DAM: A Bayesian Method for Detecting Genome-wide Associations on Multiple Diseases

Xuan Guo[1], Jing Zhang[2] *, Zhipeng Cai[1], Ding-Zhu Du[3], and Yi Pan[14] *

[1] Department of Computer Science, Georgia State University, Atlanta, GA
[2] Department of Mathematics and Statistics, Georgia State University, Atlanta, GA
[3] Department of Computer Science, The Univ. of Texas at Dallas, Richardson, TX
[4] Department of Biology, Georgia State University, Atlanta, GA

Abstract. Taking the advantage of high-throughput single nucleotide polymorphism (SNP) genotyping technology, large genome-wide association studies (GWASs) have been considered to hold promise for unraveling complex relationships between genotypes and phenotypes. Current multi-locus-based methods are insufficient to detect interactions with diverse genetic effects on multifarious diseases. In addition, statistic tests for high order epistasis ($\geq$ 2 SNPs) raise huge computational and analytical challenges because the computation increases exponentially as the growth of the cardinality of SNPs combinations. In this paper, we provide a simple, fast and powerful method, DAM, using Bayesian inference to detect genome-wide multi-locus epistatic interactions on multiple diseases. Experimental results on simulated data demonstrate that our method is powerful and efficient. We also apply DAM on two GWAS datasets from WTCCC, *i.e.* Rheumatoid Arthritis and Type 1 Diabetes, and identify some novel findings. Therefore, we believe that our method is suitable and effective for the full-scale analysis of multi-disease-related interactions in GWASs.

Keywords: Bayesian inference, Genome-wide association studies, Genetic factors, Epistasis

1 INTRODUCTION

Genome-wide association study (GWAS) has been proved to be a powerful genomic and statistical inference tool. The goal is to identify genetic susceptibility through statistical tests on associations between a trait of interests and the genetic information of unrelated individuals [1]. In genetics, genotype-phenotype association studies have established that single nucleotide polymorphisms (SNPs) [2], one type of genetic variants, are associated with a variety of diseases [3]. The current primary analysis paradigm for GWAS is dominated by the analysis on susceptibility of individual SNPs to one disease a time, which might only explain a small part of genetic causal effects and relations for multiple complex diseases [4]. The word, epistasis, has been defined generally as

* To whom correspondence should be addressed.

the interaction among different genes [5]. Many studies [9] have demonstrated that the epistasis is an important contributor to genetic variation in complex diseases, such as asthma [6][8], breast cancer [10], diabetes[7], coronary heart disease [11], and obesity [12]. In this article, we consider epistatic interactions as the statistically significant associations of d-SNP modules ($d \geq 2$) with multiple phenotypes [13].

Recently, the problem of detecting high-order genome-wide epistatic interaction for case-control data has attracted extensive research interests. Generally, there are two challenges in mapping genome-wide associations for multiple diseases on large GWAS dataset [14]: the first is arose from the heavy computational burden, *i.e.* the number of association patterns increases exponentially as the order of interaction goes up. For example, around 6.25×10^{11} statistical tests are required to detect pairwise interactions for a dataset with 500,000 SNPs. The second challenge is that existing approaches lack statistical powers for searching high-order multi-locus models of disease. Because of the huge number of hypotheses and the limited sample size, a large proportion of significant associations are expected to be false positives. Many computational algorithms have been proposed to overcome the above difficulties. More details about these tools can be found in a recent survey [15]. To the best of our knowledge, current epistasis detecting tools are only capable of identifying interactions on GWAS data with two groups, *i.e.* case-control studies. Thus, they are incompetent to discover genetic factors with diverse effects on multiple diseases. Moreover, they lose the benefit of alleviating deficiency of statistical powers by pooling different disease samples together.

In this paper, we design and implement a Bayesian inference method for Detecting genome-wide Association on Multiple diseases, named DAM, to address above challenges. DAM employs Markov Chain Monte Carlo (MCMC) sampling based on the Bayesian variable partition model, and makes use of stepwise condition evaluation to identify significant disease(s)-specific interactions. It first generates a candidate set of SNPs based on our Bayesian variable partition model by applying Metropolis-Hastings (MH) algorithm. A stepwise evaluation of association is engaged to further detect the genetic effect types for each interaction. Systematic experiments on both simulated and real GWAS datasets demonstrate that our method is feasible for identify multi-locus interaction on GWAS datasets and enriches some novel, significant high-order epistatic interactions with specialties on various diseases.

