates. The model used (Figure 4a) contained interactions among SNP C6-153 from the DR locus (spanning SNPs C6-152-C6-155) on Chromosome 6 and SNP C16-30 from Locus A (spanning SNPs 30-31 on Chromosome 16), age and smoking. The other SNPs were uninformative, that is, they were not associated with the vector phenotype in the simulations. The covariance matrices for the mixture of bivariate normal distributions arising from the interactions in Figure 4 were

$$\Sigma_0 = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}$$

The  $P$ -value of the KWII of each combination was obtained using 10 000 independent permutations of the phenotype MQT vector for each combination assessed. The permutation  $P$ -values reported are for the first replicate.

We analyzed all 50 replicates individually to obtain KWII values. The mean and standard deviations for the KWII of each combination of variables were calculated from these results.

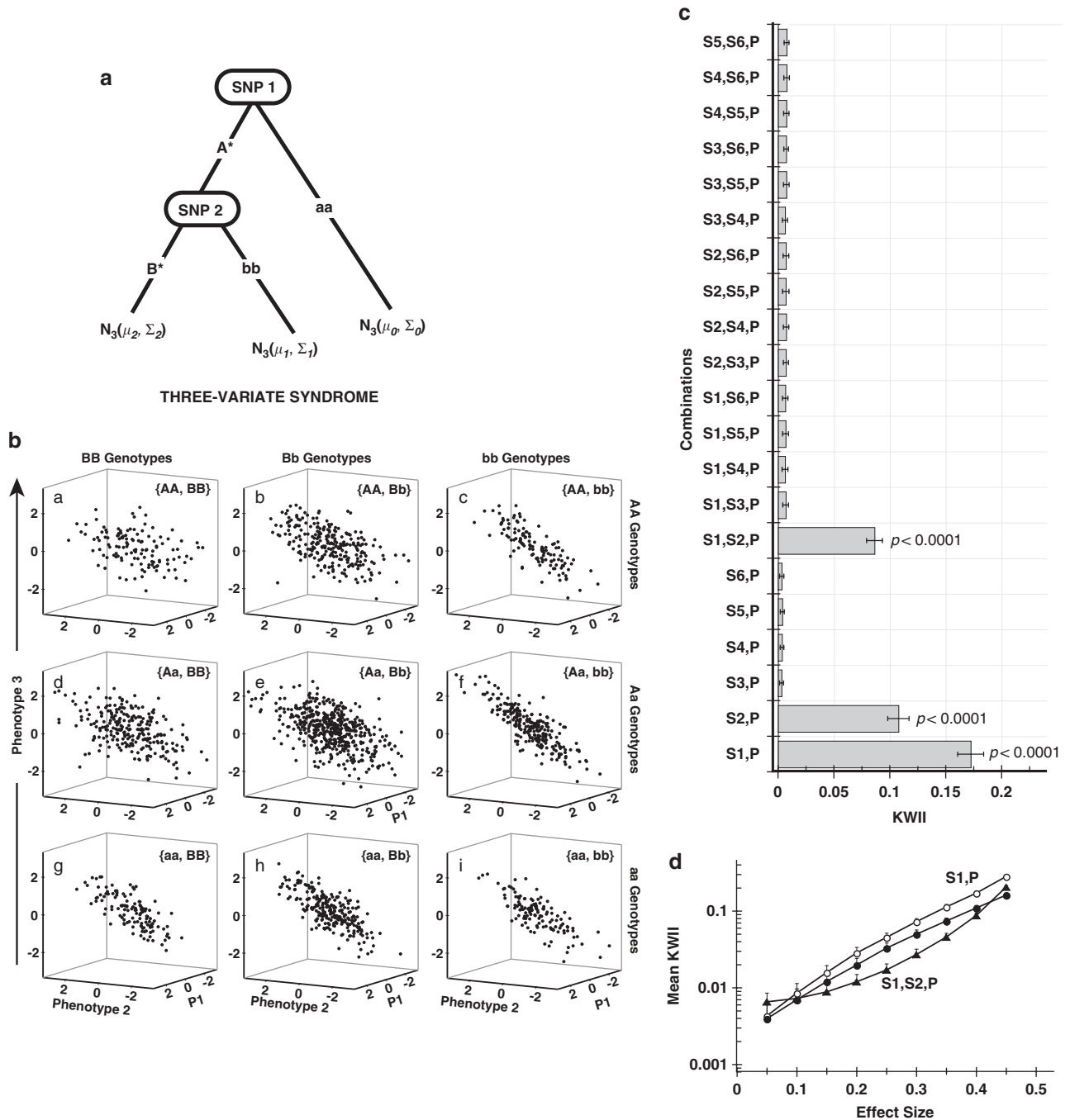
All the analyses were performed with SYMPHONY input parameters values of  $\theta = 50$  and  $\tau = 3$ .



**Figure 1** Case Study 1: two-SNP interactions with a bivariate phenotype. (a) The gene–gene interaction model for Case Study 1. The alleles of SNP 1 are *A* and *a*, whereas those of SNP 2 are *B* and *b*; the asterisk is a ‘wild card’ indicating either allele. For Case Study 1, the Pearson correlation ( $R^2$ ) values for the correlation between the traits  $P_1$  and  $P_2$  was 0.5 for the *aa* genotype at SNP 1, 0.9 for the  $\{A^*, bb\}$  genotypes and 0.1 for all  $\{A^*, B^*\}$  genotypes; the means were all zero and the variances were all unity. (b) The relationship between the syndrome  $P$ , which is comprised of traits  $P_1$  and  $P_2$ , for various genotype combinations of SNP 1 and SNP 2; each plot in the panel represents the genotype combination indicated. The data points correspond to a single subject and the regression lines are shown. (c) The mean KWII values at an effect size of 0.4 for the combinations indicated in the y-axis. The error bars are standard deviations of the KWII. The  $P$ -values from permutation testing are shown for combinations with a  $P$ -value  $\leq 0.05$ . (d) The KWII values of the combinations in (c), but with only phenotype  $P_2$  considered. (e) is a plot of KWII vs effect size for the combinations  $\{1, P\}$ ,  $\{2, P\}$  and  $\{1, 2, P\}$ , which are indicated as open circles, filled circles and triangles, respectively.

