IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 9, 2020, accepted January 22, 2020, date of publication January 27, 2020, date of current version February 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969465

SHEIB-AGM: A Novel Stochastic Approach for Detecting High-Order Epistatic Interactions Using Bioinformation With Automatic Gene Matrix in Genome-Wide Association Studies

LIYAN SUN, GUIXIA LIU, AND RONGQUAN WANG
Department of Computational Intelligence, College of Computer Science and Technology, Jilin University, Changchun 130600, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130600, China

Corresponding author: Guixia Liu (liugx@jlu.edu.cn)

This work was supported in part by the U.S. Department of Commerce under Grant BS123456, in part by the National Nature Science Foundation of China under Grant 61772226, Grant 61373051, and Grant 61862056, in part by the Science and Technology Development Program of Jilin Province under Grant 20140204004GX, in part by the Project of Science and Technology Innovation Platform of Computing and Software Science (985 Engineering), in part by the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China, and in part by the Fundamental Research Funds for the Central.

ABSTRACT Detecting epistatic interactions in GWAS (genome-wide association studies) data is of great significance in studying common and complex diseases; however, the ability to detect high-order epistatic interactions in GWAS data is still insufficient. Existing methods are usually used to identify two-order interactions, and they cannot detect a large number of interactions. In this article, we propose a novel stochastic approach named SHEIB-AGM (stochastic approach for detecting high-order epistatic interactions using bioinformation with automatic gene matrix). SHEIB-AGM utilizes bioinformation to construct a gene matrix. In each iteration, it randomly generate a high-order SNP combination based on the gene matrix. SHEIB-AGM utilizes k2 (the Bayesian network scoring criterion) and G-test to detect epistasis in the generated combination and automatically update the gene matrix. We have compared SHEIB-AGM with six other methods, i.e., DECMDR, SNPHarvester, MACOED, AntEpiSeeker, HS-MMGKG and SEE, on simulated data including 108 epistatic models and 17,600 files. The results demonstrate that SHEIB-AGM greatly outperforms the above methods in terms of F-measure and power. We utilized SHEIB-AGM (with and without bioinformation) to analyze a real GWAS dataset from the Wellcome Trust Case Control Consortium. The results indicate that SHEIB-AGM with bioinformation can detect 33.94∼3069.40-times more epistatic interactions. We have found numerous genes and gene pairs that may play an important role in seven complex diseases. Some of them have been found in the CTD database (the Comparative Toxicogenomics Database).

INDEX TERMS Epistasis, genome-wide association studies, single-nucleotide polymorphism.

I. INTRODUCTION

Thanks to the development of high-throughput sequencing technology, it is feasible to measure hundreds of thousands of SNP (single nucleotide polymorphism) [1], [2] genotypes from thousands of individuals. Genome-wide association studies (GWAS) [3]–[8] play a very important role in identifying the causes of common and complex diseases. They aim to detect relationships between SNPs and phenotype (disease status) by analyzing GWAS data. The GWAS data typically contain thousands of samples (diseased samples and normal samples) and hundreds of thousands of SNPs. Many SNPs related to a certain phenotype have been discovered [9]–[14]. To understand the underlying causes of common and complex diseases, considering joint genetic effects (epistasis) across the whole genome is necessary. However, this creates huge computational complexity in the analysis. Epistasis [15]–[20] is a phenomenon in which the effect of an SNP depends on other SNPs. It is widely accepted that complex traits or diseases may be caused by many SNPs. The pathogenic SNPs may show minimal effects individually but strong effects jointly. These are epistatic interactions.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

In recent years, numerous methods have been proposed for detecting epistatic interactions [21]–[30]. MDR (multifactor-dimensionality reduction) [21] is a method for reducing the dimensionality of multilocus information to improve the identification of polymorphism combinations associated with disease risk. MDR is nonparametric and can be utilized to detect high-order epistatic interactions. The original MDR is very time consuming. It can only be used on data containing dozens of SNPs. DECMDR [22] is a method that combines the DE (differential evolution) algorithm with CMDR (classification-based multifactor-dimensionality reduction). It uses the CMDR as a fitness measure to evaluate the solutions in the DE process for scanning the epistatic interactions in GWAS. SNPHarvester [24] is a stochastic search method used to detect epistatic interactions. SNPHarvester greatly reduces the number of SNPs. MACOED [25] is a multi-objective heuristic optimization methodology for detecting epistatic interactions. MACOED combines two complementary evaluation objectives from logistical regression and Bayesian network methods to evaluate SNP combinations. MACOED uses a memory-based multi-objective ACO (ant colony optimization) algorithm. AntEpiSeeker [26] is a two-stage ant colony optimization algorithm. In the first stage, AntEpiSeeker searches SNP combinations of sufficient size using ACO. In the second stage, it conducts an exhaustive search of epistatic interactions within the highly suspected SNP combinations and the reduced set of SNPs with top ranking pheromone levels. HS-MMGKG [29] is also a multi-objective heuristic optimization methodology. It uses harmony search with five objective functions. SEE [30] is a multi-objective evolutionary algorithm that uses eight evolution objectives. Four of these objectives are widely used to measure the relationship between SNP combinations and phenotype in GWAS. The other four objectives are measures of the difference between an SNP combination and its best element. Although a variety of methods have been proposed, the ability to detect epistatic interactions is still insufficient, especially in detecting high-order interactions.

