American Journal of Biostatistics 2 (1): 1-10, 2011
ISSN 1948-9889

A Bayesian Method for Disentangling Dependent Structure of Epistatic Interaction

[1]Jing Zhang, [2]Qu Zhang, [3]Darrin Lewis and [4,5]Michael Q. Zhang
[1]Department of Statistics, Yale University,
24 Hillhouse Ave., New Haven, CT 06511, USA
[2]Department of Human Evolutionary Biology,
11 Divinity Avenue, Harvard University, Cambridge, MA 02138
[3]Cold Spring Harbor Laboratory, 1 Bungtown Road,
Cold Spring Harbor, New York, 11724, USA
[4]MCB and Center for Systems Biology, The University of Texas at Dallas,
800 West Campbell Road, RL11, Richardson, Dallas, 75080, USA
[5]NTLIST, Tsinghua University, Beijing, 100084, China

Abstract: Problem statement: We propose a Bayesian method (RBP) to recursively infer the independence structure of epistatic interactions in case-control study. **Approach:** Based on the results of BEAM2, RBP can powerfully detect the marginal and conditional independence within interacting SNPs even in the complicated interaction cases. **Results:** We did extensive simulations to test RBP and compare it with stepwise logistic regression. Simulation results show that this approach is more powerful than stepwise logistic regression in detecting in marginal independence and conditional independence as well as more complicated dependence structure. We then applied BEAM2 and RBP on dbMHC Type 1 Diabetes (T1D) data and we found in MHC region, genes DRB1 and DQB1 are associated with T1D with saturated interaction structure which is consistent with the current knowledge of haplotype effect of these two genes on T1D. **Conclusion:** RBP is a powerful method to infer detailed dependence structures in epistatic interactions.

Key words: Type 1 Diabetes (T1D), Recursive Bayesian Partition (RBP), Genome-Wide Association (GWA), Independence Partition Model (IPM), Chain-Dependence Model (CDM)

INTRODUCTION

Recently the methodology of Genome-Wide Association (GWA) has been greatly improved (WTCCC, 2007a; Zhang and Liu, 2007; Zhang *et al*., 2011; Yang *et al*., 2009). A Bayesian method (BEAM, Zhang and Liu, 2007) equipped with Monte Carlo algorithms has been shown able to powerfully detect high-order interactions in genome-wide association studies. This method uses Markov chain Monte Carlo (MCMC) to 'interrogate' each marker conditional on the current status of other markers iteratively. But one drawback of BEAM is the assumption that SNPs are independent to each other, thus BEAM is not able to capture the block-wise structure of human genome. Zhang *et al*. (2011) extended BEAM model to BEAM2, incorporating LD blocks into the original Bayesian partition model. This BEAM2 is able to simultaneously infer haplotype-blocks and select SNPs within blocks that are associated with the disease, either individually or through epistatic interactions with others across the genome. Using WTCCC1 type 1 diabetes data, BEAM2 identified the most previous reported associated SNPs and also landscaped the two-way interactions in MHC region.

Under the concept of common (complex) diseases, genetic variants typically show very little effect independently with low penetrance, but they may interact with each other in complex ways, i.e., they have complicated interacting structure, which is probably because of the sophisticated regulatory mechanisms encoded in the human genome (WTCCC, 2007a; Chambers and Hastie, 1992; Yang *et al*., 2009; WTCCC, 2007b; Zhang *et al*., 2011; Ding *et al*., 2005a; 2005b). Both BEAM2 and BEAM2 use a saturated model to model the interaction group, thus neither of them can infer the detail interacting structure. However, knowing the detail structure is very important for

Corresponding Author: Jing Zhang, Department of Statistics, Yale University, 24 Hillhouse Ave., New Haven, CT 06511, USA

investigating the etiopathogenesis and genetic mechanisms of the complex disease.

In this study, we propose a Bayesian method called Recursive Bayesian Partition (RBP) to recursively infer the marginal and conditional independence among interacting genetic markers. Given the associated markers inferred by BEAM2 or BEAM2, RBP first uses Independence Partition Model (IPM) to recursively infer all the marginally independent interaction groups, such that there is no interaction between/across different groups. That is, only within each group, there are some interactions. Then, RBP uses Chain-Dependence Model (CDM) to recursively infer the conditional independence within each interaction group (Zhang *et al*., 2010).

Using simulations we showed our method is more powerful than stepwise logistic regression using AIC or BIC in both marginal independence and conditional independence detections. In the complicated interaction simulations, our method is much more powerful than stepwise logistic regression. We applied RBP to the dbMHC type 1 Diabetes (T1D) data and we found genes DRB1 and DQB1 are associated with T1D with saturated interaction structure which is consistent with the current knowledge of haplotype effect of these two genes on T1D (Steenkiste *et al*., 2007).

MATERIALS AND METHODS

Recursive Bayesian Partition (RBP) we propose a Bayesian Partition model to search for independence groups and conditional independence among interacting SNPs. The whole procedure is done in two steps: first, we use Independence Partition Model (IPM) to partition all the interacting SNPs into several independence groups, i.e., there is no interactions across groups; thenwe use Chain-Dependence Model (CDM) to search for conditional independence within each groups.

Suppose there are N_d sequences in the case group and N_u sequences in the control group. Each sequence is p-SNP long and each SNP position i can take L_i possible categories more generally, we can view each position as a random variable. Thus, our data consists of observations on each individual status (or response) variable Y, i.e., 0 for control and 1 for case and its p "explanatory" variables $X_1,\ldots, X_p$.