The computational speed was assessed in the context of the hybrid GAW15 data set simulations on a 3.2-GHz Intel Xeon computer with Irwindale Processor with 4Gb of main memory. We created six data sets containing 30,

100, 300, 1000, 3000 or the entire 9187 SNPs from the 10-K GAW15 data set. All data sets contained 4000 subjects and the sex, smoking, age, C6-153, C16-30 variables.

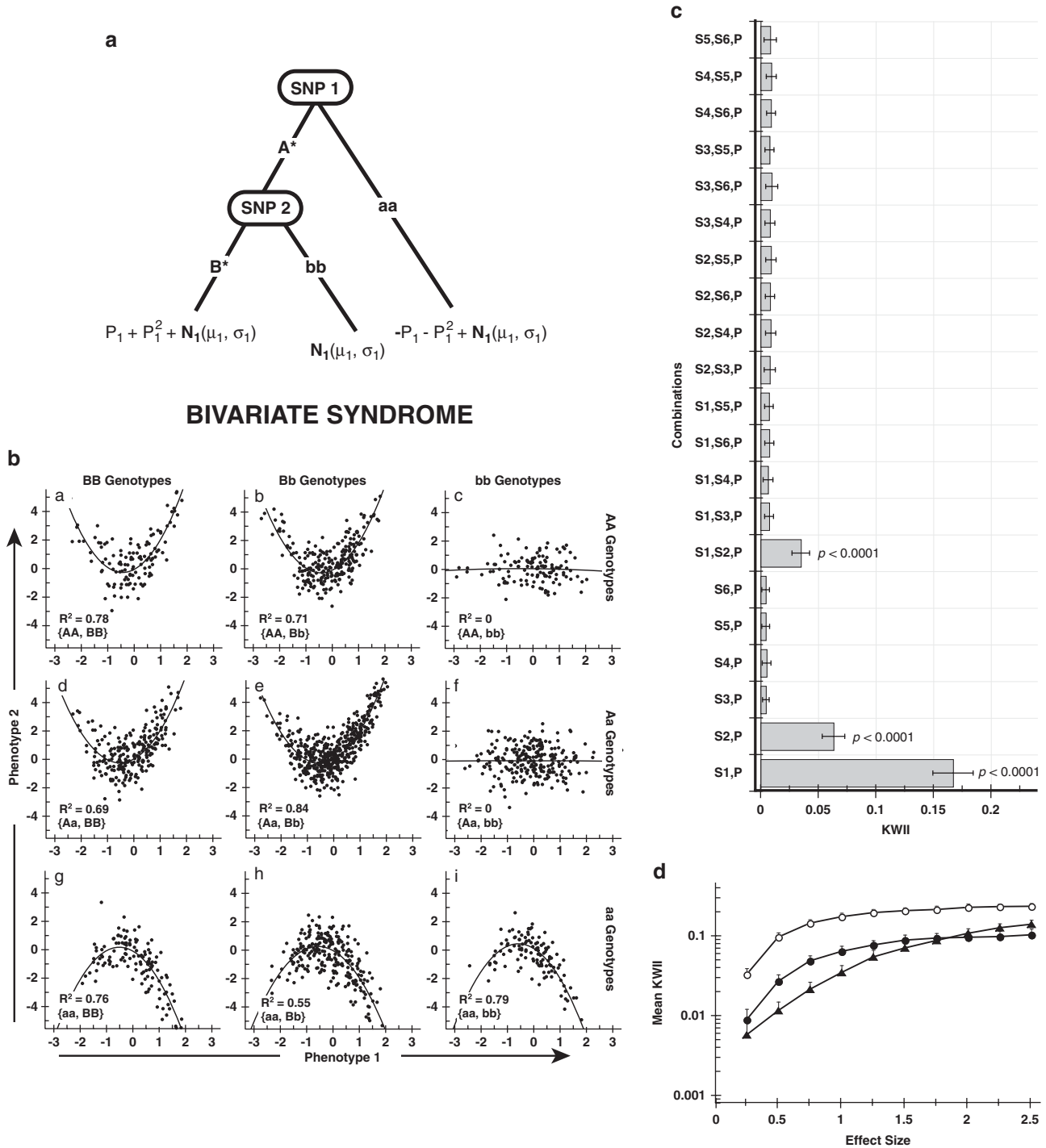


**Figure 2** Case Study 2: two-SNP interactions with a three-variate phenotype. (a) The interaction model used to generate the data for Case Study 2. The model is similar to that in Figure 1 but involves a phenotype vector containing three quantitative trait variables. (b) The data distribution of the phenotype vector for the various genotype combinations involved in the interaction model. (c) The mean KWII values at an effect size of 0.4 for the combinations indicated in the y-axis. The error bars are standard deviations of the KWII. The  $P$ -values from permutation testing are shown for combinations with  $P$ -values  $\leq 0.05$ . (d) is a plot of KWII vs effect size for the combinations {1, P}, {2, P} and {1, 2, P}, which are indicated as open circles, filled circles and triangles, respectively.

**Analysis of HDL and Atherosclerosis Data.** This data set from mice contains genotypes, HDL concentrations and size of aortic fatty streak measurements for 294 female F2 intercross progeny (derived from the C57BL/6 (B6) strain, which is susceptible to atherosclerosis and has low levels of HDL, and the 129 strain, which is relatively resistant to atherosclerosis and has high HDL levels) that were fed a high-fat diet for 14 weeks (Ishimori *et al.*, 2004). The mice were genotyped with 88 simple sequence length polymorphic markers and

subsequently 23 additional simple sequence length polymorphic markers in the QTL regions were added. The data were obtained from the Center for Genome Dynamics at the Jackson Laboratory (<http://cgd.jax.org/nav/qlarchive1.htm>).