In this work, we propose a novel stochastic approach named SHEIB-AGM (stochastic approach for detecting high-order epistatic interactions using bioinformation with automatic gene matrix). Compared with other methods, SHEIB-AGM has the following main advantages:

1) SHEIB-AGM does not need users to specify the order of the epistatic interactions. It automatically calculates *mo* (maximum order) based on the number of samples in the GWAS data, and *mo* can also be specified by the user. SHEIB-AGM can detect any-order ($\in [2, mo)$) interactions.
2) SHEIB-AGM is a stochastic approach. In each iteration, it randomly generates an SNP combination that contains *mo* SNPs. There is minimal relationship between iterations; thus, SHEIB-AGM is parallelizable. Users can specify the number of threads by setting the *local* parameter.
3) SHEIB-AGM can build a gene matrix (*associated Genes*) based on bioinformation (if provided by users). In each iteration, the SNP combination can be generated based on the gene matrix, and the matrix is updated based on whether an epistatic interaction is found in the generated combination. By using the matrix, the method greatly improves the performance in detecting epistatic interactions by using bioinformation.
4) SHEIB-AGM utilizes k2 (the Bayesian network scoring criterion) to find an epistatic interaction in the generated SNP combination and G-test to determine whether the interaction is significant. Thus, it can detect any-order ($\in [2, mo)$) epistatic interactions.
5) In the implementation of SHEIB-AGM, it utilizes a Boolean representation to save the GWAS data. SHEIB-AGM utilizes logical operations to calculate k2 and G-test based on the representation. Because the gene matrix is symmetrical, to avoid wasting memory, it utilizes an array to save the gene matrix. All these details of the implementation improve the speed and reduce the memory consumption of SHEIB-AGM.

To show the performance of SHEIB-AGM, we have conducted a lot of experiments both on simulated GWAS data and real GWAS data. We have compared SHEIB with DECMDR, SNPHarvester, MACOED, AntEpiSeeker, HS-MMGKG and SEE on 3 simulated datasets including 108 epistatic models and 17600 files. The results indicate that SHEIB-AGM greatly outperforms the other six methods in terms of F-measure and power, especially in detecting 3rd-order epistatic interactions.

We have utilized SHEIB-AGM (with and without bioinformation) to analyze a real GWAS dataset from WTCCC (the Wellcome Trust Case Control Consortium) [31]. The results demonstrate that SHEIB-AGM can greatly improve the detection ability by using bioinformation. SHEIB-AGM found many epistatic interactions of varies of order. Some of the detected genes have evidence in the CTD database (the Comparative Toxicogenomics Database) [32]. We have drawn SNP networks and gene networks based on the epistatic interactions found by SHEIB-AGM.We have detected many novel genes, which may play a key role in the seven complex diseases studied in the WTCCC dataset, including STK32A-AS1, FAM155B, MTRNR2L10, SNHG14, NCK1-AS1, MIR1254-1, CSAG4, MIR1254-1, and MEIOB. We believe that SHEIB-AGM is a powerful tool to help us in understanding pathogenesis of common and complex diseases.

II. MATERIALS AND METHODS

A. HARDWARE

All experiments were performed on a computer using a Linux system with 48G of RAM and AMD Ryzen Threadripper 1950X CPU.

