Heredity (2011) 107, 320–327

www.nature.com/hdy

ORIGINAL ARTICLE

Modeling of environmental and genetic interactions with AMBROSIA, an information-theoretic model synthesis method

P Chanda[1], A Zhang[1] and M Ramanathan[2]
[1]*Department of Computer Science and Engineering, State University of New York, Buffalo, NY, USA and* [2]*Department of Pharmaceutical Sciences, State University of New York, Buffalo, NY, USA*

To develop a model synthesis method for parsimoniously modeling gene–environmental interactions (GEI) associated with clinical outcomes and phenotypes. The AMBROSIA model synthesis approach utilizes the *k*-way interaction information (KWII), an information-theoretic metric capable of identifying variable combinations associated with GEI. For model synthesis, AMBROSIA considers relevance of combinations to the phenotype, it precludes entry of combinations with redundant information, and penalizes for unjustifiable complexity; each step is KWII based. The performance and power of AMBROSIA were evaluated with simulations and Genetic Association Workshop 15 (GAW15) data sets of rheumatoid arthritis (RA). AMBROSIA identified parsimonious models in data sets containing multiple interactions with linkage disequilibrium present. For the GAW15 data set containing 9187 single-nucleotide polymorphisms, the parsimonious AMBROSIA model identified nine RA-associated combinations with power $>90\%$. AMBROSIA was compared with multifactor dimensionality reduction across several diverse models and had satisfactory power. Software source code is available from http://www.cse.buffalo.edu/DBGROUP/bioinformatics/resources.html. AMBROSIA is a promising method for GEI model synthesis.
Heredity (2011) **107,** 320–327; doi:10.1038/hdy.2011.18; published online 23 March 2011

Keywords: gene–environment interactions; gene–gene interactions; *k*-way interaction information

Introduction

The potentially ubiquitous roles of environmental factors, gene–gene interactions (GGI) and gene–environmental interactions (GEI) in pathophysiology could be contributing factors to the somewhat limited success of genome-wide association studies of complex diseases (Weiss and Terwilliger, 2000; Ambrosone *et al*., 2007; Goldstein, 2009).

Methodologies for measuring environmental exposures and determining genetic variations have advanced rapidly. However, analysis tools to evaluate the combined effects of numerous environmental exposures and multiple genetic variations on disease risk and on clinical and treatment outcomes are lacking.

Several critical computational problems limit GEI analysis in clinical pharmacogenetics and human epidemiology. These include the following: (i) identifying effective metrics to detect disease risk-associated GEI, (ii) developing efficient algorithms to tackle the inherent combinatorial complexity of interaction analysis, and (iii) implementing modeling strategies to identify the critical GEI and to systematically extract knowledge. This manuscript presents on a modeling strategy for GEI analysis based on a novel and generalizable information-theoretic framework.

Modeling methods complement GEI identification approaches by reducing data complexity and enabling users to assess the key relationships of different genetic and environmental factor interactions to the disease. Several factors make GEI modeling difficult. For example, the presence of linkage disequilibrium among the genetic and correlations among the environmental variables burdens modeling methodologies with redundant information. Many complex diseases exhibit genetic heterogeneity (GH) with different underlying causes for the same disease or treatment phenotype present in the data.

The usefulness of the information-theoretic *k*-way interaction information (KWII) metric and the AMBIENCE and CHORUS algorithms for GEI analysis of discrete phenotypes and quantitative traits has recently been demonstrated (Chanda *et al*., 2007, 2008). This paper focuses on development and critical assessment of AMBROSIA, an information-theoretic model-synthesis method for GEI. The methodology differs uniquely from available approaches in its conceptual framework, versatility and computational efficiency.

Correspondence: Dr M Ramanathan, Department of Pharmaceutical Sciences, State University of New York, 427 Cooke Hall, Buffalo, NY 14260, USA.
E-mail: Murali@Buffalo.Edu
Received 25 August 2010; revised 30 December 2010; accepted 1 February 2011; published online 23 March 2011

Materials and methods

Definitions and terminology

The terminology, developed by Chanda *et al*. (2007, 2008) is concisely recapitulated here.

GGI and GEI: The methods are applicable to both GEI and GGI analyses; we will use the term GEI for both.

Model and model synthesis: A model is 'a parsimonious set of variable combinations capable of explaining the phenotype'. Model synthesis is the procedure for identifying a model.

Entropy: The entropy, $H(X)$, of a discrete random variable X can be computed from the probabilities of $p(x)$ using:

$$H(X) = -\sum_{x} p(x) \ln p(x)$$

k-Way interaction information (KWII): The KWII is used to define interactions (Chanda *et al*., 2008). For the k-variable case on the set of predictors $v = \{X_1;X_2;\ldots;X_K\}$ and a phenotype variable Y, the KWII can be expressed as an alternating sum (Han, 1980) over all possible subsets T of $\{v,Y\}$:

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

The $|\{v,Y\}|$ and $|T|$ terms in the exponent represent the size of the combination $\{v,Y\}$ and the size of the subsets T. The number of predictor variables $K = |\{v\}|$ in a combination is called the order of the combination.

The KWII represents 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, multivariate interaction metric and does not contain contributions arising from the KWII of other lower order combinations of these variables.

Interaction: Our operational definition is as follows: 'a positive KWII value for a variable combination indicates the presence of an interaction, negative KWII values indicate the presence of redundancy, and a KWII value of zero denotes the absence of k-way interactions' (Chanda *et al*., 2008).