2 METHOD

2.1 Notations

Suppose a GWAS dataset D has M diallelic SNPs and N samples. In general, bi-allelic genetic markers use uppercase letters (e.g. A, B,...) to denote major alleles and lowercase letters (e.g. a, b) to denote minor alleles. For encoding three genotypes, one popular way is to use $\{1, 2, 3\}$ to represent $\{aa, Aa, AA\}$, respectively. For a GWAS dataset with L groups, it includes one shared control group

and $L-1$ case groups. We use $N^{(L)}$ denotes the number of controls (*i.e.* normal individuals) and $N^{(i)}$ denotes the number of cases (*i.e.* disease individuals) in i-th groups ($i = 1 \ldots L-1$). X is utilized to indicate the ordered set of SNPs, and x_i represents i-th SNP in X.

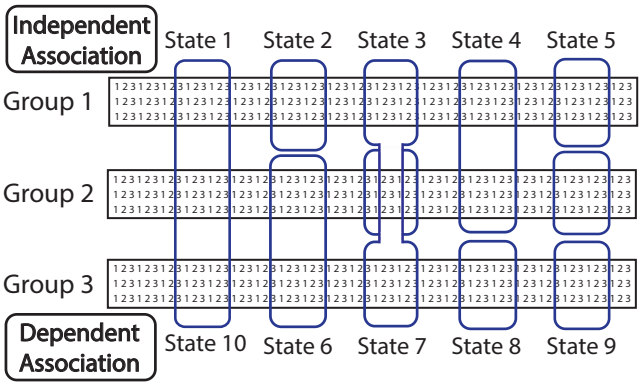

Fig. 1. Illustration for 10 states on 3 groups.

For a set of L groups, there are B_L partitions, and here we also refer partition to state. Let S denote the set of states, and s_k is the k-th state with $|s_k|$ non-empty sets of groups. In general, the M markers are assigned into $2B_L$ states, and all states belong to two categories: $s_{k_1} \in \{s_1, \ldots, s_{B_L}\}$ indicates SNP markers contributing independently to the phenotypes, and $s_{k_2} \in \{s_{B_L+1}, \ldots, s_{2B_L}\}$ indicates SNP markers that jointly influence the phenotypes. An example for a three groups dataset with 10 possible states is showed in Figure 1, where states 1 to 5 indicate that SNPs are independently associated with certain phenotypes, and states 6 to 10 indicate that SNPs are dependently associated with the phenotypes. In our experiments, group 1 and 2 are cases and group 3 is control. Since we want to identify SNPs associated with phenotypes, SNPs in states 2 to 5 and states 6 to 10 are the desired ones with disease associations. Let $I = (I_1, \ldots, I_M)$ record the memberships of SNP with $I_m \in \{1, \ldots, 2B_L\}$, $\mathbb{M}_k$ denote the number of SNP markers in k-th state ($\sum_{k=1}^{2B_L} \mathbb{M}_k = M$), and $D^{(k)}$ denote genotypes of SNPs in k-th state.

2.2 Bayesian variable partition model

Consider a categorical variable X, which can be sampled at t different states $\{\Delta_1, \Delta_2, \ldots, \Delta_t\}$ with t different distribution $\{\Theta_1, \Theta_2, \ldots, \Theta_t\}$, where Θ_k is the distribution of X at k-th state. The model describing the sums of independently and identically distributed mixture categorical variables at different states is referred as a 'multinomial model', meaning that it can be partitioned into t

inseparable multinomial models. Consider a model for a vector of M categorical variables $X = \{x_1, x_2, \ldots, x_M\}$. If all variables are independent, the model can be simply treated as the union of M univariate multinomial models. If interactions exist among multiple variables, a new model with a single variable by collapsing the interacting variables can replace the model for these multiple variables. The sample space of the collapsed variable is the product of the sample spaces of the variables before collapsing. Bayesian variable partition model (BVP) is a multinomial model based on Bayesian theorem. The likelihood for the multinomial model by given i-th state is