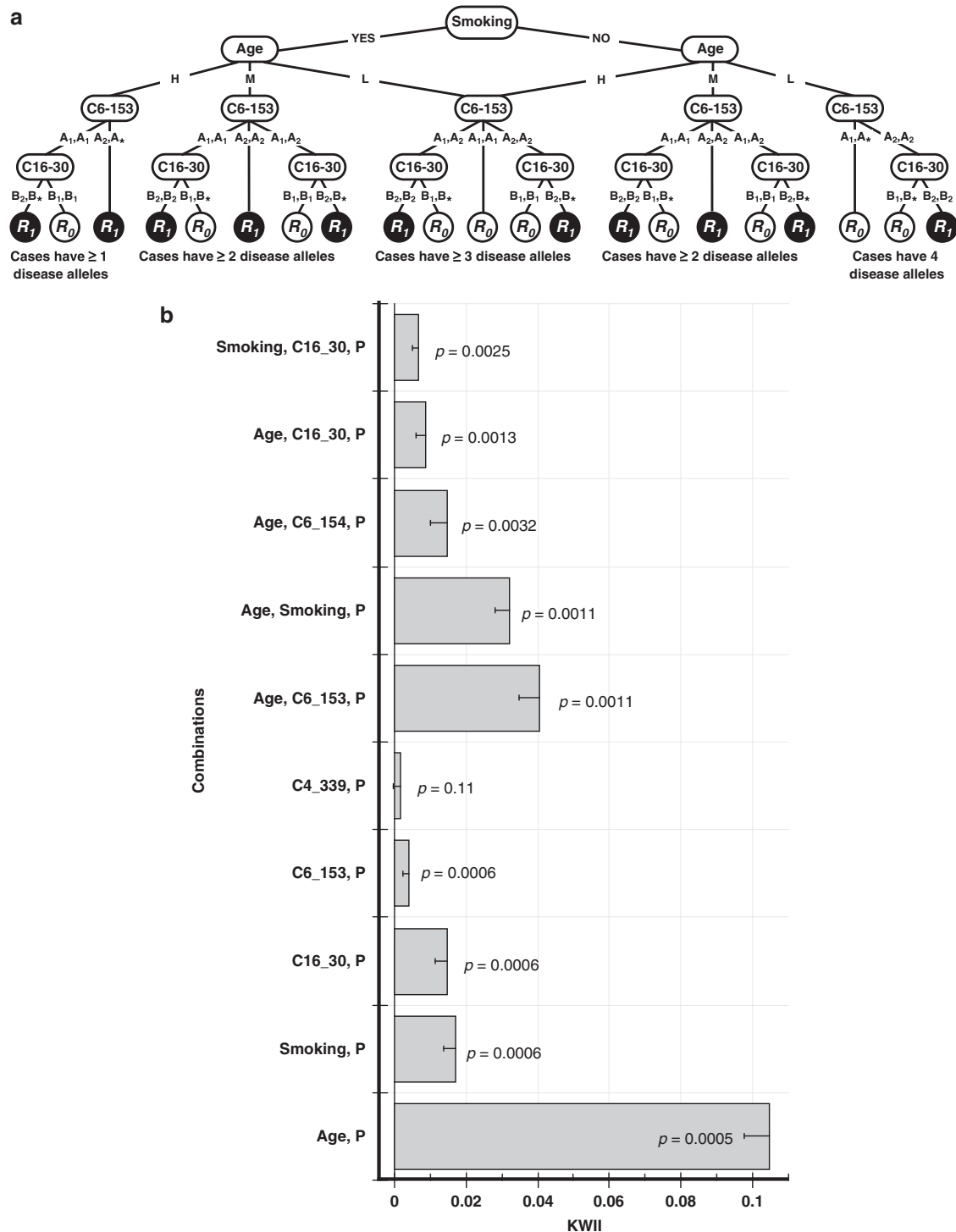
We analyzed the HDL and atherosclerotic aortic fatty streak lesion size as the MQT vector phenotype of interest. The atherosclerotic aortic fatty streak lesion size variable was logarithmically transformed using the function  $\log_{10}(1 +$



**Figure 3** Case Study 3: bivariate phenotype with nonlinear dependencies. (a) The interaction model for Case Study 3. The model is similar to that in Figure 1 except for the phenotype with non-linear dependence between the two quantitative trait variables. (b) The data distribution of the phenotype vector for the various genotype combinations involved in the interaction model; the regression curves for each of the genotypes are also shown. (c) The mean KWII values at an effect size of 1.0 for the combinations indicated in the y-axis. The error bars are standard deviations of the KWII. The *P*-values from permutation testing are shown for combinations with *P*-values  $\leq 0.05$ . (d) is a plot of KWII vs effect size for the combinations {1, P}, {2, P} and {1, 2, P}, which are indicated as open circles, filled circles and triangles, respectively.

Atherosclerotic Aortic Fatty Streak Lesion Size) before analysis; the 1 was added to accommodate individuals with no atherosclerotic streaks within the same transformation framework. The vector phenotype was obtained by normalizing

the means and standard deviations of both the HDL and transformed atherosclerotic aortic fatty streak lesion size variable to zero and 1, respectively.



**Figure 4** Results for hybrid simulations with the GAW15 Problem 3 data set: The model used for simulating the vector phenotypes is shown in (a). The value of the covariance corresponding to the correlation coefficient  $R_1$  was 0.75 and that corresponding to  $R_0$  was 0.25. The mean KWII values for the five one- and two-predictor-containing combinations with the highest KWII values are shown in (b). The permutation  $P$ -values are shown against the bars and the error bars are standard deviations of the KWII.

The  $P$ -value of the KWII of each combination was obtained using 10 000 independent permutations of the MQT vector phenotype for each combination assessed.

The results were compared with the findings on the same data set using regression presented in Tables 1–3 of Ishimori *et al.* (2004) and to the information-theoretic findings in Chanda *et al.* (2009).

*Analysis of 1000 genomes data.* Data from the 1000 Genomes project was downloaded from <http://www.1000genomes.org/data>. PLINK was used to compute minor allele frequency (MAF) and pairwise LD values.