AMBIENCE algorithm: AMBIENCE is an information-theoretic search method for detecting GEI. It uses a computationally efficient hill-climbing algorithm to identify the most promising regions in combinatorial space so that the number of KWII computations is reduced. The inputs to AMBIENCE are θ, the number of combinations retained in the each search iteration, and τ, the number of iterations, which determines the highest order of variable combination detected. For details, see Chanda *et al*. (2008).

The output from AMBIENCE consists of $\tau\theta$ combinations and their associated KWII values. The order of the combinations ranges from 1 through τ, and the θ combinations with the highest KWII values within each order identified by AMBIENCE are provided as a sorted list. AMBIENCE also provides permutation-derived p-values as described in the following section.

KWII *p*-values: The p-value of the KWII of each combination was determined using 10 000 permutations of the phenotype (Sucheston *et al*., 2010). The permutations for each combination were conducted independently of the other combinations using a fast algorithm from Patefield (1981). The permutation procedure provides the null distribution of the KWII, that is, when the combination of variables was not associated with the phenotype. The p-value for the combination was defined as the proportion of permutations with KWII values that were greater than or equal to the observed KWII.

AMBROSIA algorithm

The most promising variable combinations identified in the AMBIENCE output are input for the model synthesis step, AMBROSIA.

AMBROSIA employs an iterative procedure to efficiently identify the set of variable combinations capable of explaining the phenotype. It prospectively assesses measures of the relevancy, redundancy and parsimony when specific variable combinations are included in the model. The entire model synthesis method is based on the KWII, and repeated refitting of each model to the data is unnecessary. The pseudocode for AMBROSIA is summarized in Supplementary Figure 1.

The input to AMBROSIA is the set S of promising variable combinations and their KWII values that were output from AMBIENCE. The output from AMBROSIA is the parsimonious set of combinations M that explain the phenotype, denoted by Y.

The number of parameters P of a model corresponds to the number of degrees of freedom of M. The Model information content (MIC) is defined as the sum of the KWIIs of the combinations C in M:

$$MIC(M) = \sum_{C \in M} KWII(C)$$

Initially, M is empty. In step 1, combinations with negative KWII and combinations with nonsignificant KWII values as assessed by permutation testing are eliminated from set S. The significance of the KWII for each combination was conducted with 10 000 independent permutations of the phenotype Y.

Step 2: Each AMBROSIA iteration contains three distinct parts that employ the KWII, that is: 2A) Combination selection, 2B) Parsimony evaluation and 2C) Redundancy evaluation.

Step 2A, Combination selection: In the first iteration, *Combination 1*, the combination with the highest KWII identified by AMBIENCE present in S, is initially added to the model for evaluation. In the subsequent iterations, the *Combination j* with highest KWII among the remaining combinations is added to M for evaluation in Step 2B.

Further, a combination remaining in S is added directly to M if all of its proper phenotype-containing subsets are already present in M. This is possible because the number of parameters in the model does not increase when such a combination is added.

Step 2B, Parsimony evaluation: For parsimony evaluation, we employ a heuristic metric, the corrected MIC (MICC), motivated by the Akaike Information Criterion (Hurvich and Tsai, 1995). The MICC rewards goodness of fit as assessed by the KWII and penalizes for the number of parameters P in the model.

$$MICC = 2N\Delta MIC - 2\Delta P$$

In the MICC definition above, N is the sample size, MIC is the change in MIC and P is the increase in the number

of parameters in the model due to the inclusion of a new combination. The first term is analogous to a log likelihood, whereas the second term is a penalty for increased model complexity. The heuristic was based on two known results. First, the KWII is the Kullback–Leibler divergence between the joint probability density and the Kirkwood superposition approximation (Jakulin and Bratko, 2004), and second, that the Kullback–Leibler divergence is the expected log likelihood. For combinations containing three variables, the KWII is the Kullback–Leibler divergence between the joint probability density and the model constructed using all pairwise dependences (Jakulin and Bratko, 2004).

The combination is eliminated from both *M* and *S* if the $e^{-MICC} \geqslant \eta$, where η is a user-defined threshold.

Step 2C, Redundancy evaluation: Let {*Combination i*} represent a combination already in *M*. The KWII value defined as, *KWII*({*Combination i*}, {*Combination j*}, *Y*) was computed for each combination in *M*. *Combination j* was eliminated from both *M* and *S* if the KWII was negative, because it indicates redundancy with *Combination i*. This step effectively eliminates inclusion of new combinations in strong linkage disequilibrium (LD) with combinations already in the model. *Combination j* is retained in *M* if it is not eliminated after Step 2C.

Step 2 is repeated until *S* is empty. The output is the list of combinations in *M*. Model size was defined as the number of combinations in *M*.

Step 3, Model significance evaluation: The significance of MIC was assessed with 10 000 independent permutations of the phenotype.

A demonstration of the AMBROSIA is presented in results that provide a step-by-step description of the method.

Simulations for evaluating AMBROSIA

Case study 1 and case study 2: The GGI models are shown in Supplementary Figure 2A. The case-control design with 2500 cases and 2500 controls was used for both case studies. Details of the simulation methods are in supplementary methods. The case-control status phenotype variable was denoted by *Y*.

Case study 1 contained 60 single-nucleotide polymorphisms (SNPs) in four groups, G_1 through G_4, with linkage disequilibrium within each group (Supplementary Figure 2B). The disease-causing SNPs were *SNP 7* and *22*.