Independence Partition Model (IPM): Our Bayesian Independence variable partition model seeks to partition the p variables into two groups: A and B. We say that the joint distribution for X_G is a Independence Partition model if the index set G can be partitioned into disjoint subsets A and B, such that X_A and X_B are mutually independent, i.e., $P(X_G) = P(X_A)P(X_B)$. We let Π denote the partitioning, i.e., indicating which indices in G belong to which subset.

We let D denote n iid observations ($n = N_d$ in cases and $n = N_u$ in controls) on $(X_G) = (X_A, X_B)$. Suppose X_A takes on A_N possible values, follows the distribution Multinom (Θ_A), where $\Theta_A = (\theta_1^A, \ldots \theta_{N_A}^A)$. The prior for Θ_A is assumed to be the Dirichlet distribution with pseudo-counts vector $\beta^A = (\beta_1^A, \ldots. \beta_{NA}^A)$ denoted as Dir (β^A). In this study we let $\beta_k^A = 1/N_A$. Then let $n^A = (n_1^A, \ldots n_{N_A}^A)$, where n_k^A is the number observations whose X_A takes the kth categorical value. Integrating out the multinomial parameters, we obtain that Eq. 1:

$$p(D_A \mid \prod) = \frac{\Gamma\left(n^A + \beta^A\right)}{\Gamma\left(|n^A| + \beta^A|\right)} \frac{\Gamma\left(|\beta^A|\right)}{\Gamma\left(\beta^A\right)} \equiv$$

$$\left(\prod_{K=1}^{N_A} \frac{\Gamma\left(n_K^A + \beta_k^A\right)}{\Gamma\left(\beta_k^A\right)}\right) \frac{\Gamma\left(\sum_{K=1}^{N_A} \beta_K^A\right)}{\Gamma\left(n + \sum_{K=1}^{N_A} \beta_K^A\right)}, \tag{1}$$

where, we define $\Gamma(v) = \Gamma(v_1) \ldots \Gamma(v_k)$ for $v = (v_1,\ldots,v_k)$. The computation for $p(D_B|\Pi)$ is exactly the same as that for $p(D_A|\Pi)$.

Let N_B be the number of possible values X_B can take and let $X_B \sim$ Multinom (Θ_B), with $\Theta_B = (\theta_1^B, \ldots\ldots, \theta_{NB}^B)$. A Dirichlet ($\beta^B$) prior distribution is imposed on Θ_B, where $\beta^B = (\beta_1^B, \ldots.., \beta_{NB}^B)$. Let n_k^B be the number of times X_B takes on value k in our observations D_B. Thus, we have Eq. 2:

$$p(D_B \mid \prod) = \frac{\Gamma\left(n^A + \beta^A\right)}{\Gamma\left(n \ \ | + \beta^A|\right)} \frac{\Gamma\left(|\beta^B|\right)}{\Gamma\left(\beta^B\right)} \tag{2}$$

Bayesian recursive model selection (Zhang *et al*., 2010): In order to determine to include an independence model for control, we define a model indicator I_C, which is equal to 1 if the variables $\{X_{1,\ldots,} X_P\}$ in controls have the same group membership as in cases and 0 if the variables in controls are all independent of each other, in which case we have:

$$P(\{X_{1,\ldots,} X_P\} \mid I_C = 0, Y = 0) = \{\prod P(X_j \mid I_C = 0, Y = 0)\}$$

where, $p(X_j|I_c = 0, y = 0)$ is multinomial with probability vector Θ_j, Multinom(Θ_j) and Θ_j follows a Dirichlet distribution a priori with parameter $a_j = (a_{j1}, a_{j2}, \ldots, a_{jGj})$. Integrating out the multinomial parameters, we obtain that Eq. 3:

$$P(X_{1,\cdots,}X_P \mid I_C = 0, Y = 0) = \prod_{J=1}^{P}\left\{\left\{\prod_{K=1}^{G_J}\frac{\Gamma(n_{jk}+\alpha_{jk})}{\Gamma\alpha_{jk}}\right\}\frac{\Gamma(|\alpha_j|)}{\Gamma(N_u+|\alpha_j|)}\right\} \qquad (3)$$

Here the operation |a| sums over all elements in a.

We assume an equal probability prior for I_C. Then, we can write the joint distribution as (D here includes both case data and control data) Eq. 4:

$$P(D \mid \prod, I_C, Y)p(\prod\ )p(I_c)P(Y) \qquad (4)$$

In which p (D|Π,l$_c$ = 1,Y = 0), p (D|Π,l$_c$ = 0,Y = 1) and p (D|Π,l$_c$ = 1,Y = 1)are equal to expression (1) times (2) and p (D|Π,l$_c$ = 0,Y = 1) is equal to (3) . We use a MCMC algorithm to simulate from (4) so as to estimate the posterior distribution of I_C and Π. After partitioning all the variables into A or B, we reapply the method to A and B superlatively. The procedure is applied recursively until only single-variable nodes are available and thus all the variables are grouped in several independent disjoint subsets.

Chain-Dependence Model (CDM): After identifying all the independence groups, we use Chain-dependence Model to discover the conditional independence within each group. In the above IPM model, variables in A or B are not imposed any simplifying dependence structure, which in statistical term means that a "fully saturated" model was used. However, in practice often a much more desirable and simpler model, which takes advantage of conditional independence relationships among the variables, can fit the data well. In complex disease scenario, SNP1 could interact with SNP2, but does not directly interact with SNP3. SNP2 interacts with SNP3. Thus all the three SNPs are in the same independent group (i.e., IPM cannot separate them), but conditional on SNP2, SNP1 is conditionally independent of SNP3. Therefore, detecting conditional independence within each group can tell more detail about the relationship within each independence group.

