TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 12/13 pp678–689
DOI: 10.26599/TST.2019.9010064
Volume 25, Number 5, October 2020

BAM: A Block-Based Bayesian Method for Detecting Genome-Wide Associations with Multiple Diseases

Guanying Wu, Xuan Guo*, and Baohua Xu*

Abstract: Many human diseases involve multiple genes in complex interactions. Large Genome-Wide Association Studies (GWASs) have been considered to hold promise for unraveling such interactions. However, statistic tests for high-order epistatic interactions ($\geqslant 2$ Single Nucleotide Polymorphisms (SNPs)) raise enormous computational and analytical challenges. It is well known that the block-wise structure exists in the human genome due to Linkage Disequilibrium (LD) between adjacent SNPs. In this paper, we propose a novel Bayesian method, named BAM, for simultaneously partitioning SNPs into LD-blocks and detecting genome-wide multi-locus epistatic interactions that are associated with multiple diseases. Experimental results on the simulated datasets demonstrate that BAM is powerful and efficient. We also applied BAM on two GWAS datasets from WTCCC, i.e., Rheumatoid Arthritis and Type 1 Diabetes, and accurately recovered the LD-block structure. Therefore, we believe that BAM is suitable and efficient for the full-scale analysis of multi-disease-related interactions in GWASs.

Key words: disease association study; epistasis; Linkage Disequilibrium (LD) block; Bayesian methods

1 Introduction

Most common diseases, such as hypertension, cancer, diabetes, and heart disease, are resulting from the joint effects of various genetic variants, environmental factors, or their interactions. It is of great interest to identify the genetic risk factors for understanding disease mechanisms to develop effective treatments and improve public health. Genome-Wide Association Study (GWAS) has been proved to be a powerful genomic and statistical inference tool to identify genetic susceptibility on associations between traits of interests and genetic information of unrelated individuals[1,2]. In genetics, many genotype-phenotype association studies have established that Single Nucleotide Polymorphisms (SNPs)[3], a common type of genetic variants, are associated with a variety of diseases[4]. In a case-control study, an SNP is said to be associated with a disease if the genotype distributions at that SNP in cases and controls are different. In addition to test SNPs individually, it has been anticipated that epistatic interactions among SNPs, defined as multiple SNPs jointly associated with a disease, may be responsible for significantly elevating the risks of some human complex diseases. Moreover, there has been major progress in identifying the genetic variants that influence a diverse range of complex human phenotypes. These so-called pleiotropic effects have been found in many diseases, including cardiovascular disease[5,6], neurological disease[7–9], and psychiatric illness[10,11]. In this paper, we consider epistatic interactions as the statistically significant associations of d genetic markers ($d \geqslant 2$) with multiple phenotypes.

The problem of detecting high-order genome-wide epistatic interactions using case-control data has attracted extensive research interests recently. Many

• Guanying Wu and Baohua Xu are with the Dental Center of China-Japan Friendship Hospital, Beijing 100029, China. E-mail: wuguanying2008@163.com; orthodontist_wu@163.com.
• Xuan Guo is with the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA. E-mail: xuan.guo@unt.edu.
∗ To whom correspondence should be addressed.
This work is an extension of Xuan Guo's PhD dissertation at Georgia State University, Athens, GA, USA.
Manuscript received: 2019-10-29; revised: 2019-12-05; accepted: 2020-01-02

computational algorithms have been proposed. These methods can be broadly grouped into three categories: exhaustive search, stepwise search, and heuristic method. Exhaustive search methods enumerate and test all SNPs and their combinations under different statistical models, such as χ^2 test, entropy-based test, and exact likelihood ratio test[12–14]. Stepwise search methods avoid the massive computation burden by selecting a subset of SNPs or SNP combinations based on certain filtering criteria, for example, low-order measurement tests, then extending the SNP modules that pass criteria to higher-order interactions[15, 16]. Heuristic methods usually utilize machine learning techniques or stochastic procedures to restrict the search space of interactions to speed up the detection of genome-wide epistasis[17–19]. More details about the popular GWAS mapping tools can be found in recent surveys[20–24]. All these methods are facing two fundamental challenges in detecting genome-wide interactions related to multiple diseases: the first arises from a heavy computational burden, i.e., the number of tests grows exponentially as the order of interaction goes up. For example, we need to perform around 6.25^{11} statistical tests to detect pairwise interactions for a moderate dataset with 500 000 SNPs. The second challenge is that existing methods are not statistically powerful enough to test high-order multi-locus models of multiple diseases. Because of an enormous number of hypotheses with limited sample size, a high proportion of significant associations is expected to be false positives.

The current primary analysis paradigm for GWAS is dominated by the analysis of the susceptibility of SNPs to one disease a time, which might only explain a small part of genetic causal effects and relations for multiple complex diseases. Most current tools for high-order epistasis detection are only capable of identifying interactions for GWAS data with two groups, i.e., case-control study, except for some recently developed methods, such as SAM, a Jensen-Shannon-divergence-based method[25], and DAM, a Bayesian inference method for multi-locus associations[26]. A limitation of DAM is the model assumes that a Markov chain can capture the dependence structure of the SNPs in the data. Linkage Disequilibrium (LD) is the well-known dependence structure between adjacent SNPs in the human genome[27–29]. Two loci are at LD if they are non-randomly inherited, and they are at Linkage Equilibrium (LE) if the inheritances are independent to each other. One of the applications by applying LD in GWASs is the SNP tagging[30, 31]. An SNP is selected as a tag SNP to represent a region of the genome if it is in high LD with the group of SNPs in that region. Although the SNP data from large GWASs are already using tag SNPs to sample the genetic variants, it is inevitable to have SNPs at strong LD. The LD between adjacent SNPs exhibits a block-wise structure in the human genome. SNPs within LD-blocks are highly correlated, and the correlation is broken down by recombination events at block boundaries. If some genuinely causal SNPs are correlated with nearby SNPs due to the LD effect, most popular association detecting tools would be incapable of telling whether the significance is caused by the interaction or the LD effect. A simple Markov model used in DAM, therefore, cannot capture this block structure when analyzing dense SNPs.

In this paper, we extend DAM to model the block structures (referred as LD-block[32]) and capture the significant associations within and between the inferred blocks. We name our novel method as BAM, Block-based detection of genome-wide Association on Multiple diseases. Experimental results on simulated datasets demonstrate that BAM is capable of recovering perfectly the LD-blocks and identifying complicated embedded associations. We also applied BAM on two real WTCCC datasets[33], and some novel, interesting findings were reported. The paper is organized as follows. In Section 2, we present the extended Bayesian variable partition model with the incorporation of the LD-block model. We design a two-level Markov Chain Monte Carlo (MCMC) updating scheme and a stepwise interaction evaluation in Section 2.5. In Section 3, we demonstrate the superior performance of BAM by simulation studies comparing to two other approaches: DAM and SAM. We also report the results by applying BAM to the Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D) data from WTCCC. We conclude the article in Section 4.