Case study 2 contained 120 SNPs in eight groups (Supplementary Figure 2C). Case study 2 contained GH with two pairs of interacting loci, *SNP 7* with *SNP 22* and *SNP 67* with *SNP 82*, that each increased risk in half of the cases.

The relative risk values used in simulation for Case studies 1 and 2 were 1.2, 1.5, 1.8, 2.0 and 2.5.

Robustness of AMBROSIA to LD patterns and allele frequency

The robustness of AMBROSIA in the presence of realistic LD patterns and variations in allele frequency was assessed with data from problem 2 of Genetic Analysis Workshop 15 (GAW15, http://www.gaworkshop.org/about/publications.html#gaw15) containing the dense panel of 2300 SNPs genotyped with the Illumina platform (Illumina Inc., San Diego, CA, USA) for a 10 kb region of chromosome 18q in 920 subjects.

The data set was preprocessed to remove samples with missing data and SNPs not in Hardy–Weinberg equilibrium (χ^2 test at $\alpha = 0.05$). A set of 865 tagSNPs were selected using the method of Carlson *et al*. (Carlson *et al*., 2004), with an LD threshold of $R^2 =$ this preprocessed 895-subject data set as the GAW15-P2 data set.

We generated a sample of 2500 cases and 2500 controls by re-sampling with replacement from GAW15-P2 data. To assess robustness to LD patterns, we identified SNPs in the data set with MAF of 0.5 ± 0.01. We selected a random pair of such SNPs, say SNP *i* and SNP *j*. For each individual in the population, the case-control status was randomly assigned on the basis of penetrance matrix for interaction model of Case study 1 with the genotypes of SNP *i* and SNP *j* as inputs. The relative risk value was 2.0. This process was repeated for 100 random pairs of SNPs with MAF of 0.5 ± 0.01.

To assess the robustness of AMBROSIA to allele frequency variations, we identified SNPs with MAF values within ± 0.01 of 0.4, 0.3, 0.2 and 0.1. The computational strategy described in the previous paragraph was used.

The AMBIENCE input parameter values for robustness assessments were $\theta = 50$ and $\tau = 2$. Robustness was assessed by comparing power to the results of Case study 1.

Comparisons with multifactor dimensionality reduction (MDR)

AMBROSIA was compared head to head against MDR (http://sourceforge.net/projects/mdr/) (Ritchie *et al*., 2001). Two separate sets of comparisons with MDR were conducted. The first set of comparisons used a model based on case study 2, with a relative risk of 2.0. For the second set of comparisons, four two-locus interaction models, 1-GH, 2-GH, 3-GH and 4-GH, employed in the original MDR power evaluation paper by Ritchie *et al*. were used (Ritchie *et al*., 2003). The penetrance matrices for the latter set of models from Ritchie *et al*. (Ritchie *et al*., 2003) are summarized in Table 1.

The detailed methodology for these comparisons is presented in Supplementary Methods.

Table 1 Penetrance matrices for the comparisons of AMBROSIA to MDR

Model 1-GH $K_p = 0.05$, $h^2 = 0.013$				*Model 2-GH* $K_p = 0.025$, $h^2 = 0.013$			
	BB	*Bb*	*bb*		*BB*	*Bb*	*bb*
AA	0.0	0.1	0.0	*AA*	0.0	0.0	0.1
Aa	0.1	0.0	0.1	*Aa*	0.0	0.05	0.0
aa	0.0	0.1	0.0	*aa*	0.1	0.0	0.0
Model 3-GH $K_p = 0.06$, $h^2 = 0.007$				*Model 4-GH* $K_p = 0.025$, $h^2 = 0.003$			
	BB	*Bb*	*bb*		*BB*	*Bb*	*bb*
AA	0.08	0.07	0.05	*AA*	0.07	0.05	0.02
Aa	0.1	0.0	0.1	*Aa*	0.05	0.09	0.01
aa	0.03	0.1	0.04	*aa*	0.02	0.01	0.03

Abbreviations: GH, genetic heterogeneity; MDR, multifactor dimensionality reduction.
The penetrance values are based on the models in Ritchie *et al*. (2003). K_p denotes disease prevalence, whereas h^2 denotes heritability.

Analysis of public domain data sets

The data for problem 3 of GAW15 contains 9187 di-allelic SNPs distributed on the genome to mimic a 10-K SNP chip. There are 100 replicates modeled after rheumatoid arthritis (RA) data (Miller *et al*., 2007). The data were treated as case-control data and the unphased genotypes were analyzed.

The interactions in the simulation framework (from http://genetsim.org/gaw15/answers/) are summarized in Supplementary Table 2. There are interactions involving nine loci as follows: C, DR and D on chromosome 6, A on chromosome 16, B on chromosome 8, E on chromosome 18, F on chromosome 11, and G and H on chromosome 9.

We refer to this data set as the GAW15-P3 data set. We conducted analyses with RA affection status as the phenotype and sex, age, smoking status as nongenetic variables. Age was discretized by binning into five intervals of equal width.

The AMBIENCE input parameter values were $\theta = 50$ and $\tau = 2$. We used all 100 replicates to obtain the power of AMBROSIA.

Results

Demonstration AMBROSIA run

The pseudocode for AMBROSIA is summarized in Supplementary Figure 1. Here, we present a step-by-step demonstration run of AMBROSIA for the GAW15-P3 data set.