*Minor allele frequency.* To investigate the effect of MAF on the KWII, we used the Southern Han Chinese (CHS) data set. Each autosome was divided into

five segments of equal length and 20 contiguous SNPs were randomly selected from each segment. A data set containing 2200 SNPs and 4000 subjects was created by independently resampling each of the 22 autosomes. The Model in Case Study 1 was used to assign bivariate vector phenotypes. The informative SNP 1 in this Case Study was randomly selected from among the SNPs with MAF of  $0.49 \pm 0.01$ . The informative SNP 2 in Case Study 1 was among the SNPs with MAF of  $0.49 \pm 0.01$ ,  $0.33 \pm 0.01$ ,  $0.25 \pm 0.01$  and  $0.1 \pm 0.01$ .

**Linkage disequilibrium.** The dependence of KWII on LD was evaluated using the same general method for MAF evaluation; however, both informative SNPs (SNP 1, SNP 2) had MAFs between 0.48 and 0.5 and were randomly selected from among the SNP pairs in LD within the following ranges of  $R^2$ : 0–0.3, 0.35–0.4, 0.45–0.5, 0.65–0.8 and 0.85–1.

**Admixture/population stratification.** To assess the effect of race admixture/population stratification on the KWII, the 1000 Genomes data from the Southern Han Chinese (CHS), Finnish (FIN) and Toscani (TSI) groups were filtered to obtain a single data set containing genotypes for all the SNPs that were shared between the three groups. From each autosome, 100 SNPs were randomly selected as described for the MAF and LD simulations. The informative SNP 1 and SNP 2 in Case Study 1 were selected from among the SNPs with MAFs between 0.48 and 0.5.

The Model in Case Study 1 was used to randomly assign bivariate vector phenotypes. The covariance matrices for different genotype combinations of the three groups, however, were different:

Genotype	{aa, **}	{A*, bb}	{A*, B*}
CHS	$\Sigma_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\Sigma_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$
FIN	$\Sigma_0 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$	$\Sigma_1 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$
TSI	$\Sigma_0 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$	$\Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\Sigma_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$

The covariance matrix of the bivariate phenotype for the FIN group was independent of the genotype. Saliiently, the covariance matrices of the CHS and TSI groups for the {A\*, bb} and the {aa, \*\*} genotypes differed.

Data sets containing 4200 subjects were obtained by resampling the shared genotypes. A race variable was included for each subject based on the group (CHS, FIN or TSI) from which the resampled genotypes were obtained. To assess the effects of admixture, the percent of the FIN group in the population was varied (0, 10, 25, 33, 50 and 80%) and the relative percentages of subjects from the CHS and TSI groups were kept equal. The effects of admixture were assessed through the dependence of the KWII on the percentage of the FIN group in the population.

## RESULTS

### Analysis of case studies

**Case Study 1.** To assess the usefulness of our information-theoretic approach, we used the GGI model in Figure 1a to generate simulated data. The interaction model was inspired by the genetics of coat color in Labrador retrievers, which involves interactions between two loci involved in skin pigmentation. The MQT phenotype vector consisted of two QTs.

To exclude the possibility that the information-theoretic approach was detecting interactions by identifying associations between the individual QT in the phenotype vector, we ensured that there were no differences in the mean values and standard deviations among the genotypes on either of the two QTs: the mean values and standard deviations for both QT were zero and unity for each of the nine genotypes. The strength of the associations between the QT, however, was dependent on the genotype. These relationships are summarized in Figure 1b.

Figure 1c shows the KWII values for each combination. We found three significant peaks {1, P}, {2, P} and {1, 2, P}. From permutation analysis, the *P*-value of the KWII for the {1, P} combination was 0.004

and the *P*-values for the KWII of the {2, P} and {1, 2, P} combinations were both  $<0.0001$ . The *P*-values for the KWII of all other combinations were  $>0.05$ . Figure 1d shows the KWII values for the same data when the quantitative trait *P*2 was analyzed individually without considering *P*1: there were no prominent peaks. The results for individual analysis of *P*1 without considering *P*2 were similar and are not shown.

Figure 1d shows the dependence of the KWII values of the {1, P}, {2, P} and {1, 2, P} combinations when the effect size (defined as the difference in the covariance between *P*1 and *P*2) between the three outcomes in Figure 1a was varied. The logarithm of KWII values for the {1, P}, {2, P} and {1, 2, P} combinations increased monotonically with increasing effect size.

The power of the KWII for detecting the interactions for the {1, P}, {2, P} and {1, 2, P} combinations was computed at a sample size of  $n = 2000$  subjects and effect sizes of 0.2 and 0.4. At the effect size of 0.2, the power to detect the {1, P}, {2, P} and {1, 2, P} was 32, 100 and 88%, respectively. At the higher effect size of 0.4, the power to detect the {1, P}, {2, P} and {1, 2, P} combinations was 96, 100 and 100%, respectively.

The results from this motivating example demonstrate that the information-theoretic approach is capable of detecting GGI associated with vector phenotypes representing syndromes. High KWII values were observed for SNPs involved in interactions compared with uninformative SNPs, which had low KWII values; the value of the KWII was also dependent on the effect size.

### Case Study 2

Syndromes such as metabolic syndrome can contain more than two QTs. We therefore evaluated the ability of our method to identify interacting SNPs in a simulation framework similar to Case Study 1 but with an MQT phenotype comprising three QT. Figure 2a shows model used to generate the simulated data.

Figure 2b is a panel of three-dimensional scatter plots summarizing the generated data and shows that the mean values and standard deviations for the QT were identical for all the genotype combinations of SNP 1 and SNP 2. The KWII method therefore relies on the dependencies among the constituent phenotypes in the syndrome MQT vector to detect the underlying GGI.

Figure 2c shows the KWII values, their confidence intervals and *P*-values for various combinations in the model. As would be expected from the simulation model, the KWII values for the informative combinations {1, P}, {2, P} and {1, 2, P} had the highest peaks (all *P*-values  $\leq 0.001$ ). Figure 2d shows the dependence of the KWII values for the {1, P}, {2, P} and {1, 2, P} combinations when the effect size was varied. As for Case Study 1, the logarithm of KWII values for the {1, P}, {2, P} and {1, 2, P} combinations increased monotonically with increasing effect size.