$$\begin{aligned} P(D_m|\Delta_k) &= \int P(D_m|\Delta_k, \Theta_k) d\Theta_k \\ &= \int_{\theta_1,\theta_2,\ldots,\theta_g} P(D_m|\theta_1, \theta_2, \ldots, \theta_g) P(\theta_1, \theta_2, \ldots, \theta_g) dp \end{aligned} \tag{1}$$

where D_m is the observation for the categorical variable x_m, and g is the number of category value for the variable x_m. We set $P(\Theta = (\theta_1, \theta_2, \ldots, \theta_g))$ to be Dirichlet distribution $Dir(\alpha_1, \alpha_2, \ldots, \alpha_g)$; then we can have a closed form for Equation 1:

$$\begin{aligned} P(D_m|\Delta_k) &= \int_{\theta_1,\theta_2,\ldots,\theta_g} P(D_m|\theta_1, \theta_2, \ldots, \theta_g) P(\theta_1, \theta_2, \ldots, \theta_g) dp \\ &= \int_{\theta_1,\theta_2,\ldots,\theta_g} \frac{1}{B(\alpha_1, \alpha_2, \ldots, \alpha_g)} \prod_{i=1}^{g} p_i^{n_i + a_i - 1} dp \\ &= \left(\prod_{i=1}^{g} \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)} \right) \frac{\Gamma(|\alpha|)}{\Gamma(\mathbb{N} + |\alpha|)} \end{aligned} \tag{2}$$

where $\mathbb{N}$ is the total number of observations, and $|\alpha|$ is the sum of $(\alpha_1, \alpha_2, \ldots, \alpha_g)$. Suppose the vector I is a vector of membership of state for categorical variable vector X, we obtain the posterior distribution of I as

$$P(I|D) \propto \left(\prod_{m=1}^{M} P(D_m|I) \right) P(I) \tag{3}$$

Based on Bayesian theorem, we describe the specific Bayesian variable partition model for genome-wide association mapping as follows. For the SNPs independently associated with phenotypes, we use $\Theta_{k_1} = ((\theta_{m1}^{(\omega)}, \theta_{m2}^{(\omega)}, \theta_{m3}^{(\omega)}) : \omega \in \{1, 2, \ldots, |s_k|\}, I_{x_m} \in \{1, \ldots, B_L\})$ to denote the genotype frequencies of SNP x_m in k_1 states. Note that SNP with membership value in $\{1, \ldots, B_L\}$ does not have interaction with other SNPs. The likelihood of D^{k_1} from BVP model is that

$$P(D^{(k_1)}|\Theta_{k_1}) = \prod_{I_{x_m}=k_1} \prod_{\omega=1}^{|s_{k_1}|} \prod_{i=1}^{3} (\theta_{mi}^{(\omega)})^{n_{mi}^{(\omega)}}, \tag{4}$$

where $\left\{n_{m1}^{(\omega)}, n_{m2}^{(\omega)}, n_{m3}^{(\omega)}\right\}$ are genotype counts of SNP x_m in ω-th subset in k_1-th state. Similar to the above assumption, we set Θ_{k_1} to be a Dirichlet distribution $Dir(\alpha)$ with parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, we integrate out Θ_{k_1} and obtain the marginal probability:

$$P(D^{(k_1)}|I) = \prod_{I_{x_m}=k_1} \prod_{\omega=1}^{|s_{k_1}|} \left(\left(\prod_{i=1}^{3} \frac{\Gamma(n_{mi}^{(\omega)} + \alpha_i^{(\omega)})}{\Gamma(\alpha_i^{(\omega)})} \right) \frac{\Gamma(\left|\alpha^{(\omega)}\right|)}{\Gamma(\mathbb{N}_{k_1,\omega} + \left|\alpha^{(\omega)}\right|)} \right) \tag{5}$$

where $\mathbb{N}_{k_1,\omega}$ is the count of individuals in groups belonging to ω-th subset of k_1-th state, and $|\alpha|$ represents the sum of all elements in α.