2 Materials and Methods

2.1 Notation

Suppose we have a GWAS dataset with M SNPs, L traits, each trait with N_l samples ($l \in \{1, 2, \ldots, L\}$), and total N samples. Let $\boldsymbol{D}$ denote an $N \times M$ matrix of genotypes for all traits and samples, $\boldsymbol{D}_i$ denote vectors of genotypes observed at SNP i across samples and traits, and $\boldsymbol{D}^{(i)}$ denote an $N_i \times M$ matrix of genotypes for trait i. Note that the superscript inside a pair of parentheses is merely a label and does not represent the exponent.

As shown in Ref. [34], the number of partitions that

a set of L traits can be separated into nonempty subsets is a Bell number[35]. Let Q denote the set of distinct partitions of L traits, Q_i denote partition i, and $Q_i^{(j)}$ denote nonempty subset j in partition i. Each different partition indicates a type of association that an SNP or an SNP module may have with traits. Figure 1 shows an example for a three-trait dataset with 10 association types. For these association types, the first five indicate SNPs are independently associated with traits, and the rest five indicate SNPs are dependently associated with traits. Given that SNPs can be independent or dependent to other SNPs when associated with traits, we have independent and dependent association types. For simplicity, we use the same notation Q_i to denote association type i. When needed, we use $\tilde{Q}$ and $\bar{Q}$ to distinguish independent and dependent association types, respectively. Note that only association types $\bar{Q}_i, i \in \{2,3,4,5\}$, are epistatic interactions. Without loss of generality, we always use Q_1 to denote the association type not associated with any disease trait. The genotypes of m SNPs with association type Q_i in trait subset j are denoted by $\boldsymbol{D}^{(Q^{(j)})_i}$, which is an $N_{Q_i^{(j)}} \times m$ matrix, where $N_{Q_i^{(j)}} = \sum_l^{l \in Q_i^{(j)}} N_l$.

To model the LD effect, we seek to partition the M SNPs into $|B|$ consecutive blocks, where $B = \{B_1, B_2, \ldots, B_{|B|}\}$, and B_i is block i with $|B_i|$ consecutive SNPs. The genotype defined by B_i is denoted by D_{B_i}.

2.2 Assumptions

To make identifying epistasis with LD effects computationally possible, we make the following assumptions:

(1) Each SNP only belongs to one and only one LD-block.

(2) Each epistatic interaction contains two or more than two SNPs, each of which is from different LD-block.

(3) In one LD-block, given the causal SNPs with genotypes distributed differently between cases and controls, the genotypes of the rest SNPs follow a common distribution in both cases and controls.

2.3 Bayesian LD-block partition model

Here, we introduce a Bayesian LD-block partition model for multiple traits without considering disease association. For an LD-block B_k, there are $3^{|B_k|}$ possible genotype combinations. Note that we use genotypes and genotype combinations interchangeably. We assume that the genotype combination g of each sample follows independently from a multinomial distribution with frequency parameters θ_g, where θ_g follows a Dirichlet prior distribution with α_g as a hyper-parameter (i.e., pseudo-counts). More precisely, with n_g denoting the combined count of g observed in $\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}$, we have

$$P(\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}|\Theta) = \prod_{g=1}^{3^{|B_k|}} \theta_g^{n_g},$$

where $\Theta = \{\theta_g\}$ and

$$P(\Theta) = \frac{\Gamma(\alpha\circ)}{\prod_{g=1}^{3^{|B_k|}} \Gamma(\alpha_g)} \prod_{g=1}^{3^{|B_k|}} \theta_g^{\alpha_g - 1},$$

Association type	$\tilde{Q}_1$	$\tilde{Q}_2$	$\tilde{Q}_3$	$\tilde{Q}_4$	$\tilde{Q}_5$	$\bar{Q}_1$		$\bar{Q}_2$		$\bar{Q}_3$		$\bar{Q}_4$		$\bar{Q}_5$	
SNP ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Case 1	0	2	0	2	1	0	0	2	1	0	0	2	1	1	0
	1	2	1	2	1	1	1	2	1	1	1	2	1	1	0
	2	2	2	2	1	2	2	2	1	2	2	2	1	1	0
	0	2	0	2	1	0	0	2	1	0	0	2	1	1	0
	0	2	0	2	1	0	0	2	1	0	0	2	1	1	0
Case 2	0	0	2	2	2	0	0	0	0	2	1	2	1	2	1
	1	1	2	2	2	1	1	1	1	2	1	2	1	2	1
	2	2	2	2	2	2	2	2	2	2	1	2	1	2	1
	0	0	2	2	2	0	0	0	0	2	1	2	1	2	1
	0	0	2	2	2	0	0	0	0	2	1	2	1	2	1
Control	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Genotype combination Probability distribution

Probability distribution 1

Probability distribution 2

Probability distribution 3

Fig. 1 Illustration of ten association types in a dataset with three phenotypic traits.

where $\alpha\circ$ denotes the sum of values over all elements in $\{\alpha_g\}$. By integrating out $\{\theta_g\}$, we can obtain the marginal probability of the data in $D_{B_k}^{(Q_i^{(j)})}$ as

$$P(\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}|B_k) = \int_{\Theta} P(\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}|B_k,\Theta)P(\Theta)\mathrm{d}\Theta = \frac{\Gamma(\alpha\circ)}{\prod_{g=1}^{3^{|B_k|}}\Gamma(\alpha_g)}\int_{\Theta}\prod_{g=1}^{3^{|B_k|}}\theta^{n_g+\alpha_g-1}\mathrm{d}\Theta = \left(\prod_{g=1}^{3^{|B_k|}}\frac{\Gamma(n_g+\alpha_g)}{\Gamma(\alpha_g)}\right)\times\frac{\Gamma(\alpha\circ)}{\Gamma(N_{Q_i^{(j)}}+\alpha\circ)} \quad (1)$$

In our implementation, we set $\alpha_g = \frac{\varphi}{3^{|B_k|}}$ for every genotype combination g, and let $\varphi = 1.5$. Note that Eq. (1) can be used to model any $Q_i^{(j)}$ with two or more than two traits by counting genotypes across traits in $Q_i^{(j)}$. Further assuming independence between LD-blocks and independence between subsets of traits in a partition, we obtain the probability function $P(\boldsymbol{D}|B)$ of genotypes of all LD-blocks, which is expressed as the product of individual block probabilities as defined in Eq. (1). For example, assume all LD-blocks with the same association type Q_i, then we have

$$P(\boldsymbol{D}|B,Q_i) = \prod_k\left(\prod_j P\left(\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}|B_k\right)\right).$$

2.4 Bayesian association inference based on LD-block partition model

The SNPs in an LD-block can be broadly categorized into two groups: the SNPs truly associated with the diseases, and the rest SNPs not associated with the disease given the first group of SNPs. In this section, we first propose a Bayesian model of disease associations with only SNPs in one LD-block, and then extend it to address epistatic interactions.