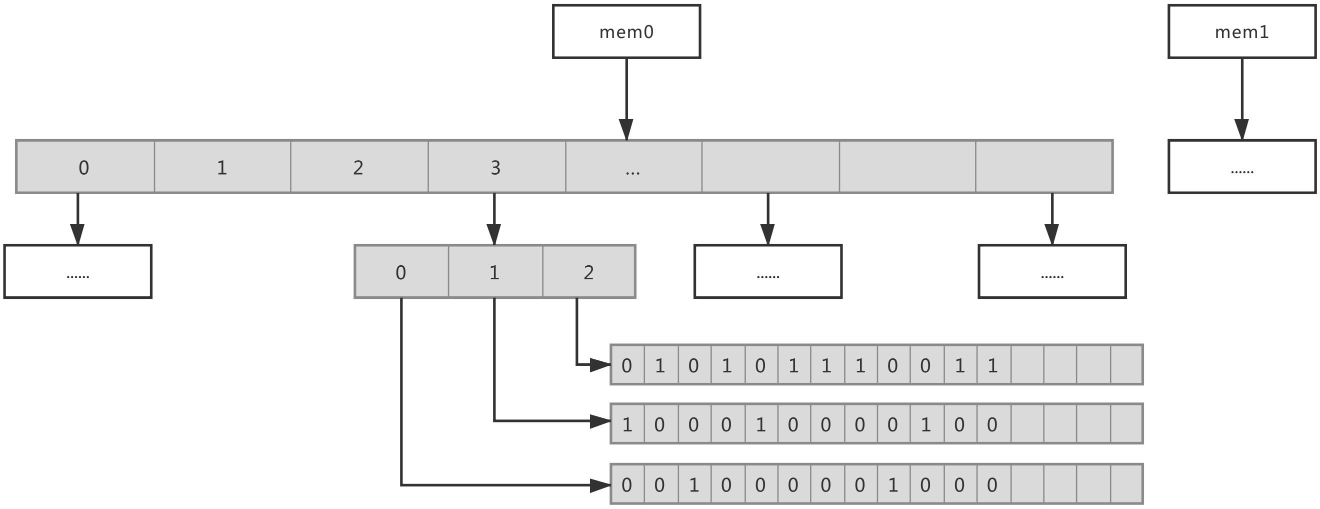


FIGURE 1. **The Boolean representation of the GWAS data.**

B. SHEIB-AGM ALGORITHM

SHEIB-AGM is a stochastic algorithm. In each iteration, it randomly generates an SNP combination containing *mo* SNPs based on bioinformation. If the k2 value of the combination is less than the average value, SHEIB-AGM will try to detect an epistatic interaction on the combination. The pseudo code is shown in Algorithm 1, and detailed descriptions are given in the subsequent subsections.

C. THE BOOLEAN REPRESENTATION AND OPERATION OF GWAS DATA

SHEIB-AGM utilizes a Boolean representation and operation of GWAS data to reduce the computing time and memory consumption, which is very similar to BOOST [33].

Fig. 1 shows the Boolean representation of the GWAS data. In Fig. 1, suppose that the GWAS data contain n SNPs, $m0$ controls, and $m1$ cases. In SHEIB-AGM, $mem0$ and $mem1$ are utilized to store genotype data of controls and cases, respectively. $mem0$ is a vector of length n. $mem0[i]$ $(i \in [0, n))$ is a vector of length 3 and stores the genotype data of the ith SNP. $mem0[i][j]$ $(j \in [0, 2])$ is a Boolean vector of length $m0$. The value of $mem0[i][j][k]$ $(k \in [0, m0))$ can only be 1 or 0. If the genotype of the ith SNP and kth control sample is j, $mem0[i][j][k]$ is 1; otherwise, it is 0. The structure of $mem1$ is similar to $mem0$. For each SNP and sample, SHEIB-AGM only needs 3 bits to store the genotype. This greatly reduces the memory consumption of the GWAS data. Fig. 2 shows how to calculate k2 or G-test for SNP combination [1,2] in SHEIB-AGM. In Fig. 2, suppose that each SNP has only two possible genotypes (0 or 1). The GWAS data contain 8 cases and 8 controls. It uses the Boolean operation to construct a contingency table for the combination. k2 and G-test can be calculated based on the table. The calculation takes advantage of the Boolean operation on the Boolean representation of the GWAS data; thus, it greatly reduces the computing time.

The computation complexity of the Boolean operation in Fig. 2 is $O(m \times o \times 3^o)$. Where o is the number of SNPs contained in the SNP combination. m is the number of samples in the GWAS data. The computation complexity seems to be very high, but most of the operations are Boolean logic operations, so the speed is very fast. The speed improvement of the storage structure and Boolean operation has been proved in other studies [29], [30], [33].

D. CALCULATE MO BASED ON THE NUMBER OF SAMPLES

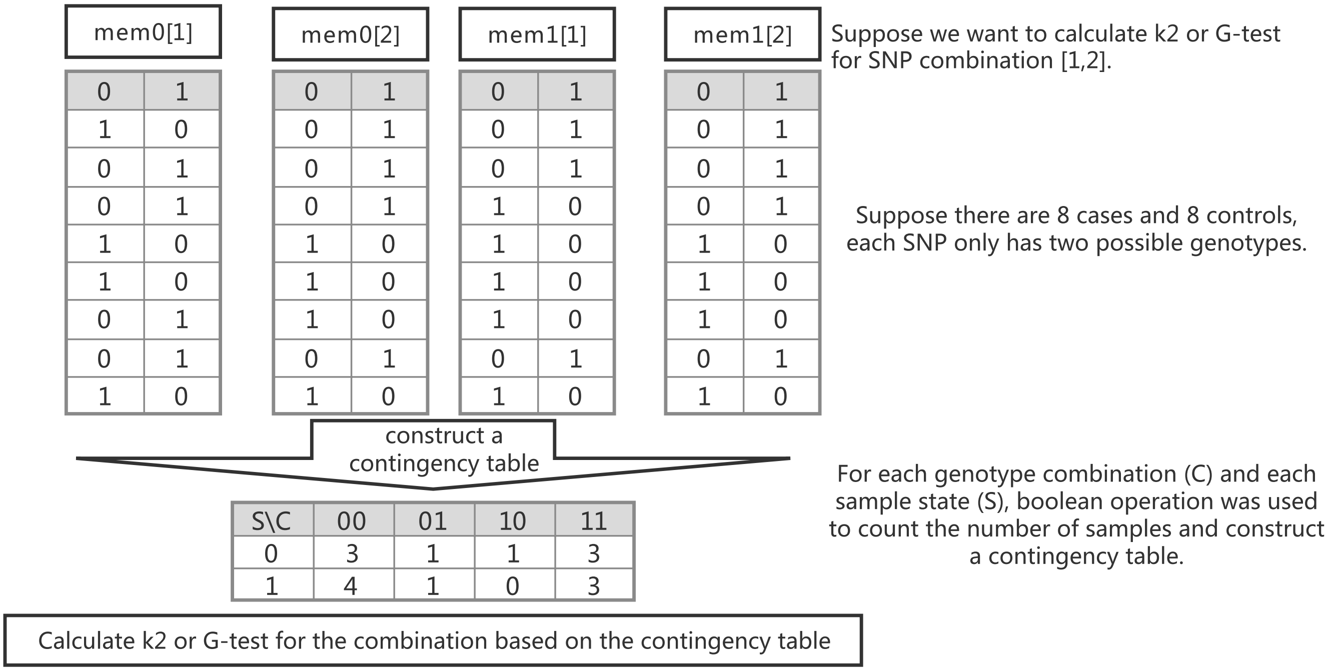
In contrast to other methods, SHEIB-AGM does not need users to specify the order of the epistatic interactions. SHEIB-AGM detects epistatic interactions whose order is less than *mo* (maximum order). *mo* can be specified by the users or calculated based on the number of samples of GWAS data [24]. If *mo* is less than 0, SHEIB-AGM will calculate *mo* as shown in (1).