SNP markers in state $\{s_{B_L}, s_{B_L+1}, \ldots, s_{2B_L}\}$ influence the disease statues through interactions. Thus, we concatenate $\mathbb{M}_{k_2}$ SNPs into a single categorical variable to resolve the interactions ($B_L+1 \leq k_2 \leq 2B_L$). Note that there are $3^{\mathbb{M}_{k_2}}$ possible concatenated genotype combinations. Let $\Theta_{k_2} = ((\phi_1^{(\omega)}, \phi_2^{(\omega)}, \ldots, \phi_{3^{\mathbb{M}_{k_2}}}^{(\omega)}) : \omega = \{1, 2, \ldots, |s_{k_2}|\})$ be the concatenated genotype frequencies over $\mathbb{M}_{k_2}$ SNPs in $s_{k_2} \in \{s_{B_L+1}, \ldots, s_{2B_L}\}$. Similarly, we use a Dirichlet prior $Dir(\beta)$ for Θ_{k_2}, $\beta = (\beta_1, \beta_2, \ldots, \beta_{3^{\mathbb{M}_{k_2}}})$. According to Equation 2, we obtain the marginal probability:

$$P(D^{(k_2)}|I) = \prod_{\omega=1}^{|s_{k_2}|} \left(\left(\prod_{i=1}^{3^{\mathbb{M}_{k_2}}} \frac{\Gamma(n_i^{(\omega)}) + \beta_i^{(\omega)}}{\Gamma(\beta_i^{(\omega)})} \right) \frac{\Gamma(|\beta^{(\omega)}|)}{\Gamma(\mathbb{N}_{k_2,\omega} + |\beta^{(\omega)}|)} \right) \tag{6}$$

where $\mathbb{N}_{k_2,\omega}$ is the count of individuals belonging to ω-th subset k_2-th state and $n_i^{(\omega)}$ is the count of i-th concatenated genotype combinations in ω-th subset in k_2-th state.

Combining Equation 3, 5 and 6 , we obtain the posterior distribution of I as

$$P(I|D) \propto \left(\prod_{k_1=1}^{B_L} P(D^{(k_1)}|I) \right) \left(\prod_{k_2=(B_L+1)}^{2B_L} P(D^{(k_2)}|I) \right) P(I) \tag{7}$$

In BVP, we set $P(I) \propto \prod_{k=1}^{2B_L} p_k^{\mathbb{M}_k}$ to embed the prior knowledge of the proportions of SNP associating with certain phenotypes. In our experiments with three groups, we set $p_k = 0.001, k \in \{2, \ldots, 10\}$, and $\alpha_i = \beta_j = 0.5, \forall i, j$.

2.3 MCMC sampling

We apply MCMC method to sample the indicator I from the distribution in Equation 7. According to the prior $P(I)$, DAM first initializes I, then use the Metropolis-Hastings (MH) algorithm [16] to construct a MCMC to update I. Three types of updating strategies are used: (i) randomly change a SNP's state, (ii) randomly exchange two SNPs' states between $(s_1, \ldots, s_{2B_L})$, or (iii) randomly shuffle the state labels between $\{s_{B_L+1}, \ldots, s_{2B_L}\}$. At each iteration, the acceptance of new indicator based on the MH ratio, a Gamma functions. DAM records the entire accepted indicator after the burn-in process, and represent it as the posterior distribution of single disease-related SNPs and interactions associated with multiple diseases. The number of iteration in burn-in process is fixed to $10M$ and the number of sampling iteration is set to M^2 in our experiments. We also apply a distance constraint that the physical distance between two SNPs in multi-locus module is at least 1Mb. This constraint is used to avoid associations that might be attributed to the LD effects [5].

2.4 Evaluation of interaction

With the candidate SNPs generated by MCMC sampling, we apply the χ^2 statistic and its conditional test to measure the significance for a dependent SNP association. Let $\mathbb{A} = (x_1, x_2, \ldots, x_d : k)$ denote an SNP module $\mathbb{A}$ with d SNPs in k-th state. We denote its χ^2 statistic as $\chi^2(x_1, x_2, \ldots, x_d : k)$ and the conditional χ^2 statistic as $\chi^2(x_1, x_2, \ldots, x_d | x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : k)$ by given a module $\mathbb{A}$ and a subset of it, $(x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}})$ with d' SNPs. The χ^2 statistic can be calculated as

$$\chi^2(x_1, x_2, \ldots, x_d : k) = \sum_{i=1}^{|s_k|} \sum_{j=1}^{3^d} \frac{(n_{g_i,j} - e_{g_i,j})^2}{e_{g_i,j}} \tag{8}$$