We used the top 10 one-SNP and two-SNP combinations with the highest KWII values identified by AMBIENCE to build a parsimonious model using AMBROSIA. The KWII values are summarized in Table 2.

In the first step, we confirmed that all combinations identified by AMBIENCE were significant using permutation-based approaches with $\alpha = 0.001$. The value of η was set to 0.001.

The combination {*C6_153*, *RA*} is the first combination in model *M* because it has the highest KWII.

The combinations {*C6_154*, *RA*} and {*C6_152*, *RA*} were evaluated in order. Because KWII(*C6_153*, *C6_154*, *RA*) and KWII(*C6_153*, *C6_152*, *RA*) were both negative, indicating redundancy with *M*, both failed to enter *M*.

The combination {*Age*, *RA*} passed the complexity test (MICC = 1050 making $e^{-MICC} < \eta$) and redundancy test (KWII(*C6_153*, *Age*, *RA*) = 0.035 > 0) and entered *M*. Combination {*C6_153*, *Age*, *RA*} enters *M* because it has positive KWII and both its subcombinations are already in *M*.

The sequential use of the complexity and the redundancy tests was continued until all the combinations remaining in *S* were evaluated. We obtained the final model *M* = {{*C6_153*, *RA*}, {*Age*, *RA*}}, {*C6_153*, *Age*, *RA*}, {*Sex*, *RA*}, {*C11_389*, *RA*}, {*C11_389*, *Age*, *RA*}, {*C11_389*, *Sex*, *RA*}, {*Smoking*, *RA*}, {*C6_162*, *RA*}, {*C6_162*, *Age*, *RA*}, {*C6_162*, *Sex*, *RA*}}. On the basis of MIC, the overall model had a permutation-derived $P < 0.0001$.

An additional AMBROSIA demonstration run for case study 1 is summarized in supplementary results.

Table 2 Top 10 one- and two-way combinations for GAW15-P3 data set based on the KWII values

One-way combinations	*KWII*	*Two-way combinations*	*KWII*
{*C6_153, RA*}	0.2550	{*C6_153, Age, RA*}	0.0347
{*C6_154, RA*}	0.2273	{*C6_154, Age, RA*}	0.0272
{*C6_152, RA*}	0.1110	{*C6_152, Age, RA*}	0.0138
{*Age, RA*}	0.0950	{*C6_155, Age, RA*}	0.0092
{*Sex, RA*}	0.0471	{*C6_136, C6_146, RA*}	0.0066
{*C6_155, RA*}	0.0377	{*C6_139, C6_150, RA*}	0.0050
{*C11_389, RA*}	0.0134	{*C6_162, Age, RA*}	0.0038
{*Smoking, RA*}	0.0117	{*C11_389, Age, RA*}	0.0033
{*C6_162, RA*}	0.0105	{*C11_389, Sex, RA*}	0.0021
{*C6_139, RA*}	0.0064	{*C6_162, Sex, RA*}	0.0014

Abbreviations: GAW15, Genetic Association Workshop 15; KWII, *k*-way interaction information; *RA*, rheumatoid arthritis.
The results are sorted by KWII values.

Performance and power of AMBROSIA

Case study 1: This case study contains the simulated LD patterns shown in Supplementary Figure 2B. The power of AMBROSIA (Figure 1a) increases sharply with increasing relative risk. At a relative risk of 1.5, the power for the {*7*, *Y*} and {*7*, *22*, *Y*} combinations were 91 and 70%, respectively. At a relative risk of 1.8 or greater, power for both {*7*, *Y*} and {*7*, *22*, *Y*} were 98% or greater.

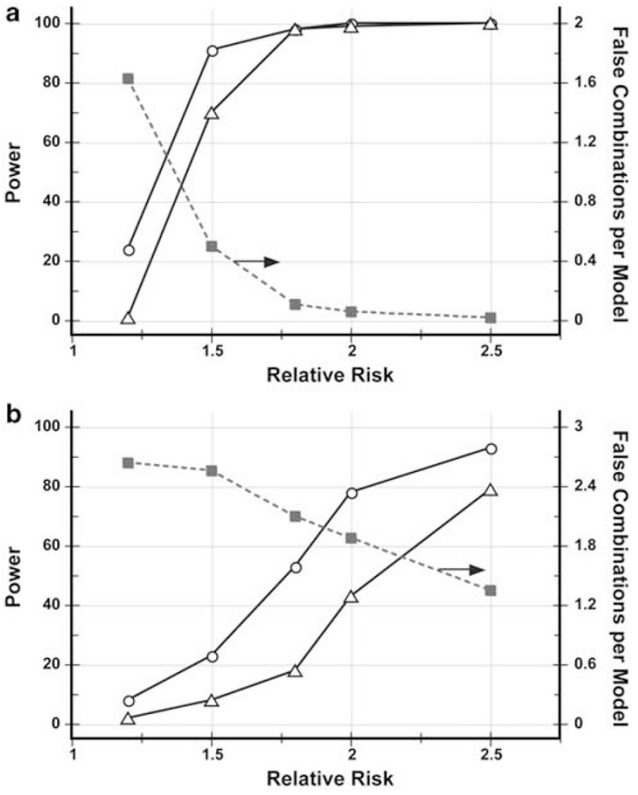

Figure 1 (**a**) Power for detecting interactions in case study 1 on the left axis and the FCM (filled squares) for case study 1 on the right axis. The open circles represent the {*7*, *Y*} combination and the open triangles represent the combination {*7*, *22*, *Y*}. The results for the {*22*, *Y*} combination were nearly identical to the {*7*, *Y*} combination and are not shown. (**b**) Power for detecting the {*7*, *Y*} and {*7*, *22*, *Y*} combinations in case study 2 and the FCM (filled squares) for case study 2 on the right axis. The results for the {*22*, *Y*}, {*67*, *Y*} and {*82*, *Y*} combinations were nearly identical to the {*7*, *Y*} combination (open circles) and are not shown; the power for the {*67*, *82*, *Y*} combination was nearly identical to the {*7*, *22*, *Y*} combination (open triangles) and is also not shown.