The power of the KWII for detecting the interactions in the {1, P}, {2, P} and {1, 2, P} combinations was computed at a sample size of  $n = 2000$  subjects and effect sizes of 0.2 and 0.4. For both effect sizes, the power to detect the {1, P} and {2, P} combinations was 100%. The power to detect the {1, 2, P} combination was 100% at the effect size of 0.4 and 62% at the effect size of 0.2.

Thus, the KWII method is capable of detecting complex dependencies in three-variate phenotypes in the absence of main effects on the constituent QTs.

**Case Study 3.** A complexity with syndrome MQT vector data that is not present in discrete phenotypes or in the analysis of a single QT is the possibility of a non-linear inter-dependence between the



constituent traits. We developed the strategy, described in Materials and methods, for identifying interactions in the presence of non-linearity by employing the residuals from regression.

Figure 3a summarizes the model. The non-linear dependence of the constituent QT for the various genotype combinations of SNP 1 and SNP 2 is highlighted in Figure 3b. Figure 3c shows that the KWII correctly identifies the main effects of SNP 1, SNP 2 and the interaction between SNP 1 and SNP 2 with the syndrome MQT vector phenotype. The variation of KWII with changes in effect size is summarized in Figure 3d. These results demonstrate that the proposed approach can be used even when non-linear dependence between the constituent phenotypes is present.

#### Performance on the hybrid 10K GAW15 data set

The goal was to demonstrate the utility of our method on a larger data set with numerous genetic markers and environmental variables. We used the hybrid simulation approach so that the ground truth required for critical assessment of the method was known. The predictors from the GAW15 simulations conducted by Miller *et al.* (2007) were retained and the vector phenotypes were randomly assigned based on the model in Figure 4a.

Figure 4b summarizes the average KWII values, standard deviations and permutation-based *P*-values for the top five one- and two-predictor containing combinations with the highest mean KWII values in the GAW15 data set. The top-ranked one-predictor containing combinations were {Age, P}, {Smoking, P}, {C16-30, P}, {C6-153, P} and {C4-330, P}. Four of the top-five combinations were present in the model (Figure 4a); the {C4-330, P} was not present in the model but had a non-significant *P*-value of 0.11. The two-predictor containing combinations with the highest mean KWII were {Age, C6-153, P}, {Age, Smoking, P}, {Age, C6-154, P}, {Age, C16-30, P} and {Smoking, C16-30, P}. The C6-154 SNP in the {Age, C6-154, P} combination is part of the same locus (HLA DR locus) that contains the C6-153 SNP present in the model and is in LD with it.

#### Computational speed

To evaluate the computational speed, we utilized the GAW15 data set as a test bed. We generated subsets of the GAW15 data set containing either 1000 or 4000 subjects, and 30, 100, 300, 3000 or all 9192 predictors in the modified GAW15 data set. The runtime requirements for obtaining the KWII values of all one- and two-predictor combinations are summarized in Table 1. The results indicate that the analysis of the GAW15 data set for 4000 subjects and 9192 predictors required ~11 h.

The SYMPHONY search algorithm is computationally much more efficient than an exhaustive search, which requires computation of all possible combinations and requires exponential time. Let *m* be the sample size of the data, *n* be the number of variables (excluding phenotype vector), *τ* be the number of iterations and *θ* be the number of combinations retained in each iteration of computing the PAI: using order notation from complexity theory, the running time of SYMPHONY can be estimated as  $O(\tau n \theta m^2 + \tau \theta 2^n m^2)$ .

#### Comparison with other methods: analysis of HDL-atherosclerosis data

The HDL-atherosclerosis data set was used to evaluate the results from SYMPHONY because this data set has been previously analyzed by two other methods for single QT.

Ishimori *et al.* (2004) used multiple regression analysis to identify main effects and interactions associated with HDL and atherosclerosis

**Table 1** Run time assessments of SYMPHONY

Number of predictors	Number of subjects	
	1000 Time (min)	4000 Time (min)
30	0.1	0.5
100	1.5	6.1
300	5.0	20.1
1000	17.0	65.2
3000	49.7	216.6
9190	143.7	655.9

individually. They found that mouse HDL concentrations were affected by six loci: two loci were located on chromosome 1 (Hdlq14 and Hdlq15) and one locus each on chromosomes 8, 9 and 12; there was also an interaction found between a locus on chromosome 2 with the Hdlq15 chromosome 1 locus. The five loci (and their combinations) associated with atherosclerosis susceptibility in mice were (1) *Ath17* on chromosome 10, (2) *Ath18* on chromosome 12, an interaction between *Ath18* and (3) *Ath19* on chromosome 11, and an interaction between (4) *Ath20* on chromosome 10 and (5) *Ath21* on chromosome 12 (Ishimori *et al.*, 2004). The same HDL and atherosclerosis phenotypes were also analyzed by Chanda *et al.* (2009) using the CHORUS information-theoretic algorithm.

Table 2 summarizes the KWII and permutation-based *P*-values for the 12 one-marker and 10 two-marker combinations with the highest KWII values for the HDL-Atherosclerosis phenotype vector. Table 2 also compares the findings from SYMPHONY to the analyses of the individual HDL and Atherosclerosis phenotypes with multiple regression and CHORUS. Among the single marker associations identified by SYMPHONY, *D10Mit28* was found to be associated with both HDL and Atherosclerosis phenotypes individually by CHORUS. The *D10Mit213* marker was found to be associated with Atherosclerosis by both CHORUS and regression, whereas *D1Mit159* was found to be associated with HDL by both CHORUS and regression. None of the other nine markers identified by SYMPHONY overlapped with CHORUS or regression.

There were five two-marker pairs that contained *D3Mit320* and two pairs that contained *D3Mit320*. Interestingly, the *D3Mit320* had been identified as being associated with Atherosclerosis by CHORUS and four of the remaining markers (*D12Mit54*, *D12Mit172*, *D12Mit84* and *D12Mit243*) were associated with HDL levels. This suggests that interactions between loci on mouse chromosomes 3 and 12 are involved in the HDL-atherosclerosis correlations. Four markers on chromosome 12 (*D12Mit170*, *D12Mit243*, *D12Mit84*, and *D12Mit91*) were also involved in interactions with the *DXMit159* marker on the X-chromosome.