where g_i is the i-th genotype combination for d SNPs, $n_{g_i,j}$ is the number of individuals having i-th genotype combination in j-th subset in k-th state, and $e_{g_i,j}$ is the corresponding expected value. The degrees of freedom for Equation 8 is $(|s_k| - 1) \cdot (3^d - 1)$. The conditional independent test via χ^2 statistic is defined as follows

$$\chi^2(x_1, x_2, \ldots, x_d | x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : k) = \sum_{\iota=1}^{3^{d'}} \sum_{i=1}^{|s_k|} \sum_{j=1}^{3^{d-d'}} \frac{(n_{g_i,j}^{(\iota)} - e_{g_i,j}^{(\iota)})^2}{e_{g_i,j}^{(\iota)}} \tag{9}$$

where we calculate χ^2 statistic separately for each genotype combination from $\mathbb{A} - \mathbb{A}'$. The degrees of freedom for Equation 9 is $3^{d'} \cdot (|s_k| - 1) \cdot (3^{d-d'} - 1)$. In order to avoid redundant SNPs in a SNP module indicating that conditional independence model fits better, we define an epistatic interaction ($d \leq 2$) as a compact significant SNP module with definition 1.

Definition 1 *A SNPs module* $\mathbb{A} = (x_1, x_2, \ldots, x_d : k)$ *is considered as a compact significant interaction by given the significant level* α_d*, if it meets the following three conditions:*
(1) the p-value of $\chi^2(x_1, x_2, \ldots, x_d : k) \leq \alpha_d$*;*
(2) the p-value of $\chi^2(x_1, x_2, \ldots, x_d : k) < \forall$ *p-value of* $\chi^2(x_1, x_2, \ldots, x_d : k'), k \neq k'$ *and* $k' \in \{1, 2, \ldots, |S|\}$*;*
(3) the p-value of $\chi^2(x_1, x_2, \ldots, x_d | x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : k) \leq \alpha_d$
for $\forall \mathbb{A}' = (x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : k)$ *whose p-value* $\leq \alpha_{d'}$*.*

Based on definition 1, we develop a stepwise algorithm to search for top-f significant d-locus compact significant interactions, where the searching space only includes the SNP markers generated by MCMC sampling. We assume that one SNP can only participate in one significant interaction in one state. So for the SNP markers with state in $\{s_1, \ldots, s_{B_L}\}$, we first searches all the modules with just one SNP based on definition 1, then the algorithm recursively tests all the possible combinations by setting the module size with one more SNP. For the SNPs reported as jointly contributing to the disease risk, we calculate the p-value under different states and use the conditional test if part of SNPs already reported as significant. All SNPs with significant marginal associations after a Bonferroni correction are reported in a list $\mathbb{L}$. The algorithm recursively searches the interaction space with larger module size until d reaches user preset value. We add all novel d-way interactions (*i.e.*, no SNPs has been reported earlier) that are significant after the Bonferroni correction for $2B_L \cdot \binom{M}{d}$ tests. For the interactions whose subsets have been reported as compact significant, we use the conditional independent test, and put the interaction in $\mathbb{L}$ if it is still significant after Bonferroni correction of $2B_L \cdot \binom{M}{d} \cdot \binom{d}{d'}$ tests.

3 RESULTS

To the best of our knowledge, DAM is the first method to detect associations on multiple diseases, so we first give definitions of 8 simulated multi-disease models and the power metric measurement, and then evaluate the effectiveness of our method. The false positive rate of DAM is showed in Supplementary Material. We also apply DAM on two real GWAS datasets, Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D), and we find not only the results reported by other literatures but also some novel interesting interactions. DAM (in Java) is conducted on a 64-bit Windows 8 platform with 1.8 GHz Intel CPU and 4 GB RAM.

3.1 Experimental design

Data simulation To evaluate the effectiveness of DAM, we perform extensive simulation experiments using eight disease models with one- and two-locus associations on three groups. The genotypes of unassociated SNP are generated by the same procedure used in previous studies [14] with Minor Allele Frequencies

(MAFs) sampled from $[0.05, 0.5]$. The odds tables for eight models are showed in Table 1 in Supplementary Material. Model 5, 6, 7, and 8 are the extensions of Model 1, 2, 3, and 4, respectively. The settings for four datasets are showed in Table 2 in Supplementary Material. In a setting, all models are using the same MAF $\in \{0.1, 0.2, 0.4\}$, we generate 100 replicas per setting. Therefore, by given a MAF, a dataset contains at most 8 associations labeled as Ep 1 to 8. Note that in model 5 there are 7 associations, because the combination of three 2-locus models does not exist when MAF $= 0.1$. Each simulated replica contained $M = 1000$ SNPs. The sizes of three groups are set to $(1000, 1000, 2000)$ or $(2000, 2000, 4000)$, where the first two groups are considered as case groups and the third one is control group. More details about model simulation can be found in Supplementary Material.