$$mo = \lfloor ln(\min(m_{*,0}, m_{*,1})) - 0.5 \rfloor \tag{1}$$

In (1), *mo* is the maximum order. $m_{*,0}$ is the number of controls. $m_{*,1}$ is the number of samples. In this work, we do not strictly deduce the formula of *mo*. The larger the value of *mo*, the stronger the detection ability of SHEIB-AGM. Howerver, the too large value of *mo* may make many cells in the multi-SNP comtingence table only have very small number of samples. This will make the calculation of $k2$ function inaccurate. In order to ensure the effectiveness of $k2$ function, we expect that the average number of samples in each unit of the contingency table is 3, so $\frac{\min(m_{*,0}, m_{*,1})}{3^{mo}} = 3$. In (1), we use e (Euler's Number) to approximate 3 and round the calculation result of *mo*.

E. LOAD BIOINFORMATION INTO MEMORY

In this work, the bioinformation is given in a file that records the mapping between SNPs and genes. It can be obtained



mem0[1]		mem0[2]		mem1[1]		mem1[2]	
0	1	0	1	0	1	0	1
1	0	0	1	0	1	0	1
0	1	0	1	0	1	0	1
0	1	0	1	1	0	0	1
1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0
0	1	1	0	1	0	1	0
0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0

Suppose we want to calculate k2 or G-test for SNP combination [1,2].

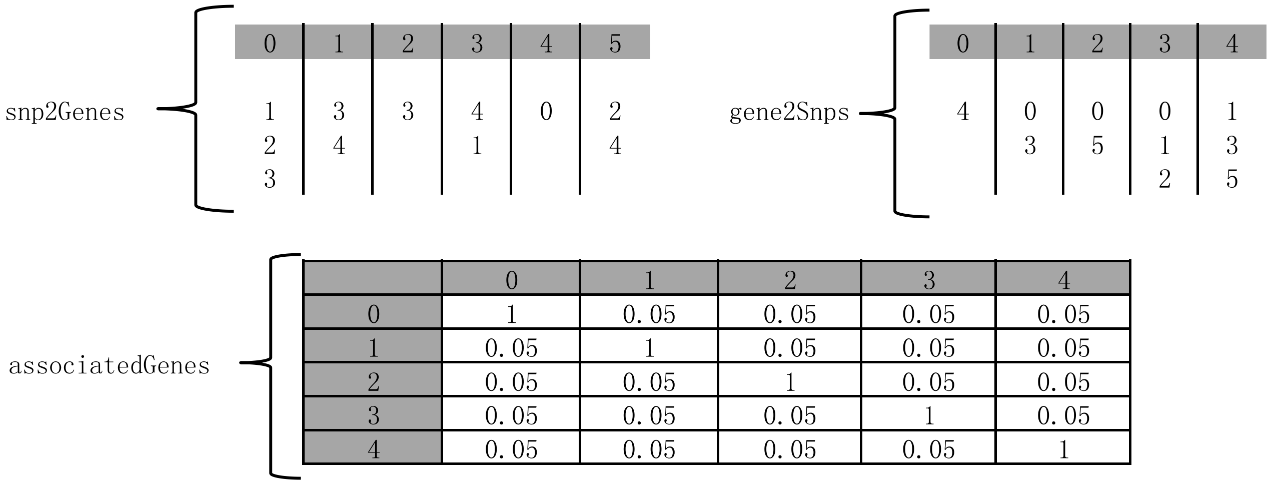
Suppose there are 8 cases and 8 controls, each SNP only has two possible genotypes.

S\C	00	01	10	11
0	3	1	1	3
1	4	1	0	3

For each genotype combination (C) and each sample state (S), boolean operation was used to count the number of samples and construct a contingency table.