Based on our notations described in Section 2.1, we know that the genotypes of SNPs with association types $Q_i (i \neq 1)$ are truly associated with diseases, and the genotypes of SNPs with association type Q_1 are not associated with diseases given $\boldsymbol{D}^{(Q_i)}$. We can infer that the genotypes $\boldsymbol{D}^{(Q_i)}(i \neq 1)$ are distributed differently and are modeled by $|Q_i|$ different distributions for each subset of L traits in a partition. Conditional on the genotypes of SNPs $\boldsymbol{D}^{(Q_i)}(i \neq 1)$, the genotypes of SNPs in $\boldsymbol{D}^{(Q_1)}$ follow a common distribution for each trait. The joint probability of the data in one LD-block can, therefore, be expressed as

$$P(\boldsymbol{D}_{B_k}) = P(\boldsymbol{D}_{B_k}^{(Q_1)}|\boldsymbol{D}_{B_k}^{(Q_{-1})})P(\boldsymbol{D}_{B_k}^{(Q_{-1})}) \quad (2)$$

where Q_{-1} denotes the association types other than Q_1, and the genotypes with association types Q_{-1} are denoted by $\boldsymbol{D}^{(Q_{-1})}$. In each subset j of a partition of L traits, for each $Q_i \in Q_{-1}$, the genotypes $\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}$ are modeled by a multinomial Dirichlet distribution as specified in Eq. (1), and by assuming independence between SNPs with different association types, we obtain

$$P(\boldsymbol{D}_{B_k}^{(Q_{-1})}) = \prod_{i\neq 1}\left(\prod_j P\left(\boldsymbol{D}_{B_k}^{(Q_i^{(j)})}\right)\right).$$

To model $P(\boldsymbol{D}_{B_k}^{(Q_1)}|\boldsymbol{D}_{B_k}^{(Q_{-1})})$, we treat all SNPs with association types Q_{-1} jointly and combine the genotypes of SNPs with association type Q_1 in all cases and controls together. These genotypes are not directly associated with any diseases given $\boldsymbol{D}^{(Q_{-1})}$, and thus have the same conditional distributions in cases and controls. Conditional on each possible genotype combination of SNPs with association types Q_{-1}, we model the conditional genotype combination distribution of SNPs with association type Q_1 again by a multinomial-Dirichlet distribution. Thus, we derive the following expression:

$$P(\boldsymbol{D}_{B_i}^{(Q_1)}|\boldsymbol{D}_{B_i}^{(Q_{-1})}) = \frac{P(\boldsymbol{D}_{B_i}^{(Q_1,Q_{-1}\to Q_1)})}{P(\boldsymbol{D}_{B_i}^{(Q_{-1}\to Q_1)})},$$

where $Q_{-1}\to Q_1$ means temporally changing the association types to Q_1 for those SNPs originally with Q_{-1} and changing back when we finish the probability calculation for $P(\boldsymbol{D}_{B_i}^{(Q_{-1}\to Q_1)})$.

To incorporate epistatic interactions in Eq. (2), we model the SNPs with the same dependent association type $\bar{Q}_i$ to be jointly associated with the traits. By assuming independence of SNPs with different association types, we obtain

$$P(\boldsymbol{D}^{(\bar{Q})}) = \prod_i P(\boldsymbol{D}^{(\bar{Q}_i)}),$$

where $\boldsymbol{D}^{(\bar{Q}_i)}$ denotes the genotypes of the SNPs with association type $\bar{Q}_i$ across LD-blocks. We introduce a latent M-dimensional indicator variable $\boldsymbol{I}$ to represent the association types for M SNPs. Given a particular block partition B and an association type indication $\boldsymbol{I}$, we obtain the joint probability function of the entire data as

$$P(\boldsymbol{D}|B,\boldsymbol{I}) = P(\boldsymbol{D}^{(\tilde{Q}_{-1})}|B,\boldsymbol{I})\times \prod_{l:I_l=\tilde{Q}_i,i\neq 1} P(\boldsymbol{D}_l^{(\tilde{Q}_i)}|B,\boldsymbol{I})\times \prod_k P(\boldsymbol{D}_{B_k}^{(Q_1)}|\boldsymbol{D}_{B_k}^{(Q_{-1})},B,\boldsymbol{I}) \quad (3)$$

Finally, the Bayesian association inference model based on LD-block model for the data $\boldsymbol{D}$, the block variable B, and the association type indicator $\boldsymbol{I}$ is written as

$$P(\boldsymbol{D},\boldsymbol{I},B) = P(\boldsymbol{D}|\boldsymbol{I},B)P(\boldsymbol{I})P(B) \quad (4)$$

where the conditional distribution $P(\boldsymbol{D}|\boldsymbol{I},B)$ is calculated in Eq. (3). We then obtain the posterior distribution of $\boldsymbol{I}$ and B as

$$P(\boldsymbol{I},B|\boldsymbol{D}) \propto P(\boldsymbol{D}|\boldsymbol{I},B)P(\boldsymbol{I})P(B) \quad (5)$$

The prior distribution of the association type indicator $\boldsymbol{I}$ is set as a product of independent multinomial distributions, $P(\boldsymbol{I}) = \prod_{i=1}^{|\tilde{Q}|}\phi_{\tilde{Q}_i}^{|l:I_l=\tilde{Q}_i|}\times \prod_{i=1}^{|\bar{Q}|}\psi_{\bar{Q}_i}^{|l:I_l=\bar{Q}_i|}$, where $\phi_{\tilde{Q}}$ and $\psi_{\bar{Q}}$ denote the prior probability of each SNP belonging to independent association type and dependent association type, respectively. By default, we set $\phi_{\tilde{Q}_{-1}} = \psi_{\bar{Q}_{-1}} = 5/M$ and $\phi_{\tilde{Q}_1} = \psi_{\bar{Q}_1} = (1-(\sum_{i\neq 1}\phi_{\tilde{Q}_i})-(\sum_{i\neq 1}\psi_{\bar{Q}_i}))/2$. That is, we assume a priori that there are 5 SNPs associated with diseases for each association type. Increasing this prior may identify additional SNPs of moderate to low effect sizes. We imposed a restriction that the maximum number of SNPs with epistatic association types must be smaller than $\log_3(N/10)$ to avoid overfitting epistasis mapping. Similar to the association type indicator vector $\boldsymbol{I}$, the block variable B in our model contains M binary indicators corresponding to M SNPs. An element in B is equal to 1 if the corresponding SNP is the start position of an LD-block, and 0 otherwise. The prior distribution of the block variable B is set as the product of $|B|$ independent Bernoulli probabilities $P(B) = \rho^{|B|}(1-\rho)^{M-|B|}$. Based on the block distributions estimated in European and Asian populations by Zhang et al.[32] and Gabriel et al.[36], we assume that there are 50 000 LD-blocks in the human genome, thus we set $\rho = 50\,000/(3\times 10^9) = 1.67\times 10^{-5}$, where 3×10^9 is the length of the human genome. A smaller value of ρ will result in less blocks with larger block size, and a larger ρ will lead to identify more blocks with smaller block size. We also imposed a restriction that the maximum number of SNPs in an LD-block must be smaller than $\log_3(N/10)$ to avoid overfitting the blocks, where N is the total number of samples in the data. Note that the effects of the prior choices for B and $\boldsymbol{I}$ will diminish as the sample size increases.