Statistical power In the evaluation of performances on simulated data, 100 datasets are generated for each setting. The measure of discrimination power is defined as the fraction of 100 datasets on which the ground-truth associations are identified as compact and significant by DAM.

3.2 Single-locus disease models

Test results are illustrated in Figure 2 in Supplementary Material for SNPs contributing independently to the disease risks. We can find that DAM is able to report nearly 100% of embedded single SNP associations under most settings. Carefully examining the results, we found that some SNPs are incorrectly assigned to a state by MCMC sampling, although they do have significant association with the phenotypes. After the stepwise evaluation, most mistakenly labeled SNPs are corrected.

3.3 Two-locus disease models

Test results for SNPs contributing jointly to the disease risks are illustrated in Figure 2. We can find that DAM is able to report nearly 100% of embedded interactions for dataset 1 and 2. It also obtained nearly full power when MAF is 0.1 for dataset 1, 2, and 4. Similar to the results on single-locus disease models, after stepwise procedure, more interactions were assigned to correct states.

3.4 Experiments on WTCCC data

We have applied DAM to analyze data from the WTCCC (3999 cases in total and 3004 shared controls) on two common human diseases: Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D), where RA is treated as group 1, T1D is treated as group 3, and control group is group 3. The procedure of quality control is the same as presented in the [14]. After SNP filtration the dataset contains 333,739 high quality SNPs. DAM ran about 36 hours, for a total of 1×10^{11} iterations. Because the importance of the MHC region in chromosome 6 with respect to

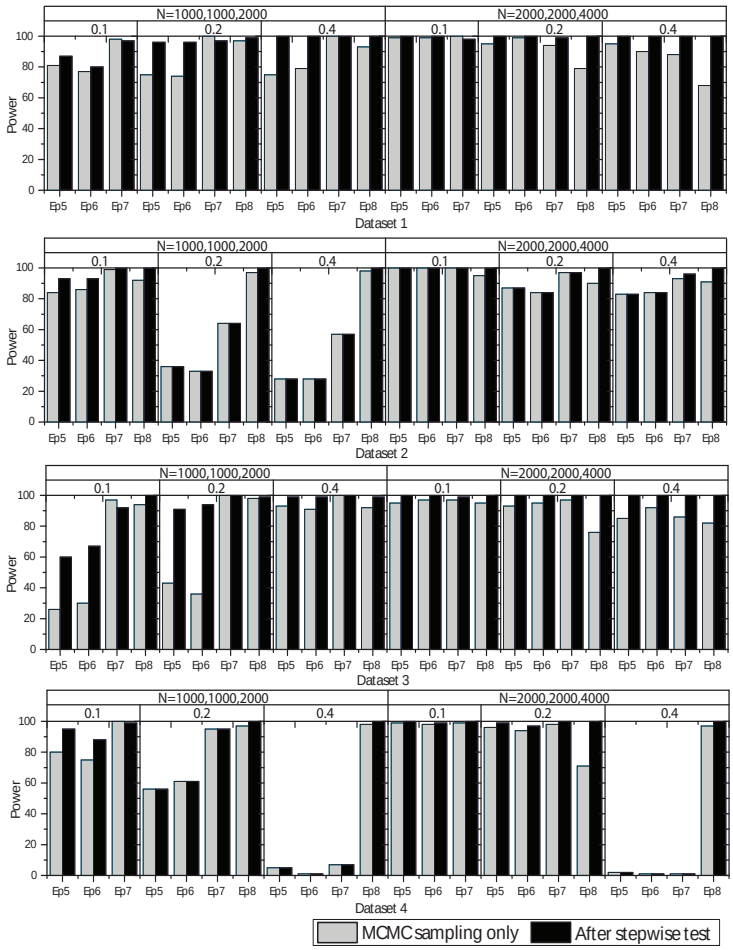

Fig. 2. Performance comparison between DAM with MCMC sampling only and with stepwise test on simulated disease datasets 1-4 embedded with Joint effect SNPs. Note that the combination of model 5 with other three 2-locus models does not exist when MAF = 0.1

infection, inflammation, autoimmunity, and transplant medicine has been heavily reported [17] [18] [19], we concentrate on the results by DAM on Chromosome 6. The posterior probabilities for SNP on Chromosome 6 are showed in Fig 3 in Supplementary Material and Figure 3.