We call it a chain-dependence model for a group of variables X_G if the index set G can be partitioned into 3 subgroups A, B and C such that X_A and X_C are independent given X_B, denoted as $X_A \rightarrow X_B \rightarrow X_C$. Only set C is allowed to be empty, in which case this model degenerates to the saturated model. Under the chain-dependence model, we can decompose the joint distribution of XG as: p (X_G) = p (X_A) p (X_A) p (X_B | X_A) p (X_C |X_B). We let Π denote the set partitioning.

Suppose X_A takes on N_A possible values, follows the distribution Multinom (Θ_A), where, $\Theta_A = (\theta_1^A, ..., \theta_{N_A}^A)$. The prior for Θ_A is assumed to be the Dirichlet distribution with pseudo-counts vector $\beta^A = (\beta_1^A, ..., \beta_{N_A}^A)$ denoted as Dir (β^A). Here we let $\beta_k^A = 1/N_A$.

Furthermore, suppose X_B takes on N_B possible values, we assume that P (X_B|X_A) is Multinom ($\Theta_{B|X^A}$), where, $\Theta_{B|X_A} = (\theta_{1,X_A}^B, \ldots, \theta_{N_B,X_A}^B)$. We let the prior distribution for ($\Theta_{B|X^A}$) be $\text{Dir}\left(\{\beta_{i|X_A}^{B|A}\}_{i=1,...,N_B}\right)$. We set $\beta_{i/X_A}^{B/A} = 1/(N_A N_B)$ in the study. Lastly, suppose X_C take on N_C possible values. Then, P (X_C|X_B) = Multinom (Θc|X_B) and $\Theta_{C|X_B} \sim \text{Dir}(\{\beta_{i,X_B}^{C|B}\}_{i=1,\ldots,N_C})$ a priori. Again, we set $\beta_{i,X_B}^{C|B} = 1/(N_B N_C)$. Suppose our data D consists of n iid observations on X_G = (X_A, X_B, X_C). We can summarize the data into counts corresponding to X_A, X_B| X_A and | X_C X_B and decompose the probability of the data conditional on the set partitioning Π as P (D|Π) = P (D_A|Π) p (D_B| D_A, Π) p (D_C | D_B, Π) accordingly. That is, we let $n^A = (n_1^A, ..., n_{N_A}^A)$ where n_k^A is the number observations whose X $_A$ takes the kth categorical value. Integrating out the multinomial parameters, we obtain that Eq. 5:

$$p(D_A \mid \prod\ ) = \frac{\Gamma\left(n^A+\beta^A\right)}{\Gamma\left(|n^A|+\beta^A|\right)}\frac{\Gamma\left(|\beta^A|\right)}{\Gamma\left(\beta^A\right)} \equiv$$

$$\left(\prod_{K=1}^{N_A}\frac{\Gamma\left(n_K^A+\beta_k^A\right)}{\Gamma\left(\beta_k^A\right)}\right)\frac{\Gamma\left(\sum_{K=1}^{N_A}\beta_K^A\right)}{\Gamma\left(n+\sum_{K=1}^{N_A}\beta_K^A\right)}, \qquad (5)$$

where, we define Γ (v) = Γ (v_1) … Γ (v_k) for v = (v_1,…,v_k). Similarly, we obtain that Eq. 6:

$$P(D_B \mid D_A, \prod\ ) = \prod_{J=1}^{N_A}\left[\frac{\Gamma(n_{|J}^{B|A}+\beta_{|J}^{B|A})}{\Gamma(n_J^A+|\beta_{|J}^{B|A}|)}\frac{\Gamma(|\beta_{|J}^{B|A}|)}{\Gamma(\beta_{|J}^{B|A})}\right] \qquad (6)$$

where, $n_{\cdot|i}^{B|A} = (n_{1|i}^{B|A}, \ldots, n_{N_B|i}^{B|A})$ with $n_{j|i}^{B|A}$ recording the number of observations in which X_A = I and X_B = j Thus, $|n_{\cdot|i}^{B|A}| = n_{1|i}^{B|A} + \cdots + n_{N_B|i}^{B|A} = n_i^A$. Finally, we get Eq. 7:

$$P(D_C \mid D_B, \Pi) = \prod_{j=1}^{N_B}\left[\frac{\Gamma\left(n_{\cdot|j}^{C|B}+\beta_{\cdot|j}^{C|B}\right)}{\Gamma\left(n_j^B+\left|\beta_{\cdot|j}^{C|B}\right|\right)}\frac{\Gamma\left(\left|\beta_{\cdot|i}^{C|B}\right|\right)}{\Gamma\left(\beta_{\cdot|j}^{C|B}\right)}\right] \qquad (7)$$

Thus, the probability of the observed data conditional on the set partitioning Π is the product of (5-7) Eq. 8:

$$P(D_A, D_B, D_C \mid \prod\ ) = P(D_A \mid \prod\ )P(D_B$$
$$\mid D_A, \prod\ )P(D_C \mid D_B, \prod\ ) \qquad (8)$$

Table 1: Interaction Models

Risk	A/A	A/a	a/a
Model 1			
B/B	1	1	1
B/b	1	$(1+q)^2$	$(1+q)^3$
b/b	1	$(1+q)^3$	$(1+q)^4$
Model 2			
Risk	A/A	A/a	a/a
B/B	1	1	1
B/b	1	1+q	1+q
b/b	1	1+q	1+q
Model 3			
Risk	A/A	A/a	a/a
B/B	1	1	1
B/b	1	$\left(1-\frac{p_1p_2}{4q_1q_2}(1-\theta)\right)$	$\left(1+\frac{p_2}{2q_2}(1-\theta)\right)$
b/b	1	$\left(1-\frac{p_1p_2}{4q_1q_2}(1-\theta)\right)$	q

Two-locus interaction models with no marginal effect $p_1 = p(A) = 1 - q_1$ and $p_2 = p(B) = 1-q_2$

Thus, if we assign a prior on Π, such as uniform, we obtain its posterior distribution by combining the prior with (8). We then design an MCMC algorithm to sample from this posterior distribution.