2.5 MCMC sampling

We apply the Metropolis-Hastings (MH) algorithm[37] to sample the indicators $\boldsymbol{I}$ and B from the distribution defined by Eq. (5). According to the prior $P(B)$, we first initialize B randomly, then use the MH algorithm to construct an MCMC to update B. To explore all possible LD-block partitions, we propose three MH updates: (1) randomly select an LD-block and split it into two new LD-blocks, (2) randomly select two adjacent LD-blocks to merge into one LD-block, and (3) randomly select two adjacent LD-blocks and shift their shared boundary by a random position. In our implementation, we set the three types of MH updates with probabilities 0.1, 0.1, and 0.8, respectively, and require each LD-block to contain at least one SNP. The update is accepted based on the MH ratio, which is a Gamma function. When updating B, simultaneously, we use a method that mixes the MH algorithm and a Gibbs sampler[38] to update the association type indicator $\boldsymbol{I}$. In each MCMC iteration, a Gibbs sampler is used to update the association type of each SNP by calculating the posterior distribution of $\boldsymbol{I} = \{\tilde{Q},\bar{Q}\}$ given all other model parameters and the data. We also propose an MH update to switch the association types of two SNPs and accept the update based on MH ratios. In each MCMC iteration, we first run the Gibbs sampler to update the association types for all SNPs once, and then run the MH algorithm to switch two SNPs with different association types. We set a minimum threshold on the posterior probabilities of $\boldsymbol{I}$ and use those SNPs with posterior probabilities larger than the threshold as candidates in the interaction evaluation process described in the next section.

2.6 Evaluation of epistasis interaction

With the candidate SNPs generated by MCMC, we apply the χ^2 statistic and the conditional χ^2 test to measure the significance for a module of SNPs. Let $A = (x_1,\ldots,x_d:Q_i)$ denote an SNP module that contains d SNPs with association type Q_i. We use $\chi^2(x_1,\ldots,x_d:Q_i)$ to denote the χ^2 statistic of A and $\chi^2(x_1,\ldots,x_d|x_{c_1},\ldots,x_{c_{d'}}:Q_i)$ to denote the conditional χ^2 statistic given a subset $A' = \{x_{c_1}, x_{c_2},\ldots,x_{c_{d'}}\}$ of A with d' SNPs. The χ^2 statistic is calculated as

$$\chi^2(x_1,\ldots,x_d:Q_i)=\sum_{j=1}^{|Q_i|}\sum_{g=1}^{3^d}\frac{(n_{j,g}-e_{j,g})^2}{e_{j,g}} \quad (6)$$

where $n_{j,g}$ denotes the frequency of genotype combination g in a trait subset j, and $e_{j,g}$ denotes the corresponding expected frequency of genotype combination g. The degree of freedom for Eq. (6) is $(|Q_i|-1)\times(3^d-1)$.

The conditional independent test based on the χ^2 statistic is defined as follows:

$$\chi^2(x_1,\ldots,x_d|x_{c_1},\ldots,x_{c_{d'}}:Q_i)=\sum_{\iota=1}^{3^{d'}}\sum_{j=1}^{|Q_i|}\sum_{g=1}^{3^{d-d'}}\frac{(n_{j,g}^{(\iota)}-e_{j,g}^{(\iota)})^2}{e_{j,g}^{(\iota)}} \quad (7)$$

The degree of freedom for Eq. (7) is $3^{d'}\times(|Q_i|-1)\times(3^{d-d'}-1)$. We treat those SNPs as redundant SNPs when they are conditionally independent by giving a subset of the SNPs in one module. To avoid these redundant SNPs, we define an SNP module ($d\geqslant 2$) as a compact, significant epistatic interaction by the following definition.

Definition 1 An SNP module $A=(x_1,\ldots,x_d:Q_i)$ is considered as a significant, compact epistatic interaction by giving a significant level α_d, if it meets the following three conditions:

(1) the p-value of $\chi^2(x_1,\ldots,x_d:Q_i)=$ the minimum p-value of $\chi^2(x_1,\ldots,x_d:Q)$, $\forall Q_i$;

(2) the p-value of $\chi^2(x_1,\ldots,x_d:Q_i)\leqslant\alpha_d$;

(3) the p-value of $\chi^2(x_1,\ldots,x_d|x_{c_1},\ldots,x_{c_{d'}}:Q_i)\leqslant\alpha_d$, for $\forall A'=\{x_{c_1},\ldots,x_{c_{d'}}\}$.

Based on Definition 1, we develop a stepwise procedure to search top-f d-locus significant, compact interactions, where the search space only includes the SNPs yielded by MCMC sampling. Here, f is a user-defined number. We restrict that each SNP can only present in one significant, compact interaction. We first search all the modules with just one SNP based on Definition 1. All SNPs with significant marginal associations after a Bonferroni correction are reported in a list L. And then, we recursively test all the possible combinations of SNPs in L by increasing the module size by one at a time until the module size reaches a user pre-set value. Bonferroni correction is applied to both independent associations and epistatic associations.

3 Experiment and Result

To show the performance of BAM with consideration of LD effects, we first introduce the simulation using HapMap data[39]. We use the null simulation to test the Type I error rate of BAM. We compared BAM to two other methods (i.e., DAM[26,34] and SAM[25]) that can find epistatic interactions associated with multiple diseases. In this comparison, we use a set of simulated datasets in which four types of known disease-specific associations were embedded in each dataset. We also apply BAM to two real WTCCC datasets, RA and T1D, and BAM produced accurate block partitions that match well to the visual blocks displayed by Haploview[40].

3.1 Experimental design

Data simulation: To evaluate the effectiveness of BAM, we performed extensive simulation experiments using four disease models of two-locus associations. We set the number of traits to be three. The sample sizes for three traits were set to $(800,800,800)$ or $(1600,1600,1600)$, where we consider the first two groups as the case groups and the third one as the control group. As the example shown in Fig. 1, we have ten different association types, and four of them are epistatic interactions associated with diseases. We labeled these four types of epistatic interactions as $\bar{Q}_2$, $\bar{Q}_3$, $\bar{Q}_4$, and $\bar{Q}_5$. We used four disease models (Table 1) to simulate epistatic interactions. These four disease models are identical to the first four models in Ref. [26]. $\bar{Q}_2$, $\bar{Q}_3$, and $\bar{Q}_4$ only need one disease model to simulate epistatic interactions since we only need two different genotype distributions. $\bar{Q}_5$ needs three distributions, thus we use four pairs of models (i.e., Models 1 & 2, Models 2 & 3, Models 3 & 4, and Models 4 & 1) to simulate epistatic interactions for $\bar{Q}_5$. For each epistatic interaction, we varied the Minor Allele Frequency (MAF) ranged in

Table 1 Odds tables of disease Models 1 – 4.