The number of false combinations per model (FCM) decreased with increasing relative risk (Figure 1a, right axis). The FCM value at a relative risk of 1.2 was 1.6, but it decreased to 0.5 at relative risk of 1.5, and the FCM value was 0.06 for relative risk of 2.0. The model size at a relative risk of 1.2 was 2.1 and the model size at a relative risk of 1.5 was 3.0. The results indicate that AMBROSIA is a promising model-synthesis method.

Case study 2, GH: Case study 2 (Supplementary Figure 2C) contained high levels of short-range LD and also GH with two pairs of interacting loci, *SNP 7* with *SNP 22* and *SNP 67* with *SNP 82*. The AMBROSIA model correctly contained four interactions involving single SNPs, {*7*, *Y*}, {*22*, *Y*}, {*67*, *Y*} and {*82*, *Y*}, and two interactions involving two SNPs, {*7*, *22*, *Y*} and {*67*, *82*, *Y*}. Figure 1b shows the power and FCM as a function of relative risk for the interactions in case study 2.

For a relative risk of 2.0, the power of AMBROSIA for detecting the first-order interactions {*7*, *Y*}, {*22*, *Y*}, {*67*, *Y*} and {*82*, *Y*} were 78, 80, 76 and 78%, respectively (Figure 1b). The powers of detecting the second-order interactions {*7*, *22*, *Y*} and {*67*, *82*, *Y*} were 43 and 42%, respectively—the decrease relative to the power of detecting first-order combinations can be attributed to GH. The FCM observed was 1.9 and the average model size was 5.9. Thus, AMBROSIA has appreciable power and low error rate in the presence of LD as well as GH.

Comparisons with MDR

Comparisons on case study 2: Because it was difficult to complete MDR analysis with 120 SNPs from case study 2, we created a smaller data set containing 24 SNPs that consisted of the central SNP, and one SNP with $R^2 = 0.9$ and one SNP with $R^2 = 0.8$ from each of the eight groups (Supplementary Figure 2D).

Table 3 shows the power of AMBROSIA compared with MDR. The presence of {*1*, *Y*}, {*4*, *Y*}, {*13*, *Y*} and {*16*, *Y*} in the AMBROSIA model represents the main effects of the individual SNPs, whereas the second-order combinations {*1*, *4*, *Y*}, and {*13*, *16*, *Y*} represent the GGI in Supplementary Figure 2D. AMBROSIA has higher power than MDR in detecting each of the four first-order interactions and both second-order interactions.

The power of MDR to detect the two-SNP combinations, {*1*, *4*, *Y*} and {*13*, *16*, *Y*} was greater compared to its power to detect the one-way combinations {*1*, *Y*}, {*4*, *Y*}, {*13*, *Y*} and {*16*, Y}. The power of MDR to detect the four-way combination or {*1*, *4*, *13*, *16*, *Y*} was 39%. It is important to note that in the context of the underlying model, combinations such as {*1*, *13*, *Y*} or {*1*, *4*, *13*, *16*, *Y*} that contain a mixture of SNPs from interaction 1 and interaction 2 are false combinations in our simulation model.

Table 3 Comparison of the power of AMBROSIA to MDR for case study 2

Combinations	*AMBROSIA*	*MDR*
{*1*, *Y*}	89	23
{*4*, *Y*}	90	17
{*1*, *4*, *Y*}	55	36
{*13*, *Y*}	89	22
{*16*, *Y*}	90	18
{*13*, *16*, *Y*}	57	34
{*1*, *4*, *13*, *16*, *Y*}	—	39

Abbreviation: MDR, multifactor dimensionality reduction.
The percentage of replicates in which the two methods correctly identified the combinations involved in the gene–gene interactions is shown.

The FCM of AMBROSIA (mean ± s.d. = 0.95 ± 1.53) was lower than MDR (mean ± s.d. = 2.24 ± 0.91) when fourth-order combinations were considered for MDR (fourth-order MDR FCM). When combinations up to second order were considered, the FCM for MDR (second-order MDR FCM) was 0.63 ± 0.72. AMBROSIA, however, does not require evaluation of combinations of order greater than two to model the interacting loci.

This example suggests that AMBROSIA can identify interacting variables at a lower interaction order consistent with the underlying simulation framework than MDR. This is advantageous because the number of combinations to be searched increases rapidly with interaction order.

Comparisons on models based on the MDR power paper: The power, frequency of FCM for models 1-GH, 2-GH, 3-GH and 4-GH are summarized in Table 4. AMBROSIA consistently achieved better power and lower FCM compared with MDR for each two-SNP interaction in these models. Further, MDR can only detect all four loci with improved power when the larger four-SNP combination {*1*, *4*, *13*, *16*, *Y*}, which does not distinguish the two sets of pairwise interactons, is considered. Even upon including the larger four-SNP combination for MDR, the power of AMBROSIA was equivalent to or better than the power of MDR.