Like IPM, we first partition variables into three disjoint subsets which follow the above chain-dependence model and then recursively apply this partition until the data do not allow us to discern more model details.

Simulations from HapMap data: We did three sets of simulations to evaluate the powers of IPM, CDM and RBP and compared them with stepwise logistic regressions by AIC and BIC. First, we simulated two independent interaction groups Fig. 7a, conditional independent group Fig. 7b for IPM and CDM respectively. And then we simulated complicated interactions which contains 9 independent groups: 6 of them are composed of two-locus interactions Fig. 7a following three different disease models in Table 1-3 of them are composed of three-loci conditional independent interactions Fig. 7b with pairwise interactions (like interaction between D1 and D2, interaction between D2 and D3 in Fig. 5b following three different disease models in Table 1. For details see the following sections.

Three different disease interaction models: We simulated case control data according to 3 disease models given in Table 1. There are 2 disease loci involved in each model. In model 1, the disease risk is present only when both disease loci contain some mutations and the disease risk increases as the number of mutations increases. In model 2, the disease risk is again present only when both disease loci contain some mutations, but the risk is a constant corresponding to a threshold model. In model 3, there is no marginal effect but only interactions for two loci.

Marginal independent groups simulation for IPM: For marginal independent groups A and B, (Group A and Group B are marginally independent in both populations (Y =1case and Y = 0 control)), the odds can be written as:

$$w_{AB} = \frac{P(Y=1|AB)}{P(Y=0|AB)} = \frac{P(Y=1)}{P(Y=0)}\frac{P(AB|Y=1)}{P(AB|Y=0)} = \frac{P(Y=1)}{P(Y=0)}\frac{P(A|Y=1)}{P(A|Y=0)}\frac{P(B|Y=1)}{P(B|Y=0)}$$

The odds ratio of a mutant type AB to wildtype A_0 B_0 is:

$$\frac{W_{AB}}{W_{A0B0}} = \frac{\frac{p(Y=1)}{p(Y=0)}\frac{P(A|Y=1)}{P(A|Y=0)}\frac{P(B|Y=1)}{P(B|Y=0)}}{\frac{p(Y=1)}{p(Y=0)}\frac{P(A_0|Y=1)}{P(A_0|Y=0)}\frac{P(B_0|Y=1)}{P(B_0|Y=0)}} =$$

$$\frac{\frac{P(Y=1|A)}{P(Y=0|A)}\frac{P(Y=1|B)}{P(Y=0|B)}}{\frac{P(Y=1|A_0)}{P(Y=0|A_0)}\frac{P(Y=1|B_0)}{P(Y=0|B_0)}}$$

$$= \frac{W_A W_B}{W_{A0} W_{B0}}$$

Which is the product of odds ratios of two groups. After taking log arithm, $\log\frac{w_{AB}}{wA_0B_0} = \log\frac{w_A}{w_{A0}} + \log\frac{w_B}{w_{B0}}$ There are two SNPs in group A and B following Model 1, 2 or 3 Table 1.

Conditional independent group simulation for CDM: For conditional independence, A and C are conditionally independent given B, the odds ratio for three groups A, B and C, is $\frac{w_{ABC}}{w_{A_0B_0C_0}} = \frac{w_B}{w_{B_0}}\frac{w_{AB}}{w_{A_0B_0}}\frac{w_{BC}}{w_{B_0C_0}}$ So logarithm of odds ratio is $\log\frac{w_{ABC}}{w_{A_0B_0C_0}} = \log\frac{w_B}{w_{B_0}} + \log\frac{w_{AB}}{w_{A_0B_0}} + \log\frac{w_{BC}}{w_{B_0C_0}}$.

There is one SNP in group A, B and C, the interaction model of AB and BC follows Model1, 2 or 3 Table 1.

Complicated interaction simulation for RBP: Finally we simulated a complicated interaction data which contains multiple marginal independent groups and within some groups there are conditional independence interactions. The logarithm of odds ratio is the summation of all the marginal independence groups and conditional independence groups. In our simulation there are totally 21 SNPs in 6 two-locus marginal independence groups and 3 three-locus conditional independence groups.

Simulations with block structures and interactions: To simulate case control data, we first randomly select a region in the human genome that contains 1000 tagged SNPs in Illumina HapMap 300k tagSNPs. We then use HAPGEN (Marchini *et al*., 2007) to generate a pool of 10000 individuals and their genotypes within the region for tagged SNPs using HapMap European individuals (parents only). Four SNPs are then randomly selected as disease loci for marginal independence simulations and three SNPs are selected for conditional independence simulations. We choose the disease allele frequency in the population to be 0.05, 0.10, 0.20 and 0.50, respectively and we specify the marginal effect size (log odds ratio-1) of each disease loci to be 0.5 for Model 1 and 2 ($\theta = 0.5$ in Table 1) and for Model 3 there is always no marginal effect, so we set interaction effect (Yang *et al*., 2009) parameter $\theta = 0.5$. Given these parameters, we calculate the joint allele frequencies of both disease loci in cases and controls using the same method used in BEAM (Zhang and Liu, 2007) and BEAM2 (Zhang *et al*., 2011). We then randomly sample 2000 cases and 2000 controls according to the disease allele frequencies from the pool of 10000 individuals without replacement.