Recent studies [18] [20] has shown that both T1D and RA strongly associated with the MHC region via single-locus association mapping, which is also verified by our results that a large portion of SNPs' posterior probabilities greater than 0.5 spreading in the region $28,477,797 - 33,448,354$. Comparing results from state 6 to state 7, we can find that many SNPs contributing to RA are not

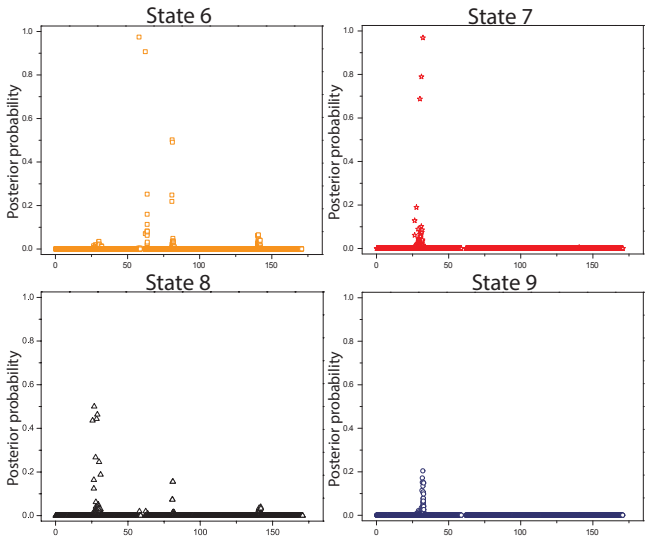

Fig. 3. Posterior probabilities of SNPs on chromosome 6. States 6 to 9 indicate joint association probabilities per SNP. X-axis indicates the chromosomal position (Mb), y-axis shows the posterior probability.

located inside the MHC region, while the SNPs associated with T1D gather in MHC region. We select top 50 SNPs according to their posterior probabilities and analyze them with the stepwise evaluation procedure introduced in Section 2.4. Table 1 summarizes some novel findings of the significant interactions with p-values adjusted by 1.61×10^{18} for three loci and 1.93×10^{23} for four loci interactions, respectively. Take the four-locus interaction (rs1977, rs707974, rs10755544, rs2322635) for example. rs1977 is located inside gene BTN3A2, which encodes a member of the immunoglobulin superfamily that may be involved in the adaptive immune response. rs707974 is in gene GPANK1, encoding a protein which plays a role in immunity. rs10755544 is at the upstream of gene KHDRBS2, which is thought to involve SH2 domain binding and protein heterodimerization activity. rs2322635 is located in gene BCKDHB for encoding branched-chain keto acid dehydrogenase, which is a multienzyme complex associated with the inner membrane of mitochondria. BTN3A2 has been shown to associate with T1D in [21]. And mutations in the BCKD gene, BCKDHA, is also known to result in maple syrup urine disease, which is related to T1D [22].

4 CONCLUSIONS

The large number of SNPs genotyped in genome-wide case-control studies poses a great computational challenge in the identification of gene-gene interactions. During the last few years, many computational and statistical tools are developed to finding gene-gene interactions for data with only two groups, *i.e.* case

Table 1. Significant interactions obtained from theWTCCC data. Following each SNP is its location.

State Index	DAM p-value	SNP 1	SNP 2	SNP 3	SNP 4
6	1.35E-26	rs4634439	rs707974	rs4236164	rs2322635
6	1.61E-26	rs6931858	rs707974	rs10755544	rs3805878
7	3.31E-26	rs1977	rs707974	rs10755544	rs2322635
7	1.86E-35	rs3117425	rs1150753	rs239494	
7	5.79E-24	rs200481	rs1150753	rs12194665	

and control groups. In this paper, we present a method, named "DAM", to address the computation and statistical power issues for multiple diseases GWASs based on Bayesian theory. We have successfully applied our method to systematic simulation and also analyzed two datasets from WTCCC. Our experimental results on both simulated and real data demonstrate that DAM is capable of detecting high order epistatic interactions for multiple diseases at genome-wide scale.