Calculate k2 or G-test for the combination based on the contingency table

FIGURE 2. The Boolean operation of GWAS data.



snp2Genes

0	1	2	3	4	5
1 2 3	3 4	3	4 1	0	2 4

gene2Snps

0	1	2	3	4
4	0 3	0 5	0 1 2	1 3 5

associatedGenes

	0	1	2	3	4
0	1	0. 05	0. 05	0. 05	0. 05
1	0. 05	1	0. 05	0. 05	0. 05
2	0. 05	0. 05	1	0. 05	0. 05
3	0. 05	0. 05	0. 05	1	0. 05
4	0. 05	0. 05	0. 05	0. 05	1

FIGURE 3. How SHEIB-AGM stores bioinformation in memory.

from dbSNP [34], which is a database established by NCBI (the National Center for Biotechnology Information) [35]. For each SNP in the WTCCC data (the real GWAS data), we obtained the gene or genes related to it in the dbSNP database.

To use bioinformation in SHEIB-AGM, we have made two very reasonable assumptions. According to Assumption 1, the algorithm should have bias such that it can tend to detect epistasis between SNPs on the same gene. According to Assumption 2, the algorithm should have bias such that it can tend to detect epistasis between SNPs on the genes in which an epistatic interaction has been found.

Assumption 1: Epistasis usually occurs within genes.

Assumption 2: The epistasis between different genes is regular. If we have found epistatic interactions on different genes, we will likely detect more interactions on the genes.

In SHEIB-AGM, as shown in Fig. 3, three variables are constructed based on gene-mapping data. *snp2Genes* is utilized to obtain SNPs located in a gene. *gene2Snps* is utilized to obtains genes in which an SNP is located. *associatedGenes* is a gene matrix. In each iteration, SHEIB-AGM tends to detect epistasis in an SNP combination whose corresponding genes have lager values in the gene matrix. According to Assumption 1, during initialization, we make the diagonal values of the matrix 1 and the non-diagonal values *minRate*. The formula for *minRate* is shown in (2). In Fig. 3,

Algorithm 1 The Pseudo Code of SHEIB-AGM Algorithm

```
Require: $pathIn$ is the path of GWAS data; $pathOut$ is the path of result; $pathS2G$ is the path of the file which records
    bioinformation; $mo$ is the maximum order, SHEIB-AGM will detect epistatic interactions whose orders are less than $mo$;
    $maxGen$ is the maximum number of iterations;
Ensure: epistatic interactions;
procedure SHEIB-AGM($pathIn$,$pathOut$,$pathS2G$,$mo$,$maxGen$)
    read from $pathIn$ and save GWAS data in memory using the Boolean representation;
    if $mo == -1$ then
        calculate $mo$ based on the number of samples;
    end if
    initialize $snp2Genes = null$; $gene2Snps = null$; $associatedGenes = null$; $G = 0$; $meanK2 = 0$; $results = empty\ set$;
    if $pathS2G \neq null$ then
        construct $snp2Genes$ and $gene2Snps$ based on the content of $pathS2G$;
        construct $gene\ matrix$ ($associatedGenes$);
    end if
    while $G < maxG$ do
        randomly generate an SNP combination $x$ based on the bioinformation;
        calculate k2 of $x$ as $k2x$;
        $meanK2 = \frac{meanK2*G+k2x}{G+1}$;
        if $k2x < meanK2$ then
            try to detect epistatic interaction on $x$;
            if an interaction is found then
                determine whether the interaction is significant based on G-test;
                if the interaction is significant then
                    add the interaction into $results$;
                end if
            end if
        end if
        update $associatedGenes$ based on whether a new interaction is found in this iteration.
        $G = G + 1$;
    end while
end procedure
```

suppose that there are 6 SNPs and 4 genes in the GWAS data. *snp2Genes* is a hash map. Its key represents an SNP, and its value represents the genes associated with the SNP. *gene2Snps* is also a hash map. Its key represents a gene, and its value represents the SNPs associated with the gene. In the three variables, the 0th gene represents the unknown gene. All SNPs that are not located in any genes are thought to be located in the unknown gene. *associatedGenes* is a symmetric matrix. It maintains a value for each pair of genes (including the unknown gene). In this figure, *pb* is set to 0.8.

$$minRate = \frac{1 - pb}{nGenes} \tag{2}$$

In (2), *pb* is a parameter specified by the users. *nGenes* is the number of genes in the GWAS data.

F. GENERATE AN SNP COMBINATION BASED ON THE GENE MATRIX

In each iteration of SHEIB-AGM, it generates an SNP combination based on *associatedGenes*. As shown in Algorithm 2, when x visits a new gene, *nextGenes* is updated based on *associatedGenes*. While generating each SNP for x, except for the first SNP, SHEIB-AGM uses roulette to randomly select a gene based on *nextGenes* and randomly selects an SNP in the selected gene to insert into x. If bioinformation is not provided, it will generate a completely random SNP combination.