In the same way, we simulated a complicated interaction case-control data, where 21 SNPs are selected for 9 independent groups. Totally 6 groups follow two-locus interaction models (2 groups for each of Model 1, 2 and 3 in Table 1). 3 groups arecomposed of three-locus conditional independent interactions (one group for each of Model 1, 2 and 3 in Table 1).

RESULTS AND DISCUSSION

Results of simulations from HapMap: We simulated case control data from HapMap data (see Methods) under three disease models and compared the power of IPM, CDM and RBP with stepwise logistic regression (Chambers and Hastie, 1992) using AIC and BIC. We first used BEAM2 (Zhang *et al*., 2011) to search for associated SNPs. We ranked SNPs in each dataset according to their association posterior probabilities, then calculate the power of each program from 50 datasets. The power is defined as the fraction of disease related SNPs ranked among the top ranked SNPs. A SNP is regarded as disease related if it is within 5 SNPs on either side of a true disease locus. The results are shown in Fig. 1 for marginal independence and Fig. 2 for conditional independence simulations.

Consistent with the simulation results from Zhang *et al*. (2011) , we observed that BEAM2 can identify most of the associated SNPs with good power for Model 1 and 2, but the performance for Model 3 is worse than the other two due to the fact there is no marginal effect in Model 3 at all.

Results for IPM and CDM compared with stepwise logistic regression: Then we focus on the associated SNPs (four in marginal independence simulations and three in conditional independence simulations) and compare the powers of IPM and CDM with stepwise logistic regression for correctly identifying interacting structures (marginal independence groups and conditional independence). For marginal independence simulations (see Methods), Fig. 3 show the powers of IPM and stepwise logistic regression (general model, i.e., genotypes are treated as categorical variables and additive model, i.e., genotypes (AA, Aa, aa) are treated as 0, 1, 2 numerical variables) using AIC and BIC. For conditional independence simulations, Fig. 4 shows the powers of CDM and stepwise logistic regression (general and additive models). Here IPM and CDM only report the model with highest posterior probability and stepwise logistic regression starts from the model with all the main effect terms, greedily adding or deleting one term at each step, until AIC or BIC stop dropping. The final model is reported by stepwise logistic regression. Then the power is calculated as the fraction of correctly inferred the interacting structures (models) among all the 50 simulated datasetsunder each parameter settings. From Fig. 3 and 4, it is clear that IPM or CDM outperform stepwise logistic regression in most parameter settings. And BIC is more powerful than AIC.

Results for RBP compared with stepwise logistic regression: Then we simulated a complicated interaction model to access the power of RBP (see Methods). In this simulation, there are totally 21 disease-associated SNPs in 6 two-locus marginal independence groups (2 for each of the models in Table 1) and 3 three-locus conditional independence groups (one for each of the models in Table 1. RBP first uses IPM to infer the independence groups and then uses CDM to infer conditional independence within each group. Figure 5 shows the power boxplot of RBP for 50 simulation datasets. The power is the fraction of SNPs whoseinteractions are correctly inferred (i.e., the SNPs are inferred in the correct independent group with correct neighbors and relationships). It is clear that as interaction parameter (θ) increases, the power substantially increase and the minor allele frequencies (f) make little difference except that when f = 0.5, there is a big drop in power. We also tried stepwise logistic regression on these simulations, but due to complicated interactions, neither AIC or BIC (general model or additive model) can have positive power.

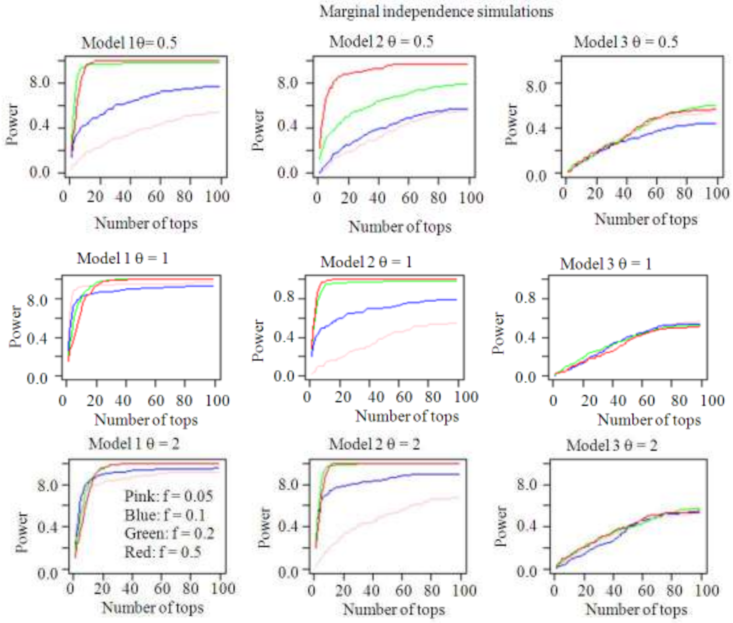


Fig. 1: Power curves of BEAM2 on marginal independence simulations. Under each setting, the power is calculated as the proportion of disease-associated SNPs in 50 datasets identified within 5 SNPs from the top m SNPs ranked by posterior probability (m ranges from 1-100). Each data set contains 1,000 SNPs in 1000 cases and 1000 controls. The disease allele frequency in the population is 0.05 (pink), 0.10 (blue), 0.20 (green) and 0.50 (red), respectively