		BB	Bb	bb
Model 1	AA	μ	μ	μ
	Aa	μ	$\mu(1+\sigma)$	$\mu(1+\sigma)^2$
	aa	μ	$\mu(1+\sigma)^2$	$\mu(1+\sigma)^4$
Model 2	AA	μ	$\mu(1+\sigma)$	$\mu(1+\sigma)$
	Aa	$\mu(1+\sigma)$	μ	μ
	aa	$\mu(1+\sigma)$	μ	μ
Model 3	AA	μ	μ	$\mu(1+\sigma)$
	Aa	μ	$\mu(1+\sigma)$	μ
	aa	$\mu(1+\sigma)$	μ	μ
Model 4	AA	μ	$\mu(1+\sigma)$	μ
	Aa	$\mu(1+\sigma)$	μ	$\mu(1+\sigma)$
	aa	μ	$\mu(1+\sigma)$	μ

Note: The disease prevalence $p(\boldsymbol{D})$, the genetic heritability h^2, and the MAFs determine the parameters (μ and σ) as in Ref. [41]. In the simulation, we set $p(\boldsymbol{D})=0.1$ for all models, $h^2=0.03$ for Model 1, $h^2=0.02$ for Models 2 – 4, and the MAFs of disease-associated SNPs to be 0.1, 0.2, and 0.4.

$\{0.1, 0.2, 0.4\}$. We generated 100 replicas per MAF setting. Note that there are only three types of epistatic interactions for $\bar{Q}_5$ when MAF $= 0.1$, because the other model pairs do not have feasible solutions for μ and σ. In total, we generated 92 epistatic interactions. Each simulation replica contained $M = 1000$ SNPs and an epistatic interaction of two SNPs. To mimic real genetic data in human populations, we select an arbitrary region in Chromosome 22. Given the genotypes of two SNPs in an epistatic interaction, we randomly sampled up to 4800 individuals from a pool of controls generate by HAPGEN[42] using HapMap European samples[39] with odds ratio $= 1$. The genotypes of unassociated SNPs are sampled from that control pool with the ordering of these SNPs kept the same as in Chromosome 22. We applied a quality control procedure by removing SNPs with MAF < 0.05 or p-values less than 0.001 from the Hardy-Weinberg Equilibrium test.

Statistical power: In the evaluation of performances on simulated data, for each setting, we generated 100 datasets, each of which contains one epistatic interaction. We define the measurement of discrimination power as the fraction of 100 datasets on which the ground-truth epistatic interactions are ranked as the top one and passed the significance threshold by the method of interest.

3.2 Null simulation to test Type I errors

To evaluate the Type I errors of BAM, we conducted the null simulation experiments. We simulated 1000 datasets without disease association embedded. The false-positive rate of BAM is shown in Fig. 2. By setting the significance level to 0.1, BAM always obtained low false-positive rates under the different sample sizes and the numbers of SNPs. Under most of the situations, BAM did not report any significant interaction. For the situations with zero false-positive rate, we found that there is no significant association with the posterior probability larger than the threshold 0.5.

3.3 Simulation experiments for four epistatic association types

To verify the effectiveness of BAM, we used the simulation strategy introduced in Section 3.1, where the LD-block structure is from a real human genome. Although the detailed boundary information is unavailable, the block structure widely exists across the entire human genome. The performance comparison between BAM, DAM, and SAM is shown in Fig. 3. We can find that BAM achieves better performance

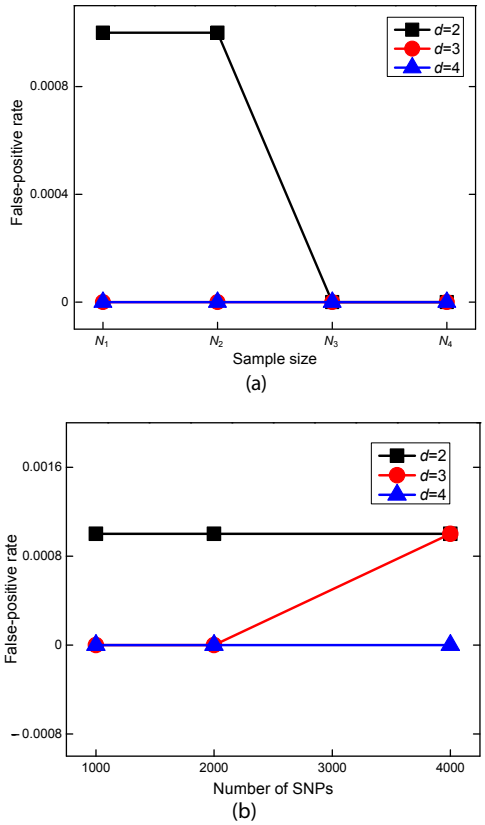

Fig. 2 False-positive rates of BAM under null models. The plots in (a) and (b) show the false-positive rates of BAM for different d when the sample sizes and the numbers of SNPs vary. In (a), $N_1 = \{200, 200, 400\}$, $N_2 = \{400, 400, 800\}$, $N_3 = \{800, 800, 1600\}$, and $N_4 = \{1600, 1600, 3200\}$.

when sample size increases from 2400 to 4800. For Models 1 & 3, BAM had steady power regardless of MAF. For Models 2 & 4, the power of BAM decreased as MAF increased. By examining the composition of heritability of these four models that also used in Ref. [26], we found that Models 2 & 4 have the same amount of heritability, but the main effects decrease as MAF goes up. Thus, the performance of BAM is positively related to the amount of main effects in the epistatic interactions. For interaction $\bar{Q}_5$ that has distinct effects on each trait, BAM obtained higher power than the other two methods regardless of the sample size. To show the overall performance for all four epistasis types, we used a metric called overall quality introduced in Ref. [26] to evaluate the overall performance of these three methods. The overall quality is defined by $n_{\text{correct}}/n_{\text{total}}$, where n_{correct} is the number of datasets where the method successfully detects the ground-truth interaction and n_{total} is the total number of datasets. The overall qualities of BAM, SAM, and DAM are 0.564,