G. DETECT AN EPISTATIC INTERACTION ON AN SNP COMBINATION BASED ON K2

The Bayesian network scoring criterion (k2) [36] is widely used in detecting epistatic interactions. The formula for k2 is shown in (3).

$$k2(Y, S) = \prod_{c \in C} \frac{m_{c,0}! \times m_{c,1}!}{(m_{c,*} + 1)!} \tag{3}$$

In (3), $k2$ is the score used to measure the association between an SNP combination and the phenotype. S represents an SNP combination. Y represents the phenotype. C is the set of the genotype combinations of S (if S contains l SNPs, C will be a set of 3^l). $m_{c,*}$ is the number of samples whereby the genotype combinations of the SNP combinations are c. $m_{c,0}$

Algorithm 2 Generate an SNP Combination Based on *Associatedgenes*

```
Require: $mo$ is the maximum order; $n$ is the number of SNPs in the GWAS data; $nGenes$ is the number of genes in the GWAS
    data; $snp2Genes$, $gene2Snps$ and $associatedGenes$ are the variables containing bioinformation; $rand()$ is a function which
    returns a random decimal $\in [0, 1)$;
Ensure: an SNP combination $x$;
procedure RanGen($mo$,$n$,$nGenes$,$snp2Genes$,$gene2Snps$,$associatedGenes$)
    initialize a vector $x$ of length $mo$;
    initialize $visitedGenes$ as an empty hash map;
    initialize $nextGenes$ as a vector of length $nGenes + 1$;
    $ti = rand(0, n)$;
    $x[0] = ti$;
    for $i \in [1, mo)$ do
        if $snp2Genes \neq null$ then
            for each gene $g \in snp2Genes[ti]$ do
                if $g \in visitedGenes$ then
                    $visitedGenes[g] + = 1$;
                else
                    $visitedGenes[g] = 1$;
                    $nextGenes + = associatedGenes[g]$;
                end if
            end for
            calculate the sum of $nextGenes$ as $s$;
            $s = rand() * s$;
            for $j \in [0, nGenes)$ do
                $s - = nextGenes[j]$;
                if $s < 0$ then
                    randomly select an SNP from $gene2Snps[j]$ as $ti$;
                    break;
                end if
            end for
        else
            randomly select an SNP from the SNPs which are not in $x$;
        end if
        $x[i] = ti$ and ensure that all elements in $x$ are in ascending order;
    end for
end procedure
```

is the number of controls whereby the genotype combinations are c. $m_{c,1}$ is the number of cases whereby the genotype combinations are c.

$k2(Y, S)$ can measure the quality of the Bayesian network constructed using S and Y. If $k2(Y, S)$ is smaller, the Bayesian network is more accurate, and the association between S and Y is more significant. When an SNP x is removed from S, if x is a noise variable (x has no effect on the phenotype), the quality of the Bayesian network will be improved, and $k2(Y, S)$ will be smaller. If x is associated with the phenotype or x has epistasis with any one of the other SNPs in S, $k2(Y, S)$ will be larger. The change in $k2(Y, S)$ is very useful in detecting an epistatic interaction on an SNP combination. As shown in Algorithm 3, each SNP in the combination is removed to check if the SNP is associated with the phenotype or if it is a part of an epistatic interaction. SHEIB-AGM will attempt to remove SNPs until it cannot remove anyone of them. If the final combination after the removal contains more than two SNPs, it will be an epistatic interaction.

H. DETERMINE WHETHER THE INTERACTION IS SIGNIFICANT BASED ON G-TEST

In SHEIB-AGM, epistatic interactions are divided into significant interactions and non-significant interactions. Using Algorithm 3, we can detect numerous epistatic interactions. In this subsection, SHEIB-AGM determines whether the interactions are significant by using G-test. G-test [37] is a likelihood-ratio or maximum likelihood statistical significance test. If the interaction is not associated with the phenotype, the distribution of G-statistic will be approximately a chi-squared distribution. It is widely used to screen out significant interactions. In this work, we utilize the p-value of G-test (g) and the change in g (gc) [30] to screen out the significant interactions. The formulas for g and gc are shown

Algorithm 3 Detect an Epistatic Interaction on an SNP Combination Based on k2

```
Require: $mo$ is the maximum order; $x$ is the SNP combination generated by SHEIB-AGM, it contains $mo$ SNPs; $k2x$ is the k2 value of $x$;
Ensure: an epistatic interaction whose order is in $[2, mo)$ or nothing;
procedure DetectEpi($mo$,$x$,$k2x$)
    initialize $l = mo$;
    while $l \neq 1$ do
        for $i \in [0, l)$ do
            $bBreak = false$;
            initialize $xx$ as a vector of length $l - 1$;
            copy all SNPs of $x$ into $xx$, except $x[i]$;
            calculate k2 of $xx$ as $k2xx$;
            if $k2xx < k2x$ then
                $x = xx$;
                $k2x = k2xx$;
                $l = l - 1$;
                $bBreak = true$;
                break;
            end if
        end for
        if $bBreak == false$ then
            break;
        end if
    end while
    if $l > 1$ then
        return $x$ as the epistatic interaction;
    end if
end procedure
```

in (4), as shown at the bottom of the next page. The interactions whose g and gc are both less than the thresholds specified by the users are significant. The significant interactions are recorded in the result file.

In (4), $g(Y; S)$ is the p-value of G-test. Y represents the phenotype. S represents an SNP combination. $p - value\_of$ represents the function used to compute the p-value of the chi-square distribution. C is the set of the genotype combinations of S (if S contains l SNPs, C will be a set of 3^l). m is the number of samples. $m_{c,0}$ is the number of controls whereby the genotype combinations are c. $m_{c,1}$ is the number of cases whereby the genotype combinations are c. $m_{c,*}$ is the number of samples whereby the genotype combinations of the SNP combinations are c. $m_{*,0}$ is the number of controls. $m_{*,1}$ is the number of cases. $\min_{E \in S} g(Y; E)$ represents the g of the SNP whose g is the smallest in S.