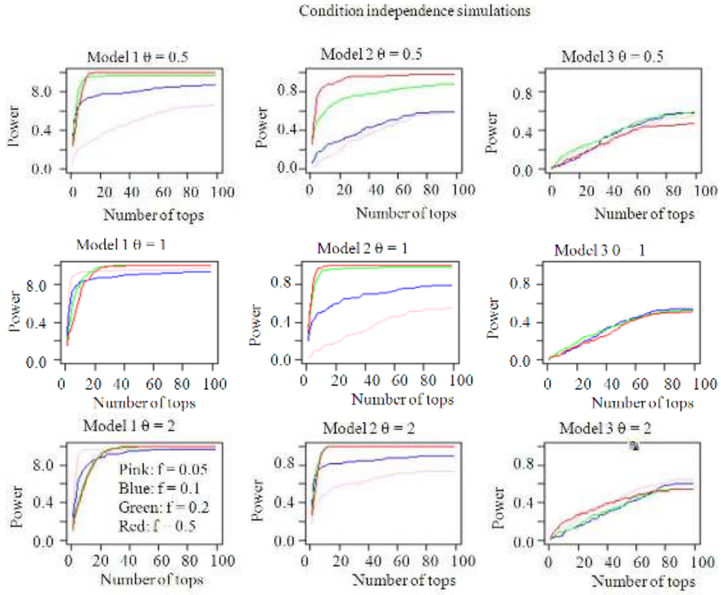


Fig. 2: Power curves of BEAM2 on conditional independence simulations. Under each setting, the power is calculated as the proportion of disease-associated SNPs in 50 datasets identified within 5 SNPs from the top m SNPs ranked by posterior probability (m ranges from 1-100). Each data set contains 1,000 SNPs in 1000 cases and 1000 controls. The disease allele frequency in the population is 0.05 (pink), 0.10 (blue), 0.20 (green) and 0.50 (red), respectively

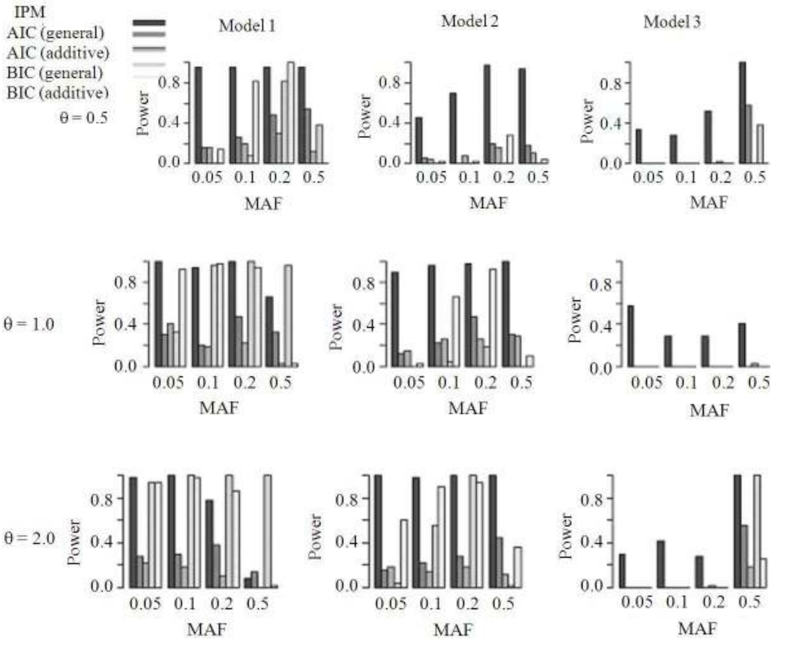


Fig.3: For marginal independence simulations (see Methods), Fig. 3 show the powers of IPM and stepwise logistic regression (general model, i.e., genotypes are treated as categorical variables and additive model, i.e., genotypes (AA, Aa, aa) are treated as 0, 1, 2 numerical variables) using AIC and BIC. Theta is the interaction parameter θ in Table 1. MAF is minor allele frequency

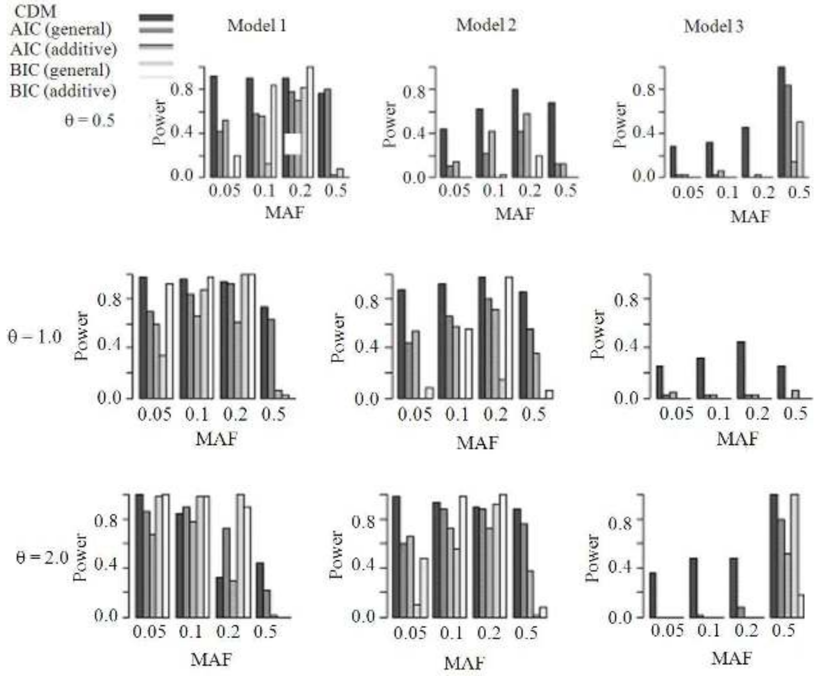


Fig. 4: For conditional independence simulations, Figure 4 shows the powers of CDM and stepwise logistic regression (general and additive models) using AIC or BIC. Theta is the interaction parameter θ in Table 1. MAF is minor allele frequency