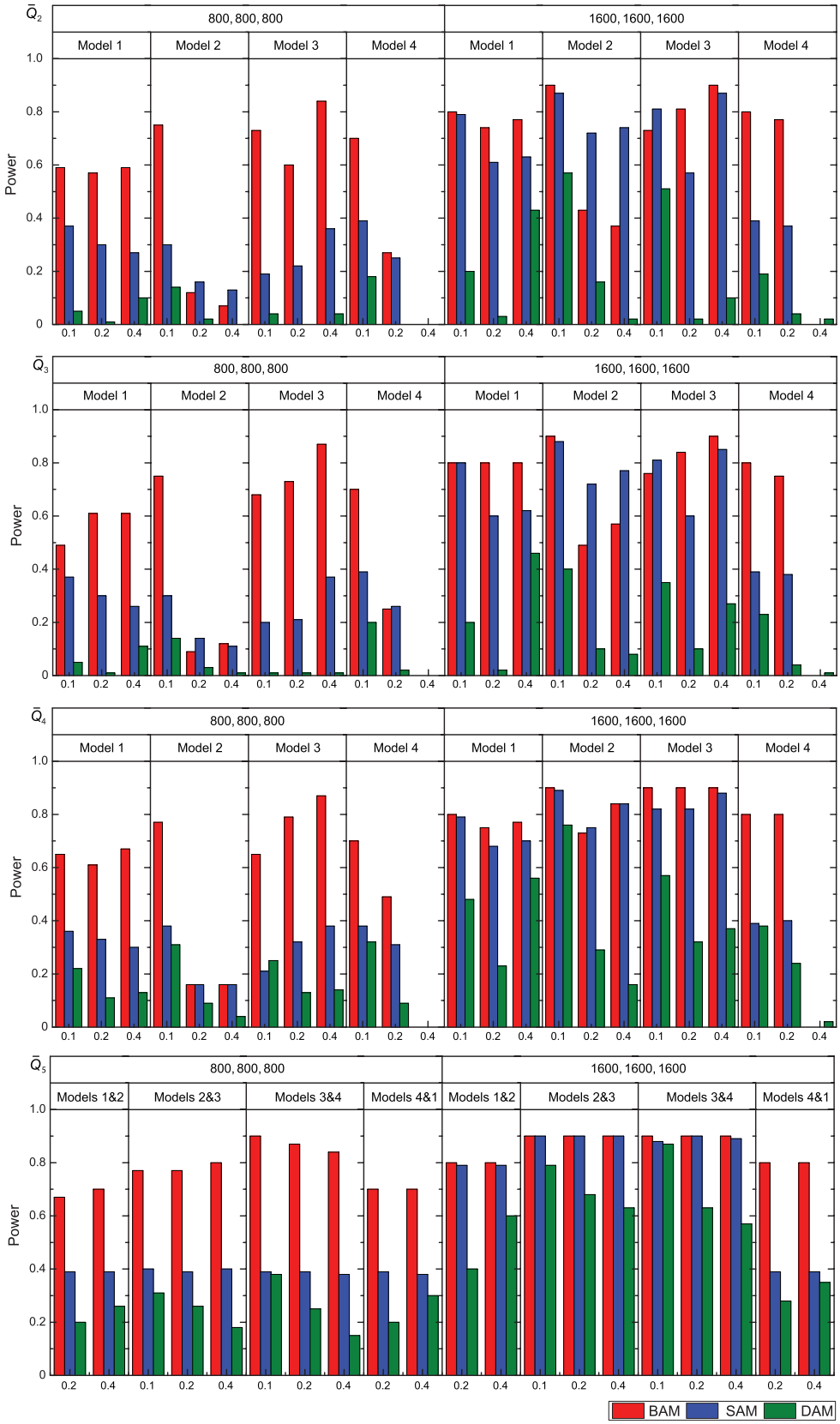

Fig. 3 Performance comparison between BAM, SAM, and DAM on the datasets with two-locus epistatic interactions. The *x*-axis shows the MAF value.

0.283, and 0.119 for the datasets with 2400 individuals, and 0.742, 0.663, and 0.321 for the datasets with 4800 individuals, respectively. Compared to DAM, BAM showed greatly power improvement with LD effects into consideration. We also found that SAM performed better than DAM when LD-block structures existed. When comparing the candidate SNP lists generated by SAM and DAM, we found that DAM had difficulty to select the disease-associated SNPs in the blocks. In contrast, SAM uses a clustering algorithm to group SNPs with different genotype distributions in different traits. So, SAM tends to place causal SNPs and their nearby SNPs into different clusters, and the disease-associated SNPs may have a chance to be evaluated in the following tests.

3.4 Experiments on the WTCCC data

We also applied BAM to two real GWAS datasets (i.e., RA and T1D) to examine the performance of BAM on LD-block structure estimation. Here, we focused on the results from Chromosome 6. The block structures due to LD effects have many different definitions. To give an intuitive idea of the performance of the block structure yielded by BAM, we randomly selected four regions with lengths about 200 kb and used HapMap to draw the corresponding Haploviews of these regions. Figure 4 shows the recovered block structures by BAM and the corresponding Haploviews. We set the number of iterations to 300 with an extra 300 in the burn-in process. Comparing to the visual blocks from Haploview, BAM produced relatively accurate LD-block partitions for these four regions.

3.5 Computational efficiency

We used two simulation datasets to evaluate the performance of computation time for BAM. Both datasets contain three traits, each with 2000 samples. The first dataset comprises 1000 SNPs, and the second one comprises 10^4 SNPs. The experiments were conducted on a Unix system powered by 2.60 GHz Intel Xeon Silver 4112 with 32 GB RAM. We run the ten independent MCMC chains for each dataset. The averaged computation time of 10 chains are 14 minutes and 10 hours, with the numbers of iteration set to 10^6 and 10^7 for these two datasets, respectively. Similar to DAM, BAM uses the boolean representation introduced by Wan et al.[12] BAM used about 21 MB memory for the second dataset. The memory usage of BAM is linear to the number of SNPs and sample size. For a typical GWAS dataset with 5×10^5 SNPs and 6000 samples, the estimation of memory consumption is around 1 GB. Therefore, memory is not a problem for BAM, given the current popular hardware configuration.

4 Conclusion

In this paper, we proposed a block-based Bayesian method, called BAM, for LD-block inference and detection of high-order epistatic interaction on multiple diseases. Extensive experimental results on simulated datasets indicate that BAM can identify embedded disease associations even with block structures present. By comparing to two other methods, i.e., SAM and DAM, we showed that BAM substantially improved the statistical power by incorporating the LD model to the Bayesian inference model. The results from experiments in human Chromosome 6 showed that BAM is capable of recovering the block structures accurately.

When using BAM on real data, we suggest applying the quality control procedures as presented in other recent studies[41,43] because sequencing bias and genotyping bias could confound BAM and lead to false-positive disease-related SNPs. For example, the coverage bias that is caused by the sequencing platforms may generate SNPs with low, uneven coverage between case and control groups. Quality control is necessary to remove those unreliable SNPs. Other issues, such as the branch effects across multiple studies, may lead to population stratification that causes the p-values of disease-unrelated SNPs inflated[44]. Therefore, in addition to BAM, we suggest employing other methods, such as those based on linear mixed model, to adjust the p-value of reported SNPs by BAM.

References

[1] H. Sabaa, Z. Cai, Y. Wang, R. Goebel, S. Moore, and G. Lin, Whole genome identity-by-descent determination, *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 2, p. 1350002, 2013.

[2] Y. Wang, Z. Cai, P. Stothard, S. Moore, R. Goebel, L. Wang, and G. Lin, Fast accurate missing SNP genotype local imputation, *BMC Research Notes*, vol. 5, no. 1, p. 404, 2012.