I. UPDATE GENE MATRIX

In this subsection, we describe how to update the gene matrix. After the initialization, the diagonal values of the matrix are 1, and the non-diagonal values are *minRate* (very small). In each iteration, if a significant epistatic interaction is found, each pair of the genes in which SNPs in the interaction are located will be set to 1 in the gene matrix (according to Assumption 2). In the subsequent iterations, the tendency to detect epistasis between the genes will increase. If SHEIB-AGM cannot detect a significant interaction in the SNP combination generated by Algorithm 2, each pair of the genes in which the SNPs in the combination are located will decrease, as shown in (5). In the subsequent iterations, the tendency to detect epistasis between the genes will decrease.

In (5), as shown at the bottom of the next page, $associatedGenes[i, j]$ is the value between the ith gene and jth gene in the gene matrix. $decRate$ is a parameter specified by the users.

III. RESULTS AND DISCUSSION

A. EXPERIMENTS ON SIMULATED DATA

1) SIMULATED DATASETS

In this work, we compared SHEIB-AGM with six other methods, DECMDR [22], SNPHarvester [24], MACOED [25], AntEpiSeeker [26], HS-MMGKG [29] and SEE [30], on three simulated datasets. All seven software packages and their parameter settings are shown in Table 1. DECMDR, MACOED, HS-MMGKG and SEE can detect any specified order epistatic interactions. SNPHarvester and AntEpiSeeker can only detect 2-order interactions. Although AntEpiSeeker was designed to detect any specified order interactions, when we executed it to detect 3-order interactions, ‘‘segment fault’’ occurred. The three simulated datasets are as follows:

TABLE 1. Introduction of the seven software used in the simulated experiment and their parameter settings.

		Parameter setting			
Algorithm	Language	All the three datasets	DME and DNME 100	DME and DNME 1000	DNME3 100
DECMDR	Java	$s = 1;$ $m = 0.5;$ $r = 0.5;$	$p = 80;$ $g = 40;$ $o = 2;$	$p = 800;$ $g = 140;$ $o = 2;$	$p = 80;$ $g = 400;$ $o = 3;$
SNPHarvester	Java	there is no parameter			not executable
MACOED	Matlab	$pvalue = 0.05/4950;$ $tau0 = 1;$ $T0 = 0.8;$ $rou = 0.9;$ $lambda = 2;$	$num\_ant = 80;$ $max\_iter = 40;$ $dim\_epi = 2;$	$num\_ant = 800;$ $max\_iter = 140;$ $dim\_epi = 2;$	$num\_ant = 80;$ $max\_iter = 400;$ $dim\_epi = 3;$
AntEpiSeeker	C++	$alpha = 1;$ $iTopModel = 80;$ $iTopLoci = 16;$ $rou = 0.05;$ $phe = 100;$ $largehapsize = 6;$ $smallhapsize = 3;$ $iEpiModel = 2;$ $pvalue = 0.01;$	$iAntCount = 80;$ $iItCountLarge = 10;$ $iItCountSmall = 30;$	$iAntCount = 800;$ $iItCountLarge = 40;$ $iItCountSmall = 100;$	not executable
HS-MMGKG	Java	$nsolution = 1;$ $hmcr = 0.8;$ $par = 0.4;$ $fold = 5;$ $p - vlaue = 0.05;$	$hms = 80/5;$ $tmax = 80 * 40;$ $order = 2;$	$hms = 800/5;$ $tmax = 800 * 140;$ $order = 2;$	$hms = 80/5;$ $tmax = 80 * 400;$ $order = 3;$
SEE	C++	$pe = 1;$ $cCec = 0;$ $cGinic = 0;$ $cK2c = 1;$ $cGc = 1;$ $cG = 0.05;$ $rn = 1;$ default values for the other parameters;	$numPop = 80;$ $maxIter = 80 * 40;$ $order = 2;$	$numPop = 800;$ $maxIter = 800 * 140;$ $order = 2;$	$numPop = 80;$ $maxIter = 80 * 400;$ $order = 3;$
SHEIB-AGM	Java	$o = -1;$ $rn = 1;$ $pb = 0.8;$ $cG = 0.05;$ $cGc = 1;$	$maxGen = 80 * 40;$	$maxGen = 800 * 140;$	$maxGen = 80 * 400;$