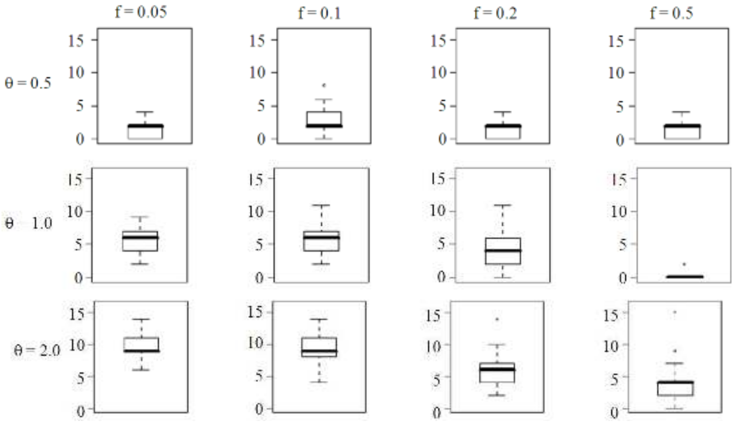
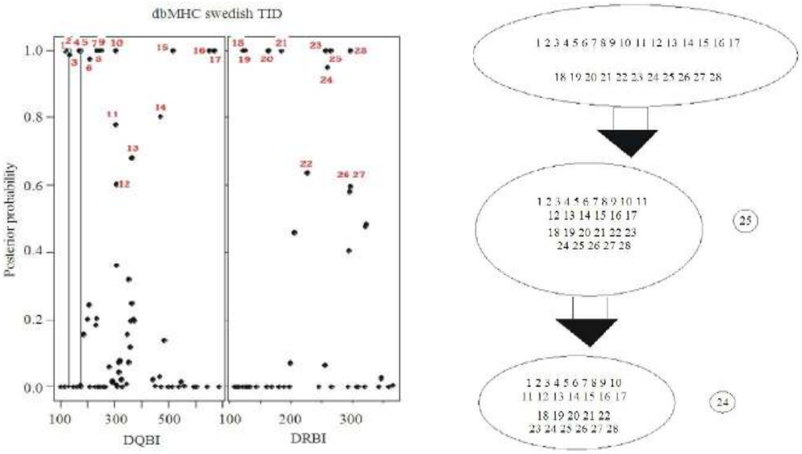


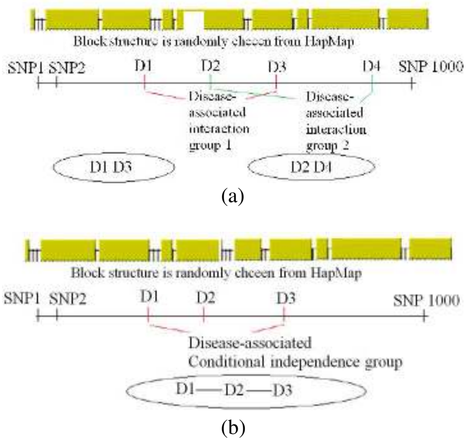
Fig. 5: Shows the power boxplot of RBP for 50 complicated interaction simulation datasets under different parameter settings



(a) (b)

Fig. 6: (a) SNP-wise posterior probabilities for dbMHC T1D data using BEAM2 and SNPs with posterior probabilities >0.5 are numbered. The dbMHC samples were collected from over 20 populations worldwide, among which 1473 individuals had unknown origins. To avoid potential confounding effects of the population structure, we only used data from the Swedish population, which contained 665 cases and 524 controls. Dot indicates the marginal posterior probability of association per SNP and circle indicates the total posterior probability of association per SNP (marginal plus joint). We connect the dot and circle for each SNP for better illustration; (b) The procedure of RBP. First we applied IPM to all of the 28 associated SNPs. Only the 25th SNP comes out independently of all the others. Then reapplying IPM to the rest of 27 SNPs, only the 24th SNP is independent of the other 26 SNPs. When reapplying IPM to the rest 26 SNPs, no independence was found (posterior probability >0.9). Then we applied CDM to the 26-SNP group, again no conditional independence was found (posterior probability >0.9)

Validation of RBP with dbMHC T1D data: We used T1D data from dbMHC (http://www.ncbi.nlm.nih.gov/gv/mhc/) to validate our RBP method. The data contained resequenced haplotypes of exons of two MHC genes DRB1 and DQB1. Since it is well-known that in HLA DQ-DR region SNPs form associated haplotypes rather than interactions, RBP should find most associated SNPs in this region cluster in one big marginal independent group and no conditional independence within the group. We first applied BEAM2 to search for associated SNPs with posterior probabilities >0.5. Figure 6a shows the posterior probabilities and number SNPs with posterior probabilities >0.5. Then we applied RBP to detect marginal independence and conditional independence recursively. Figure 6b shows the procedure of RBP. First we applied IPM to all of the 28 associated SNPs. Only the 25th SNP comes out independently of all the others.