[3] Y. He, Z. Zhang, X. Peng, F. Wu, and J. Wang, De novo assembly methods for next generation sequencing data, *Tsinghua Science and Technology*, vol. 18, no. 5, pp. 500–514, 2013.

[4] K. Peter and D. J. Hunter, Genetic risk prediction: Are we there yet? *The New England Journal of Medicine*, vol. 360, no. 17, pp. 1701–1703, 2009.

[5] M. Nikpay, A. Goel, H. H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C.

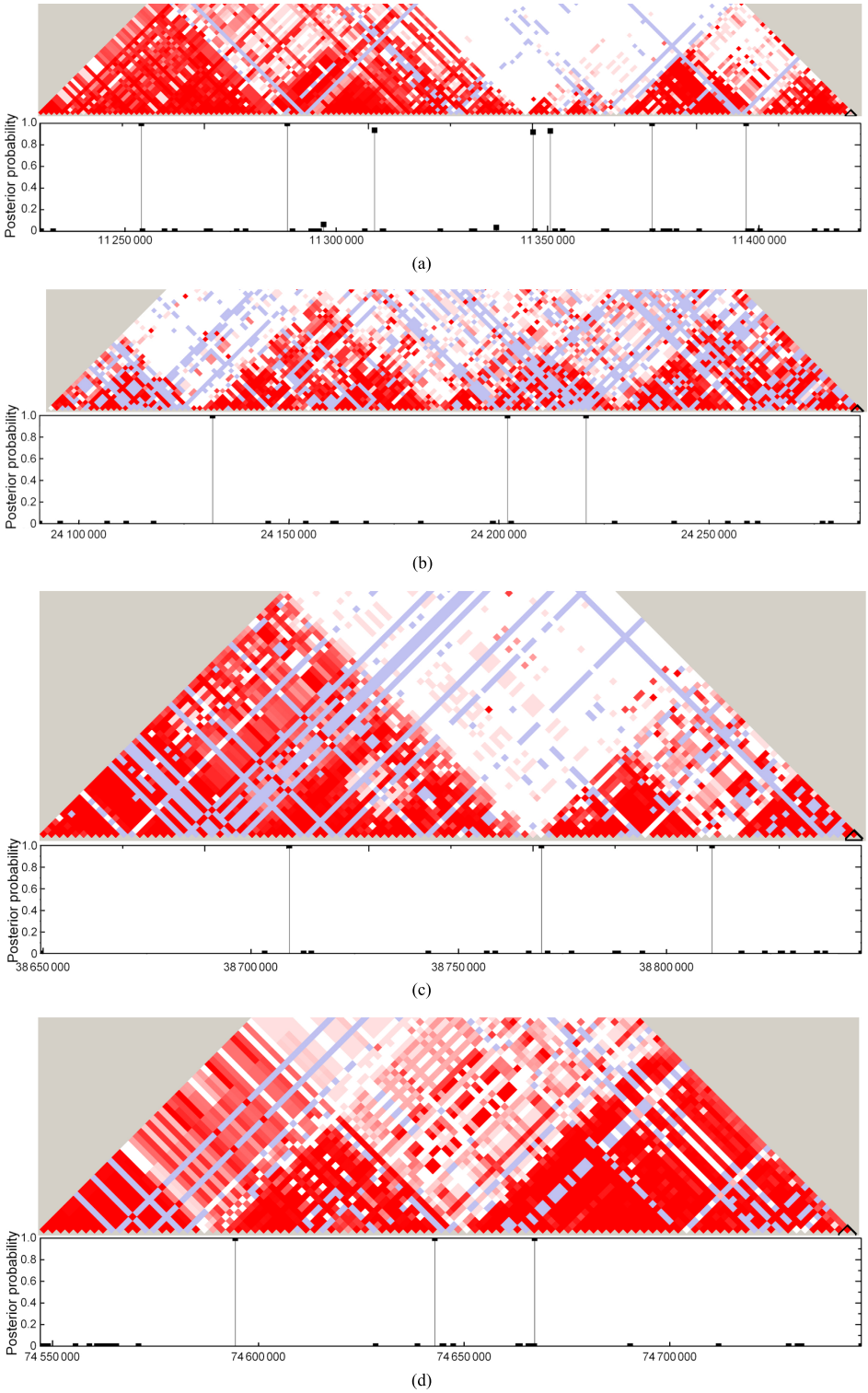

Fig. 4 Four block structures recovered by BAM in Chromosome 6. The top half figure is the Haploview. The *x*-axis in the bottom half figure is the physical locations of the SNPs, and the *y*-axis is the posterior probability of SNPs.

Hopewell, et al., A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease, *Nature Genetics*, vol. 47, no. 10, p. 1121, 2015.

[6] H. Schunkert, I. R. König, S. Kathiresan, M. P. Reilly, T. L. Assimes, H. Holm, M. Preuss, A. F. Stewart, M. Barbalic, C. Gieger, et al., Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease, *Nature Genetics*, vol. 43, no. 4, p. 333, 2011.

[7] J. C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis,

G. W. Beecham, et al., Meta-analysis of 74 046 individuals identifies 11 new susceptibility loci for alzheimer's disease, *Nature Genetics*, vol. 45, no. 12, p. 1452, 2013.

[8] C. Sun, Q. Li, L. Cui, H. Li, and Y. Shi, Heterogeneous network-based chronic disease progression mining, *Big Data Mining and Analytics*, vol. 2, no. 1, pp. 25–34, 2018.

[9] W. Van Rheenen, A. Shatunov, A. M. Dekker, R. L. McLaughlin, F. P. Diekstra, S. L. Pulit, R. A. van der Spek, U. Võsa, S. de Jong, M. R. Robinson, et al., Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis, *Nature Genetics*, vol. 48, no. 9, p. 1043, 2016.

[10] S. Ripke, N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman, G. Breen, E. M. Byrne, D. H. Blackwood, D. I. Boomsma, S. Cichon, et al., A mega-analysis of genome-wide association studies for major depressive disorder, *Molecular Psychiatry*, vol. 18, no. 4, p. 497, 2013.

[11] P. Sklar, S. Ripke, L. J. Scott, O. A. Andreassen, S. Cichon, N. Craddock, H. J. Edenberg, J. I. Nurnberger, M. Rietschel, D. Blackwood, et al., Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4, *Nature Genetics*, vol. 44, no. 9, p. 1072, 2012.

[12] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. Tang, and W. Yu, Detecting two-locus associations allowing for interactions in genome-wide association studies, *Bioinformatics*, vol. 26, no. 20, pp. 2517–2525, 2010.

[13] L. S. Yung, C. Yang, X. Wan, and W. Yu, Gboost: A GPU-based tool for detecting gene–gene interactions in genome–wide case control studies, *Bioinformatics*, vol. 27, no. 9, pp. 1309–1310, 2011.

[14] Y. Liu, H. Xu, S. Chen, X. Chen, Z. Zhang, Z. Zhu, X. Qin, L. Hu, J. Zhu, G. P. Zhao, et al., Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases, *PLoS Genetics*, vol. 7, no. 3, p. e1001338, 2011.