Overall, these comparisons suggest that the findings from SYMPHONY overlap to a greater extent with CHORUS, which is to be expected given that they share the same information-theoretic framework. These findings further demonstrate that the vector phenotype approach in SYMPHONY complements the results from analysis of single traits in important ways. Some of the results from SYMPHONY overlap partially with regression and CHORUS methods but importantly, SYMPHONY provides novel results that can direct further investigation.

**Table 2 Comparison of the one- and two-marker results from SYMPHONY for the HDL-Atherosclerosis vector phenotype.**

Marker	KWII	KWII P-value	Regression	CHORUS
D13Mit16	5.28	0.0004		
D11Mit285	5.21	0.0001		
D10Mit28	5.13	0.0008		H, A
D10Mit86	5.04	0.0008		
D12Mit7	5.04	0.0009	A	
D4Mit308	5.02	0.0008		
D10Mit213	4.93	0.0016	A	A
D1Mit65	4.92	0.0016		
D4Mit17	4.91	0.0012		
D4Mit196	4.88	0.0015		
D1Mit107	4.88	0.0017		
D1Mit159	4.87	0.0008	H	H

Marker pair		KWII	KWII P-value	Regression		CHORUS	
Marker 1	Marker 2			Marker 1	Marker 2	Marker 1	Marker 2
D3Mit320	D12Mit54	3.78	<0.0001			A	H
D3Mit320	D12Mit172	3.68	<0.0001		H	A	H, A
D3Mit320	D12Mit84	3.44	<0.0001			A	H, A
D3Mit320	D12Mit243	3.39	<0.0001		A	A	H
D3Mit320	D4Mit227	3.25	0.0001			A	
D3Mit318	D15Mit107	3.04	0.0001			A	
D3Mit318	D15Mit243	2.94	<0.0001			A	
D12Mit170	DXMit159	2.90	0.0036			H	
D12Mit243	DXMit159	2.85	0.0008	A		H	
D12Mit84	DXMit159	2.83	0.0045			H, A	
D12Mit91	DXMit159	2.80	0.0002			H, A	

Abbreviations: HDL, high-density lipoprotein; KWII, K-way interaction information. The results from the analyses of the HDL and Atherosclerosis phenotypes using multiple regression (Ishimori *et al.*, 2004) and the information-theoretic CHORUS algorithm (Chanda *et al.*, 2009) are shown. The markers associated with HDL and Atherosclerosis are denoted by H and A, respectively. Those associated with both HDL and Atherosclerosis are denoted by H, A.

**Performance on the hybrid 1000 genomes data set**

The goal of these simulations was to demonstrate the utility of SYMPHONY for analysis of genetic data from a genome-wide study and to assess the dependence of the KWII on LD, MAF and admixture. The SNP genotypes from the 1000 Genomes data were used and the vector phenotypes were simulated based on the model described for Case Study 1.

*Minor allele frequency.* Figure 5a shows the dependence of the KWII on the MAF of SNP 2 with the MAF of SNP 1 kept constant at 0.5. The KWII values of the first-order {2, P} combination and the second-order {1, 2, P} combination increase monotonically with increasing SNP 2 MAF.

*Linkage disequilibrium.* Figure 5b summarizes the effect of LD (as measured by R<sup>2</sup>) between SNP 1 and SNP 2 on the KWII. The KWII for the {1, P} and {2, P} first-order combinations converge at the highest LD value because the genotypes become highly correlated with each other. The KWII value of the second-order combination {1, 2, P} decreases monotonically with increasing LD and becomes negative, indicating that information present in this combination is rendered redundant as a result of increased LD.

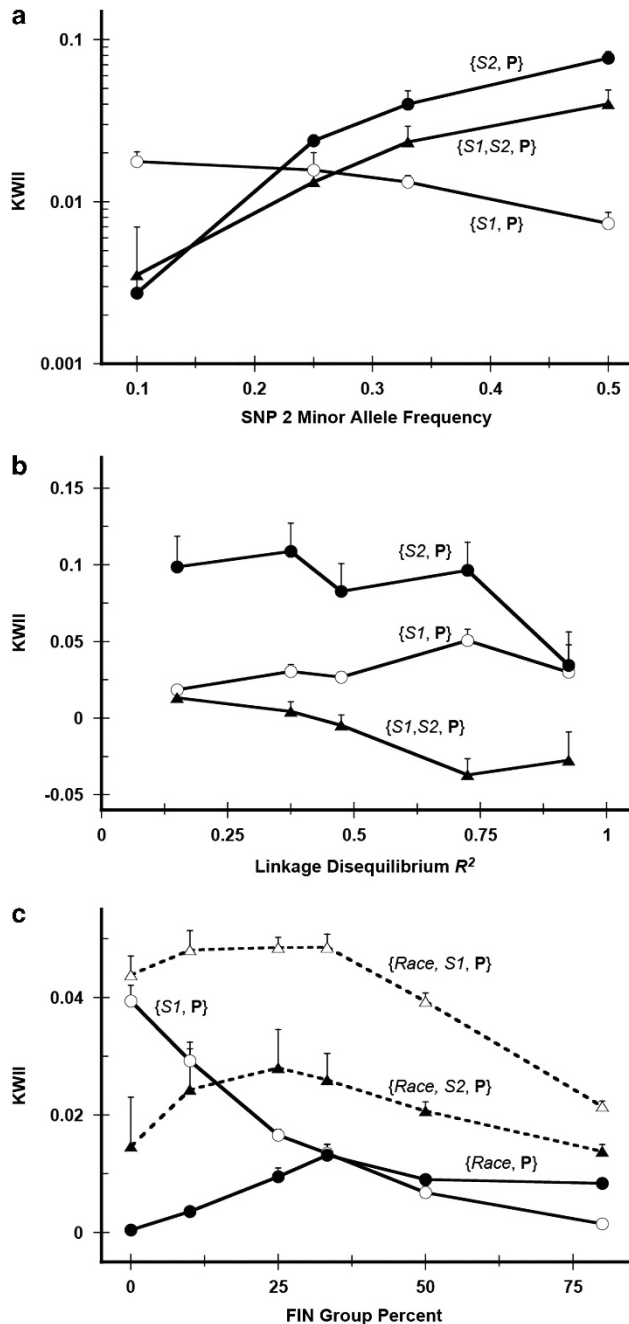
*Population stratification/admixture.* Figure 5c shows the effects of population stratification/admixture, which was assessed by increasing the percentage of the sample from the FIN group in the population. The KWII patterns for the {1, P} and {2, P} combinations were

similar (data for {2, P} not shown) and decreased monotonically as the percentage of the FIN group increased. This is a consequence of the absence of SNP 1 and SNP 2 effects in the FIN group. The {Race, P} combination reached a maximal value at a FIN group percentage of 33% when the relative proportions of all three groups—CHS, TSI and FIN—were equal. In the admixture simulation, we found high KWII values for the second-order {Race, 1, P} and {Race, 2, P} combinations. At FIN group, percentages >33% the KWII values decreased because of the absence of SNP 1 and SNP 2 effects in this group.