Block structure is randomly cheeen from HapMap
SNP1 SNP2
D1
D2
D3
D4
SNP 1000
Disease-associated interaction group 1
Disease-associated interaction group 2
D1 D3
D2 D4
Block structure is randomly cheeen from HapMap
SNP1 SNP2
D1
D2
D3
SNP 1000
Disease-associated Conditional independence group
D1—D2—D3

(a)

(b)

Fig. 7: (a) The illustration of marginal independence simulations (Ding *et al*., 2005a). To simulate case control data, we first randomly select a region in the human genome that contains 1000 tagged SNPs from HapMap. Four SNPs (D1, D2, D3, D4) are then randomly selected as disease loci for two marginal independent groups

Then reapplying IPM to the rest of 27 SNPs, only the 24th SNP is independent of the other 26 SNPs. When reapplying IPM to the rest 26 SNPs, no independence was found (posterior probability >0.9). Then we applied CDM to the 26-SNP group, again no conditional independence was found (posterior probability >0.9). So our results show that DQB1 and DRB1 do form associated haplotypes, rather than interactions.

CONCLUSION

In this study, we proposed a Bayesian method called RBP to recursively infer the interaction structure from case-control data. The method is composed of two steps: in the first step IPM was recursively used to infer the marginal independent groups; in the second step CDM was recursively used to infer the conditional independence within each independent group inferred from the first step. To our knowledge this is the first method to infer the dependence structure of interaction in GWAS recursively. We then designed several simulation studies to test IPM, CDM and RBP using marginal independence simulations, conditional independence simulations and complicated interaction simulations based on HapMap data and block structures. Using these extensive simulations we showed our method is more powerful than stepwise logistic regression using AIC or BIC in both marginal independence and conditional independencedetections. In the complicated interaction simulations, our method is much more powerful than stepwise logistic regression. We also validated RBP using dbMHC T1D data and showed DQB1 and DRB1 form strong associated haplotypes for T1D.

Although this type of Bayesian recursive partition idea has been successfully used in several difference scenarios (Zhang *et al*., 2010; Svicher *et al*., 2011a; 2011b) the current RBP model can be further improved in several ways. First, this model does not consider missing genotypes and unobserved SNPs in the case-control sample, but this can be treated via imputation or EM algorithm incorporated in RBP model. Previous studies have shown that imputing untyped SNPsand missing genotypes from a reference panel can improve the power of disease association mapping (Svicher *et al*., 2011a). Second, the current model only considers one case and one control data. Actually RBP model can be improved to incorporate multiple case-control datasets and improving the detecting power since related information can be borrowed from multiple case-control datasets. We are now developing new statistical methods for these improvements.

ACKNOWLEDGEMENT

JZ was supported by Yale University startup fund and Sesseel Fellowship. MQZ was supported by the National Basic Research Program of China (2012CB316503), National Natural Science Foundation of China (91019016,31061160497) and NIH grant (HG001696). DL was supported by fund from the Stanley Foundation. We thank valuable suggestions from Yu Zhang and Jun Liu.

REFERENCES

Chambers, J.M. and T. Hastie, 1992. Statistical Models in S. 1st Edn., Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, Calif., ISBN: 0534167640, pp: 608.

Ding, K., J. Zhang, K. Zhou, Y. Shen and X. Zhang, 2005a. htSNPer 1.0: Software for haplotype block partition and htSNPs selection. BMC Bioinform.

Ding, K., K. Zhou, J. Zhang, J. Knight and X. Zhang *et al*., 2005b. The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. Mol. Biol. Evolution, 22: 148-159. DOI: 10.1093/molbev/msh266

Marchini, J., B. Howie, S. Myers, McVean and G.P. Donnelly, 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet., 39: 906-913. DOI: 10.1038/ng2088

Steenkiste, A., A.M. Valdes, M. Feolo, D. Hoffman and P. Concannon *et al*., 2007. 14th International HLA and immunogenetics workshop: Report on the HLA component of type 1 diabetes. Tissue Antigens, 69: 214-25. DOI: 10.1111/j.1399-0039.2006.00772.x

Svicher, V., C. Alteri, A. Artese, J.M. Zhang and G. Costa *et al*., 2011b. Identification and structural characterization of novel genetic elements in the HIV-1 V3 loop regulating coreceptor usage. Antiviral Therapy, 16: 1035-1045. DOI: 10.3851/IMP1862

Svicher, V., V. Cento, M. Bernassola, M. Neumann-Fraune and F.V. Hemert *et al*., 2011a. Novel HBSAG markers tightly correlate with occult HBV infection and strongly affect HBSAG detection. Antiviral Res., 93: 86-93. DOI: 10.1016/j.antiviral.2011.10.022

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature, 437: 1299-1320. DOI: 10.1038/nature04226

WTCCC, 2007a. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447: 661-678. PMID: 17554300

WTCCC, 2007b. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat. Genet., 39: 1329-1337. PMID: 17952073

Yang, Y., C. He and J. Ott, 2009. Testing association with interactions by partitioning chi-squares. Ann. Hum. Genet., 73: 109-117. DOI: 10.1111/j.1469-1809.2008.00480.x

Zhang, J., F. Li, J. Li, M.Q. Zhang and X. Zhang, 2004. Evidence and characteristics of putative human alpha recombination hotspots. Hum. Mol. Genet., 13: 2823-2828. PMID: 15385449

Zhang, J., T. Hou, W. Wei and S.J. Liu, 2010. Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. PNAS, 107: 1321-1326. DOI: 10.1073/pnas.0907304107

Zhang, Y. and S.J. Liu, 2007. Bayesian inference of Epistatic interactions in case-control studies. Nat. Genet., 39: 1167-1173. DOI: 10.1038/ng2110

Zhang, Y., J. Zhang and S.J. Liu, 2011. Block-based bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. Ann. Applied Stat., 5: 2052-2077. DOI: 10.1214/11-AOAS469