Table 4 summarizes the second-order MDR FCM, which is associated with lower power, and also the fourth-order MDR FCM, which enables MDR to identify the larger four-SNP combination {*1*, *4*, *13*, *16*, *Y*}. The FCM of AMBROSIA compares favorably with the second-order MDR FCM for models 1-GH and 2-GH; for models 3-GH and 4-GH, the FCM of AMBROSIA was higher than the second-order MDR FCM. However, MDR had approximately 8.8 to 15-fold lower power to detect the interacting combinations than AMBROSIA for models 3-GH and 4-GH when interactions up to second-order were considered.

Robustness of AMBROSIA to LD patterns and allele frequency

We evaluated power of AMBROSIA in the context of the GAW15-P2 data, which contains allele frequencies and LD patterns representative of real data. The distribution of LD values of a subset of SNPs from a representative simulation (Figure 2a) demonstrates that a wide range of LD values are included; the R^2 values in Figure 2a ranged from 0 to 0.89. Figure 2b is a histogram showing the MAF distribution, and again a wide range of allele frequencies were included.

AMBROSIA had a power of 98–100 for allele frequencies of 0.2 or greater (Figure 2c). The proportion of FCM ranged from 1.6 at MAF = 0.1 to 1.2 at MAF = 0.5. At MAF = 0.5, the power increased rapidly with increase in relative risk (Figure 2d). The first-order {*i*, *Y*} combinations had power greater than 90% at a relative risk of 1.5 or greater, whereas second-order {*i*, *j*, *Y*} combinations had power approaching 80%. These analyses highlight the model-synthesis capabilities and robustness of

Table 4 Comparison of AMBROSIA to MDR for models based on the penetrance matrices in Table 1

Model	*Power (%)*[a]					*FCM*[b]		
	{*1*, *4*, *Y*}		{*13*, *16*, *Y*}		{*1*,*4*,*13*,*16*,*Y*}	*AMBIENCE*	*MDR*	
	AMBROSIA	*MDR*	*AMBROSIA*	*MDR*	*MDR*		*Second-order*	*Fourth-order*
1-GH	100	8	100	8	14	0.0 ± 0.0	0.0 ± 0.0	0.07 ± 0.26
2-GH	100	38	100	39	100	0.0 ± 0.0	0.0 ± 0.0	0.66 ± 0.47
3-GH	97	11	97	10	56	0.07 ± 0.26	0.02 ± 0.14	0.21 ± 0.5
4-GH	90	5	91	6	82	0.21 ± 0.43	0.03 ± 0.17	0.24 ± 0.49

Abbreviations: FCM, false combinations per model; GH, genetic heterogeneity; MDR, multifactor dimensionality reduction.
The penetrance matrix values are based on the models in Ritchie *et al*. (2003), which systematically evaluated the power of MDR.
[a]Power was computed for $\alpha = 0.001$.
[b]Combinations up to either second-order or fourth-order (last column) were considered for MDR as AMBROSIA required only second-order combinations and MDR required fourth-order combinations to detect {*1*, *4*, *13*, *16*, *Y*} with power >80%.