[15] J. Marchini, P. Donnelly, and L. R. Cardon, Genome-wide strategies for detecting multiple loci that influence complex diseases, *Nature Genetics*, vol. 37, no. 4, p. 413, 2005.

[16] J. Li, A novel strategy for detecting multiple loci in genome-wide association studies of complex diseases, *International Journal of Bioinformatics Research and Applications*, vol. 4, no. 2, p. 150, 2008.

[17] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. Tang, and W. Yu, Predictive rule inference for epistatic interaction detection in genome-wide association studies, *Bioinformatics*, vol. 26, no. 1, pp. 30–37, 2009.

[18] B. Liu, S. Feng, X. Guo, and J. Zhang, Bayesian analysis of complex mutations in hbv, hcv, and hiv studies, *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 145–158, 2019.

[19] Y. Zhang and J. S. Liu, Bayesian inference of epistatic interactions in case-control studies, *Nature Genetics*, vol. 39, no. 9, p. 1167, 2007.

[20] X. Guo, N. Yu, F. Gu, X. Ding, J. Wang, and Y. Pan, Genome-wide interaction-based association of human diseases—a survey, *Tsinghua Science and Technology*, vol. 19, no. 6, pp. 596–616, 2014.

[21] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, 10 years of GWAS discovery: Biology, function, and translation, *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.

[22] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, A survey about methods dedicated to epistasis detection, *Frontiers in Genetics*, vol. 6, p. 285, 2015.

[23] Y. J. Wen, H. Zhang, Y. L. Ni, B. Huang, J. Zhang, J. Y. Feng, S. B. Wang, J. M. Dunwell, Y. M. Zhang, and R. Wu, Methodological implementation of mixed linear models in multi-locus genome-wide association studies, *Briefings in Bioinformatics*, vol. 19, no. 4, pp. 700–712, 2017.

[24] X. Ding and X. Guo, A survey of SNP data analysis, *Big Data Mining and Analytics*, vol. 1, no. 3, pp. 173–190, 2018.

[25] X. Guo, Searching genome-wide disease association through SNP data, PhD dissertation, Georgia State University, Athens, GA, USA, 2015.

[26] X. Guo, J. Zhang, Z. Cai, D. Z. Du, and Y. Pan, Searching genome-wide multi-locus associations for multiple diseases based on bayesian inference, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* , vol. 14, no. 3, pp. 600–610, 2017.

[27] T. Berisa and J. K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations, *Bioinformatics*, vol. 32, no. 2, p. 283, 2016.

[28] S. Gazal, H. K. Finucane, N. A. Furlotte, P. R. Loh, P. F. Palamara, X. Liu, A. Schoech, B. Bulik-Sullivan, B. M. Neale, A. Gusev, et al., Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection, *Nature Genetics*, vol. 49, no. 10, p. 1421, 2017.

[29] Y. Cheng, H. Sabaa, Z. Cai, R. Goebel, and G. Lin, Efficient haplotype inference algorithms in one whole genome scan for pedigree data with non-genotyped founders, *Acta Mathematicae Applicatae Sinica, English Series*, vol. 25, no. 3, pp. 477–488, 2009.

[30] Z. Liu and S. Lin, Multilocus LD measure and tagging SNP selection with generalized mutual information, *Genetic Epidemiology*: *The Official Publication of the International Genetic Epidemiology Society*, vol. 29, no. 4, pp. 353–364, 2005.

[31] Z. Cai, H. Sabaa, Y. Wang, R. Goebel, Z. Wang, J. Xu, P. Stothard, and G. Lin, Most parsimonious haplotype allele sharing determination, *BMC Bioinformatics*, vol. 10, no. 1, p. 115, 2009.

[32] Y. Zhang, J. Zhang, and J. S. Liu, Block-based bayesian epistasis association mapping with application to WTCCC type 1 diabetes data, *The Annals of Applied Statistics*, vol. 5, no. 3, p. 2052, 2011.

[33] Wellcome Trust Case Control Consortium, Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls, *Nature*, vol. 447, no. 7145, p. 661, 2007.

[34] X. Guo, J. Zhang, Z. Cai, D. Z. Du, and Y. Pan, DAM: A bayesian method for detecting genome-wide associations on multiple diseases, in *Bioinformatics Research and Applications*. New York, NY, USA: Springer, 2015, pp. 96–107.

[35] E. T. Bell, Exponential numbers, *The American Mathematical Monthly*, vol. 41, no. 7, pp. 411–419, 1934.
[36] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, et al., The structure of haplotype blocks in the human genome, *Science*, vol. 296, no. 5576, pp. 2225–2229, 2002.
[37] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Berlin, Germany: Springer Science & Business Media, 2008.
[38] G. Casella and E. I. George, Explaining the gibbs sampler, *The American Statistician*, vol. 46, no. 3, pp.167–174, 1992.
[39] D. Altshuler and P. Donnelly, A haplotype map of the human genome, *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005.
[40] P. I. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler, Efficiency and power in genetic association studies, *Nature Genetics*, vol. 37, no. 11, pp. 1217–1223, 2005.
[41] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang, and W. Yu, Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies, *The American Journal of Human Genetics*, vol. 87, no. 3, pp. 325–340, 2010.
[42] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes, *Nature Genetics*, vol. 39, no. 7, pp. 906–913, 2007.
[43] X. Guo, Y. Meng, N. Yu, and Y. Pan, Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering, *BMC Bioinformatics*, vol. 15, no. 1, p. 102, 2014.
[44] J. K. Pritchard and N. A. Rosenberg, Use of unlinked genetic markers to detect population stratification in association studies, *The American Journal of Human Genetics*, vol. 65, no. 1, pp. 220–228, 1999.

Guanying Wu received the BS and MS degrees from Peking University Health Science Center in 2002 and 2004, respectively. He is now an associate professor at the Dental Center of China-Japan Friendship Hospital. He is a member of World Society of Lingual Orthodontics, European Society of Lingual Orthodontics, World Federation of Orthodontists, Chinese Stomatological Association, and Beijing Stomatological Association, China. He has published more than 20 technical papers in various prestigious journals.

Xuan Guo received the PhD degree from Georgia State University in 2015. He is currently an assistant professor at the University of North Texas. From 2015 to 2017, he was a postdoctoral research associate at Oak Ridge National Laboratory. His research focuses on big data mining and high-performance computing and their applications in environment, food, and health sectors. He serves many premium conferences and journals as editors, chairs, or TPC members.

Baohua Xu received the PhD degree from Peking University Health Science Center in 1999. He is now a professor at the Dental Center of China-Japan Friendship Hospital. He was a visiting professor at the University of Pennsylvania. He is also a professor at Beijing University of Chinese Medicine and Beijing University of Chemical Technology. He is a member of World Society of Lingual Orthodontics, European Society of Lingual Orthodontics, World Federation of Orthodontists, Chinese Stomatological Association, and Beijing Stomatological Association, China. He has published more than 20 technical papers in various prestigious journals.