**DISCUSSION**

In this report, we have presented results for a novel information-theoretic GGI and GEI analysis method for disease syndromes. The salient contributions of the paper include the development of an MQT vector framework for describing the phenotype of the syndrome and the demonstration that the versatile and general framework previously developed in AMBIENCE and CHORUS for the genetic analysis of discrete and QTs can be deployed for vector traits.

Our approach is based on the assumption that the MQT vector phenotype is adequately modeled by the multivariate normal distribution. Several statistical tests for assessing whether or not a given data set follows a multivariate normal distribution are available (Cox and Small, 1978; Friedman and Rafsky, 1979; Smith and Jain, 1988; Henze, 2002). In addition to its statistical underpinnings in the Central Limit theorem, the normal distribution  $N(\mu, \sigma)$  has



**Figure 5** Results for hybrid simulations with the 1000 genomes set: Case Study 1 was used to simulate the vector phenotype. (a) The dependence of the KWII of the informative  $\{1, P\}$ ,  $\{2, P\}$  and  $\{1, 2, P\}$  SNP combinations on the minor allele frequency (MAF) of *SNP 2*. (b) Dependence of the KWII on the linkage disequilibrium as assessed by  $R^2$  value. (c) Dependence of the KWII on admixture as assessed by varying the percentage of the FIN group in the population while keeping the percentages of the CHS and TSI groups equal. The *Race* variable corresponds to the label indicating group membership (CHS, TSI or FIN). The KWII for the  $\{S2, P\}$  combination was similar to that for the  $\{S1, P\}$  combination and is not shown. The error bars in each plot are standard errors from five independent simulations.

maximum entropy among all real-valued distributions with specified mean  $\mu$  and standard deviation  $\sigma$  (Srivastava and Gupta, 2008). The findings from the univariate normal distribution also generalize to the multivariate normal distribution  $N(\mu, \Sigma)$ , which has maximum

entropy among all real-valued distributions with specific mean vector  $\mu$  and covariance matrix  $\Sigma$ . Therefore, a multivariate normal distribution is reasonable if only the mean and covariance matrix of MQT vector phenotype are known. A question that is outside the scope of the current research is the nature and number of the dimensions that should comprise the syndrome vector. The choice of the dimensions should be based on clinical and biometric characteristics that provide sensitivity and specificity for diagnosing and measuring the progression of the syndrome pathology objectively in the clinical or study setting. However, there may be a lack of agreement on whether a particular syndrome is a distinct clinical entity; in such situations, there may not be clinical consensus on the best ways diagnose and measure it.

We have provided results that demonstrate proof of concept that (with modifications) the KWII method may be generalizable to certain situations wherein there are non-linear relationships between the QT constituents comprising the vector phenotype. However, the method relies on regression. More research is necessary to define the range and applicability of the approach and its power to detect the potentially diverse range of non-linearities that could occur.

A range of GEI analysis methods based on the multi-factor dimensionality reduction (MDR) method are available for identifying and analyzing GEI (Ritchie *et al.*, 2001; Hahn *et al.*, 2003; Moore *et al.*, 2006). MDR reduces the dimensionality of the multi-locus genotype systematically by pooling into high and low risk groups (Ritchie *et al.*, 2001; Hahn *et al.*, 2003; Ritchie *et al.*, 2003; Bush *et al.*, 2006; Moore *et al.*, 2006). The combinatorial partitioning method and the restricted partition method (Culverhouse *et al.*, 2004) share similarities to MDR, but identify multilocus genotypes capable of predicting QT levels (Nelson *et al.*, 2001). The generalized MDR method employs the generalized linear model framework for scoring in conjunction with MDR for dimensionality reduction. Generalized MDR enables covariate corrections and handles both discrete phenotypes and continuous traits in population-based study designs. The Pedigree Disequilibrium Test (Martin *et al.*, 2000) approach has been used to extend MDR, which was initially limited in its capacity to include potentially informative family data beyond single matched pairs in each family, to family-based study designs (Martin *et al.*, 2006). MDR and its variants including generalized MDR are not capable of GEI analysis of syndromes.

Some machine learning methods including random forests and decision trees employ information gain metrics as an aid to decision making. Random forest methodologies are now being applied to GWAS studies (Kim *et al.*, 2009; Meng *et al.*, 2009; Wu, 2011); however, the design and role of the information-theory metrics are different from that in SYMPHONY. Additionally, random forest and decision tree algorithms have generally been limited to discrete phenotypes. The Markov blanket method has been used to analyze interactions at the genome-wide level with case-control study designs (Han *et al.*, 2010). The extensions of these methods for analyzing vector phenotypes have not yet been evaluated systematically.

The KWII and PAI metrics that provide the underpinning of SYMPHONY are robust to LD. The second term on the right-hand side of the PAI definition contains the inter-dependence among the genetic (for example, pairwise LD) and environmental variables (for example, common source of multiple pollutants). Unlike traditional LD measures, which are pairwise in character (and typically are computed for contiguous SNPs), the PAI and KWII can assess the joint dependence between variables generalized to multiple, non-adjacent SNPs. We have extensively investigated the robustness

of the PAI over a range of LD values for discrete and QTs (Chanda *et al.*, 2008, 2009). The robustness of the PAI to LD does not compromise its ability to identify informative interactions.