Supplementary information Supplementary Material and DAM software are available at `http://www.cs.gsu.edu/~xguo9/research/DAM.html`

Acknowledgments This study is supported by the Molecular Basis of Disease (MBD) program at Georgia State University.

References

1. Sabaa, H., Cai, Z., Wang, Y., Goebel, R., Moore, S., Lin, G.: Whole genome identity-by-descent determination. Journal of Bioinformatics and Computational Biology **11**(02) (2013) 1350002
2. He, Y., Zhang, Z., Peng, X., Wu, F., Wang, J.: De novo assembly methods for next generation sequencing data. Tsinghua Science and Technology **18**(5) (2013) 500–514
3. Peter, K., J., H.D.: Genetic risk prediction : Are we there yet? The New England journal of medicine **360**(17) (2009) 1701 – 1703
4. He, Q., Lin, D.Y.: A variable selection method for genome-wide association studies. Bioinformatics **27**(1) (2011) 1–8
5. Cordell, H.J.: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics **11**(20) (2002) 2463–2468
6. Cai, Z., Sabaa, H., Wang, Y., Goebel, R., Wang, Z., Xu, J., Stothard, P., Lin, G.: Most parsimonious haplotype allele sharing determination. BMC Bioinformatics **10**(1) (2009) 115
7. Wang, Y., Cai, Z., Stothard, P., Moore, S., Goebel, R., Wang, L., Lin, G.: Fast accurate missing snp genotype local imputation. BMC Research Notes **5**(1) (2012) 404

8. Cheng, Y., Sabaa, H., Cai, Z., Goebel, R., Lin, G.: Efficient haplotype inference algorithms in one whole genome scan for pedigree data with non-genotyped founders. Acta Mathematicae Applicatae Sinica, English Series **25**(3) (2009) 477–488
9. Liu, W., Chen, L.: Community detection in disease-gene network based on principal component analysis. Tsinghua Science and Technology **18**(5) (2013) 454–461
10. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. The American Journal of Human Genetics **69** (July 2001) 138–147
11. Nelson, M., Kardia, S., Ferrell, R., Sing, C.: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Research **11**(3) (2001) 458–470
12. Cordell, H.J.: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet **10** (06 2009) 392–404
13. Wang, Y., Liu, G., Feng, M., Wong, L.: An empirical comparison of several recent epistatic interaction detection methods. Bioinformatics **27**(21) (2011) 2936–2943
14. Guo, X., Meng, Y., Yu, N., Pan, Y.: Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. BMC bioinformatics **15**(1) (2014) 102
15. Guo, X., Yu, N., Gu, F., Ding, X., Wang, J., Pan, Y.: Genome-wide interaction-based association of human diseases-a survey. Tsinghua Science and Technology **19**(6) (2014) 596–616
16. Liu, J.S.: Monte Carlo strategies in scientific computing. springer (2008)
17. Lechler, R., Warrens, A.N.: HLA in Health and Disease. Academic Press (2000)
18. Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L., Yu, W.: Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics **87**(3) (2010) 325–340
19. Zhang, J., Wu, Z., Gao, C., Zhang, M.Q.: High-order interactions in rheumatoid arthritis detected by bayesian method using genome-wide association studies data. American Medical Journal **3**(1) (2012) 56–66
20. Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J., Rayner, N.W., Freathy, R.M., et al.: Wellcome trust case control consortium (wtccc), mccarthy mi, hattersley at: Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. Science **316**(5829) (2007) 1336–1341
21. Viken, M.K., Blomhoff, A., Olsson, M., Akselsen, H., Pociot, F., Nerup, J., Kockum, I., Cambon-Thomsen, A., Thorsby, E., Undlien, D., et al.: Reproducible association with type 1 diabetes in the extended class i region of the major histocompatibility complex. Genes and immunity **10**(4) (2009) 323–333
22. Henneke, M., Flaschker, N., Helbling, C., Müller, M., Schadewaldt, P., Gärtner, J., Wendel, U.: Identification of twelve novel mutations in patients with classic and variant forms of maple syrup urine disease. Human mutation **22**(5) (2003) 417–417