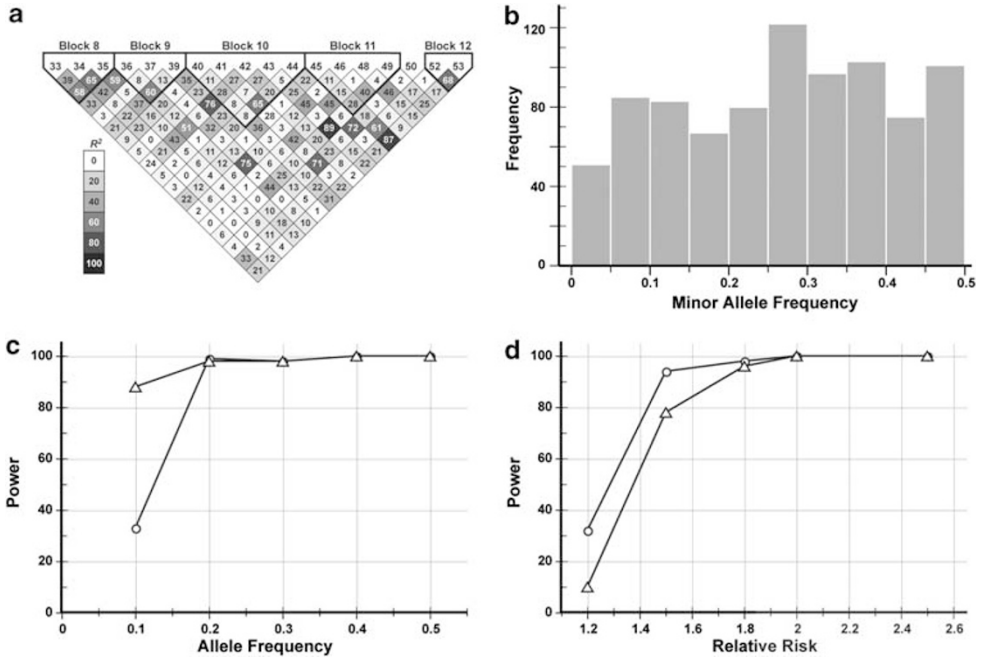

Figure 2 Results from robustness assessment. (**a**) Distribution of LD values for a representative set of haploptype blocks of a simulation from the GAW15-P2 data set. Darker shades indicate higher levels of LD (as measured by R^2, the actual R^2 values are also shown; the palette on the right provides linear gray scale used) and the lighter shades indicate lower levels of LD. (**b**) Distribution of minor allele frequencies (bin width of 0.05) in the data set. (**c**, **d**) Dependence of power in the robustness assessment experiment on allele frequency and relative risk, respectively. In (**c**) and (**d**), the open circles represent the {*i*, *Y*} combination and the open triangles represent the combination {*i*, *j*, *Y*}. The results for the {*j*, *Y*} combination were nearly identical to the {*i*, *Y*} combination and are not shown.

AMBROSIA in the face of the confounding factors such as LD in real data sets.

AMBROSIA analysis of GAW15-P3 data set

We analyzed *RA* as the phenotype for the GAW15-P3 data set. The availability of 100 replicates from repetitions of the simulation procedure (Miller *et al*., 2007) enabled us to critically assess AMBROSIA performance.

The frequencies of the combinations identified in AMBROSIA models for the 100 replicates are summarized in Table 5. The average model size across the 100 replicates was 9.54 combinations per model. The combinations {*C6_153*, *Age*, *RA*}, {*Sex*, *RA*}, {*Age*, *RA*}, {*C11_389*, *RA*}, {*Smoking*, *RA*} and {*C6_153*, *RA*} occurred in all 100 replicates, whereas the combinations {*C6_162*, *RA*}, {*C11_389*, *Age*, *RA*} and {*C6_162*, *Age*, *RA*} occurred in greater than 90% of replicates. The proportion of replicates in which the individual combinations were identified can be interpreted as the power to detect the combinations when only a single data set is available.

Notably, only a single representative from each locus spanning multiple SNPs was present in the AMBROSIA models; for example, locus DR spanned SNPs 152—155, but only SNP *C6_153* was selected. The detection of SNP *C6_162*, which represents locus D and has a rare allele that increases RA risk fivefold, suggests that AMBROSIA is capable of identifying contributions of alleles with low allele frequencies. There were single instances of two combinations {*C6_139*, *Age*, *RA*} and {*C6_139*, *RA*} that contained SNP *C6_139*, which was not present in Supplementary Table 1.

Table 5 Results for GAW15-P3 data set. Each of the 100 replicates in the GAW15-P3 data set was modeled separately

Combinations	*Number of replicates with combinations present in model*
C6_153, *Age*, *RA*	100
Sex, *RA*	100
Age, *RA*	100
C*11_389*, *RA*	100
Smoking, *RA*	100
C6_153, *RA*	100
C6_162, *RA*	99
C11_389, Age, *RA*	96
C6_162, *Age*, *RA*	92
C6_162, *Sex*, *RA*	29
Age, *Smoking*, *RA*	17
C11_389, *Sex*, *RA*	7
Sex, *Smoking*, *RA*	5
C6_162, C*11_389*, *RA*	3
C6_153, *Smoking*, *RA*	2
C6_162, *Smoking*, *RA*	1
C6_139, *Age*, *RA*	1
C6_139, *RA*	1

Abbreviations: GAW15, Genetic Association Workshop 15; *RA*, rheumatoid arthritis.
The first column shows the combinations included in the models and the number of GAW15-P3 replicates in which the combinations occurred.

Computation time and scaling characteristics

The computation time and scaling characteristics depend on three sequential analysis steps. The first step involves an algorithm, for example, AMBIENCE, to identify the most promising combinations for modeling. The second step involves permutation-based significance assessment and the third step involves AMBROSIA.

Time complexity of AMBIENCE: Let n denote the sample size and m denote the number of predictor variables (excluding the phenotype variable). The runtime complexity of AMBIENCE is given by $O(\tau \times \theta \times n \times m^2) + \times \theta \times 2^{\tau} \times m^2)$, where τ is the maximum combination size to explore and θ is the number of combinations to retain in each step of the search procedure (Chanda *et al*., 2008). In genetic applications, the range of τ-values of interest is small because of sample size constraints and limits computational complexity.

Time complexity of permutations: The permutation procedure, if implemented naively, can be very time consuming because the KWII of each combination has to be repeatedly computed. However, the permutation can be implemented in a very efficient manner for discrete data because the sufficient statistics for computing the KWII of a combination are present in the corresponding contingency table. All permutations correspond to a change in cell counts of the contingency table subject to the constraint that the row sums and column sums are unchanged. Only one scan of the data is necessary for building the contingency table for each combination (Patefield, 1981).

The creation of the contingency table T for a combination C with k variables and b states has $O(m \times b)$ complexity. The KWII(C) computation requires $O(m \times b + 2^k \times b) = O(m \times b)$ computations (for $m >> 2^k$) because entropies of all subsets of k variables are computable by marginalizing T.

The first KWII computation involves $O(m \times b)$ computations because T is constructed. For each successive permutation, we randomly vary counts in T using an efficient algorithm (Patefield, 1981) that requires $O(b)$ computations and scales linearly with the number of permutations. For N_{PERM} permutations, the time complexity for the permutation step is $N_{PERM} \times O(b)$.