In previous work, we have shown that the KWII is also effective at identifying two-locus interaction models that do not contain main effects (Chanda *et al.*, 2009; Sucheston *et al.*, 2010). Because SYMPHONY utilizes a marginal effect search strategy based on the PAI, its power is compromised when main effects are completely absent. However, interactions that lack main effects entirely are relatively rare in real data because of the stringent symmetry required to enforce absence of main effects (Chanda *et al.*, 2008). Interactions in real data generally contain traces of main effects that can be leveraged by the SYMPHONY search. SYMPHONY can also easily be modified so that the entire space of second-order combinations is searched to identify interactions which could overcome the loss of power (Chanda *et al.*, 2008).

In this report, we have focused on syndromes comprising multivariate normally distributed QT. However, many clinical syndrome vector phenotypes may comprise discrete or continuous non-normal variables. For example, subjects with autism may be described by the severity of the social and communication symptoms, the severity of fixated behaviors as well as associated features such as presence of genetic risk factors, epilepsy and intellectual disability. The framework developed in the paper can potentially be extended to syndrome phenotype vectors containing discrete and multivariate normally distributed variables in a straightforward manner since the phenotype can be represented as a mixture of multinormal distributions. Data transformations such as the logarithm and the cube root, which are well-established approaches to obtain normal distributions from non-normal data, could also be useful if an individual QT is not normally distributed. However, there are numerous challenges related to computations of the entropy expression for more complex multivariate distributions.

In conclusion, our results suggest that the availability of versatile methods based on information-theoretic metrics could enhance the GGI and GEI analysis for complex disease phenotypes.

#### DATA ARCHIVING

There were no data to deposit.

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### ACKNOWLEDGEMENTS

Support from the National Multiple Sclerosis Society (RG3743 and 4836-A-5) and the Department of Defense (MS090122) to the Ramanathan laboratory is gratefully acknowledged.

- Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C *et al.* (2007). Information-theoretic metrics for visualizing gene-environment interactions. *Am J Hum Genet* **81**: 939–963.
- Cox DR, Small NJH (1978). Testing multivariate normality. *Biometrika* **65**: 263–272.
- Culverhouse R, Klein T, Shannon W (2004). Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* **27**: 141–152.
- Dirckx JH (2001). In: Dirckx JH (ed.) *Stedman's Concise Medical Dictionary for the Health Professionals*. Lippincott, Williams and Wilkins: Philadelphia, PA.
- Fano RM (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press: Cambridge, MA.
- Ford ES, Giles WH, Dietz WH (2002). Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey. *JAMA* **287**: 356–359.
- Friedman JH, Rafsky LC (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two sample tests. *Ann Stat* **7**: 697–717.
- Gokhale DV, Ahmed NV, Res BC, Piscataway NJ (1989). Entropy expressions and their estimators for multivariate distributions. *IEEE Trans Inform Theory* **35**: 688–692.
- Hahn LW, Ritchie MD, Moore JH (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* **19**: 376–382.
- Han B, Park M, Chen XW (2010). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics* **11**: Suppl 3 S5.
- Han TS (1980). Multiple mutual informations and multiple interactions in frequency data. *Inform Control* **46**: 26–45.
- Henze (2002). Invariant tests for multivariate normality: a critical review. *Stat Pap* **43**: 467–506.
- Ishimori N, Li R, Kelmenson PM, Korstanje R, Walsh KA, Churchill GA *et al.* (2004). Quantitative trait loci analysis for plasma HDL-cholesterol concentrations and atherosclerosis susceptibility between inbred mouse strains C57BL/6J and 129S1/SvJmJ. *Arterioscler Thromb Vasc Biol* **24**: 161–166.
- Jakulin A (2005). Machine learning based on attribute interactions. *PhD thesis*. University of Ljubljana: Ljubljana, Slovenia.
- Jakulin A, Bratko I (2004). In: Greiner R, Schuurmans D (eds). *Proceedings of the Twenty-First International Conference on Machine Learning (ICML-2004)*. Banff, Canada, pp 409–416.
- Kim Y, Wojciechowski R, Sung H, Mathias RA, Wang L, Klein AP *et al.* (2009). Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proc* **3**: Suppl 7 S64.
- Knights J, Ramanathan M (2012). An information theory analysis of gene-environmental interactions in count/rate data. *Hum Hered* **73**: 123–138.
- Martin ER, Monks SA, Warren LL, Kaplan NL (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* **67**: 146–154.
- Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH (2006). A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genet Epidemiol* **30**: 111–123.
- McGill WJ (1954). Multivariate information transmission. *Psychometrika* **19**: 97–116.
- Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL (2009). Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* **10**: 78.
- Miller MB, Lind GR, Li N, Jang S-Y (2007). 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 Genet* **1**: Suppl 1 S4.
- Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N *et al.* (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* **241**: 252–261.
- Nelson MR, Kardia SL, Ferrell RE, Sing CF (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* **11**: 458–470.
- Ritchie MD, Hahn LW, Moore JH (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* **24**: 150–157.
- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF *et al.* (2001). Multifactor dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* **69**: 138–147.
- Shannon CE (1997). The mathematical theory of communication. 1963. *MD Comput* **14**: 306–317.
- Smith SP, Jain AK (1988). A test to determine the multivariate normality of a dataset. *IEEE Trans Pattern Anal Mach Intell* **10**: 757–761.
- Srivastava S, Gupta MR (2008). *IEEE International Symposium on Information Theory (ISIT)*. Toronto.
- Sucheston L, Chanda P, Zhang A, Tritchler D, Ramanathan M (2010). Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity. *BMC Genomics* **11**: 487.
- Watanabe S (1960). Information theoretical analysis of multivariate correlation. *IBM J Res Develop* **4**: 66–82.
- Wu Q (2011). *IEEE International Conference on Bioinformatics and Biomedicine*.

Bush WS, Dudek SM, Ritchie MD (2006). Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics* **22**: 2173–2174.

Chanda P, Sucheston L, Liu S, Zhang A, Ramanathan M (2009). Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits. *BMC Genomics* **10**: 509.

Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, Ambrosone C *et al.* (2008). AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* **180**: 1191–1210.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)