Time Complexity of AMBROSIA: Let a fraction, α, of the $\tau\theta$ combinations from AMBIENCE emerge as significant after permutation testing. Model building in AMBROSIA therefore proceeds with $\psi = \alpha\tau\theta$ combinations.

We consider the worst-case scenario wherein all ψ combinations are nonredundant and pass the parsimony test. In this scenario, $O(\psi^2)$ redundancy check tests and parsimony tests are necessary. Each parsimony test consumes a constant time, whereas each redundancy check is a KWII computation requiring $O(m^2)$ computations. Thus the worst-case computation costs involved in the AMBROSIA-modeling step is $O(\psi^2\ m^2)$. Typically, only a few SNPs and predictor variables are involved in interactions in the majority of gene–environmental interaction analyses; as a result, the fraction α of significant combinations is the key factor that makes $O(\psi^2\ m^2)$ small.

Overall, the runtime costs of AMBIENCE dominate over the time required for permutations and AMBROSIA. Our experiments indicate that AMBIENCE requires about 80–90% of the total time incurred.

Discussion

In this paper, we have described AMBROSIA, a model synthesis method for GEI analysis. Our approach is distinctive in the information-theoretic method used, its versatility, generalizability and scalability. In AMBROSIA, we cogently enhance relevancy of included combinations, preclude entry of combinations with redundant information, and also penalize for unjustifiable complexity in the model. The AMBROSIA model synthesis paradigm differs substantively from other GEI analysis methods in numerous respects, including the GEI identification metrics, the synthesis procedure, and the model evaluation steps.

MDR is widely used for GEI analysis and was therefore selected for head-to-head comparisons. MDR conducts exhaustive search of genotype combination space, which despite availability of a parallelized version of MDR, limits computational efficiency. MDR has limited power in the presence of GH (Ritchie *et al*., 2003) because it constructs models from the best one-way, two-way and three-way combinations. AMBROSIA does not constrain the number of combinations, combination order and structure—it evaluates the redundancy and parsimony of an ordered list of KWII values. MDR is also limited to analysis of binary phenotypes. AMBROSIA can be deployed for binary, discrete, continuous and mixed distributions for which KWII calculations are possible. GMDR is an MDR extension that uses the general linear model in conjunction with dimensionality reduction.

Logistic regression and logic regression are examples of regression-based approaches. In logistic regression, unaccounted variance is reduced because of the net result of all terms in the model. Logistic regression requires explicit model specification, repeated re-fitting

and is adversely affected by multi-collinearity with SNPs that are in high LD (Briollais *et al*., 2007). The KWII, in contrast, does not contain information regarding lower order terms, and the AMBROSIA model is assembled with non-overlapping terms. There is also no iterative re-fitting of the data. In logic regression, the Boolean outputs of logic trees constructed with genotypes are used as predictors in a regression framework (Kooperberg and Ruczinski, 2005). In our studies, logic regression had low power and high type I error; its type I error was on average two-to-three orders of magnitude greater than KWII, MDR and logistic regression (unpublished results).

Here, we have used the output of AMBIENCE as input for AMBROSIA. This requirement is not stringent, and AMBROSIA can be compatible with other GEI-identification approaches. The KWII for these inputs have to be computed in AMBROSIA. Although AMBROSIA is currently implemented to return a single 'best' model as output, it can be modified to provide a ranked ensemble of alternative models. The alternative models can be built by utilizing the combinations that are eliminated at the redundancy step. The alternative models can be ranked and compared using Akaike (or Bayesian Information Criterion) weights.

Our results indicate that AMBROSIA is a promising approach for GEI modeling in pharmacogenetics.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Support from the National Multiple Sclerosis Society (RG3743), the Department of Defense Multiple Sclerosis Program (MS090122), and the Center for Protein Therapeutics is gratefully acknowledged.

Licence for publication: The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors. *Confidentiality*: Use of the information in this manuscript 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. *Funding*: Support from the National Multiple Sclerosis Society (RG3743) and the Department of Defense (MS090122) is also gratefully acknowledged.

References

Ambrosone CB, Shields PG, Freudenheim JL, Hong CC (2007). Re: Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. *J Natl Cancer Inst* **99**: 487 (author reply 488–489).

Briollais L, Wang Y, Rajendram I, Onay V, Shi E, Knight J *et al*. (2007). Methodological issues in detecting gene-gene interactions in breast cancer susceptibility: a population-based study in Ontario. *BMC Med* **5**: 22.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**: 106–120.

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 environment interactions associated with complex phenotypes. *Genetics* **180**: 1191–1210.

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.

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

Goldstein DB (2009). Common genetic variation and human traits. *N Engl J Med* **360**: 1696–1698.

Han TS (1980). Multiple mutual informations and multiple interactions in frequency data. *Inf Control* **46**: 26–45.

Hurvich CM, Tsai CL (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51**: 1077–1084.

Jakulin A (2005). Machine learning based on attribute interactions. Ph.D. 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.

Kooperberg C, Ruczinski I (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* **28**: 157–170.

McGill WJ (1954). Multivariate information transmission. *Psychometrika* **19**: 97–116.

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 Genetics*. **1**(Suppl 1): S4.

Patefield WM (1981). Algorithm AS 159: an efficient method of generating random R x C tables with given row and column totals. *J R Stat Soc C (Appl Stat)* **30**: 91–97.

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.

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.

Weiss KM, Terwilliger JD (2000). How many diseases does it take to map a gene with SNPs? *Nat Genet* **26** (2): 151–157.

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