- DME and DNME 100: This dataset contains 8 DME (disease loci with marginal effects) and 60 DNME (disease loci without marginal effects) models. Each model contains 100 simulated GWAS files. Each file contains 100 SNPs, 800 cases, and 800 controls. The DME models were obtained from DECMDR [22] and the DNME models were generated based on a variety of MAFs [0.2,0.4] and Heritabilities [0.025,0.05,0.1,0.2,0.3,0.4] by using GAMETES_2.1 [38]. The penetrance tables of the 68 models are shown in Table S1 in the Supplementary Appendix.
- DME and DNME 1000: This dataset is the same as DME and DNME 100, except that in this dataset, each simulated GWAS file contains 1000 SNPs.
- DNME3 100: This dataset contains 40 DNME models. Each model contains 100 simulated GWAS file. Each

$$g(Y;S) = pvalue\_of\left(2\sum_{c\in C} m_{c,0} \times \frac{m_{c,0}\times m}{m_{c,*}\times m_{*,0}} + m_{c,1} \times \frac{m_{c,1}\times m}{m_{c,*}\times m_{*,1}}\right)$$

$$gc(Y;S) = \begin{cases} \dfrac{g(Y;S)}{\min\limits_{E\in S} g(Y;E)}, & if\ \min\limits_{E\in S} g(Y;S) \neq 0 \\ 1, & if\ g(Y;S) = 0\ and\ \min\limits_{E\in S} g(Y;E) = 0 \\ 4, & if\ g(Y;S) \neq 0\ and\ \min\limits_{E\in S} g(Y;E) = 0 \end{cases} \tag{4}$$

$$T = associatedGenes[i,j] \times (1 - decRate)$$

$$associatedGenes[i,j] = \begin{cases} T, & if\ T > minRate \\ minRate, & if\ T <= minRate \end{cases} \tag{5}$$

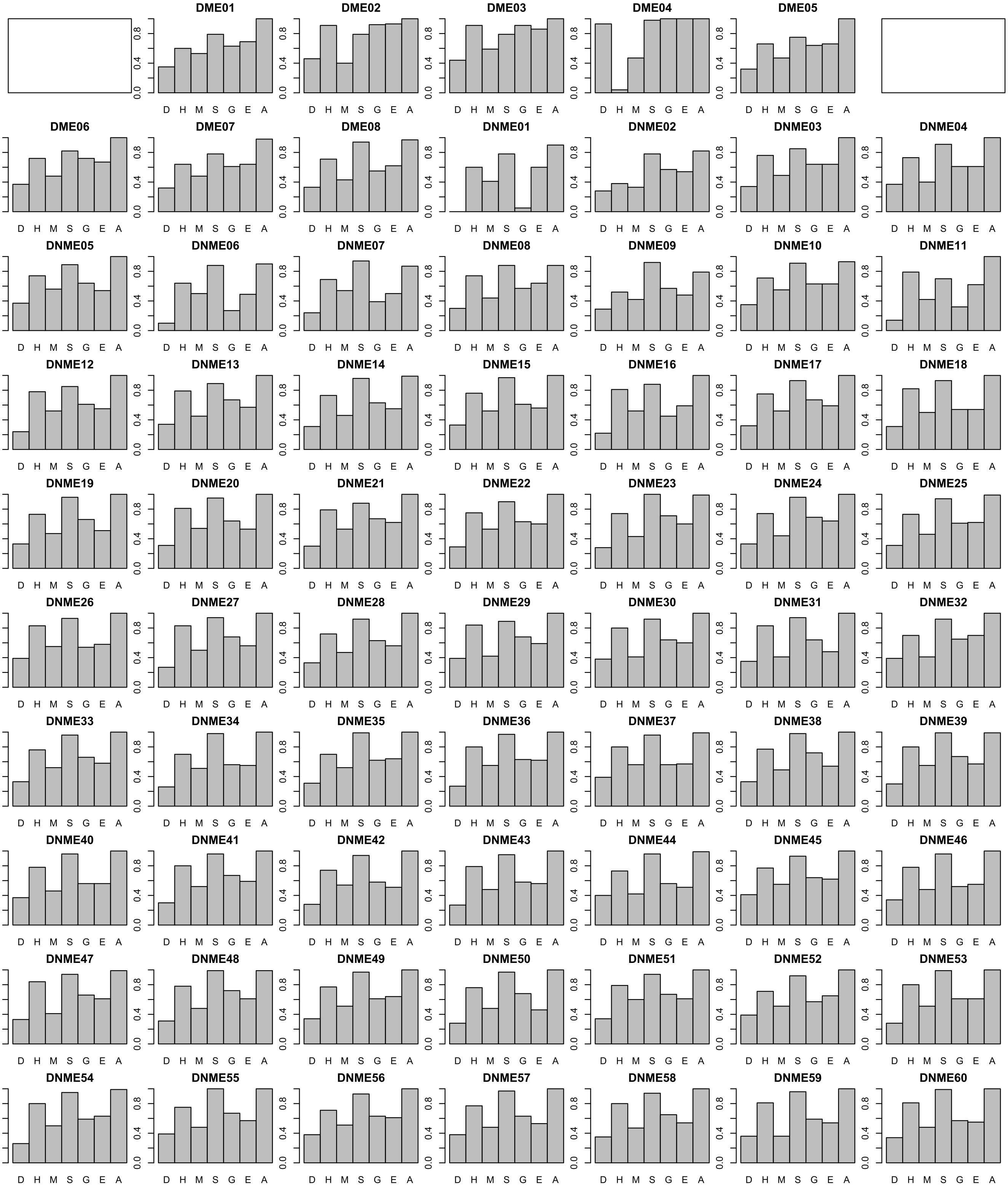


FIGURE 4. Power comparisons between DECMDR (D), SNPHarvester (H), MACOED (M), AntEpiSeeker (S), HS-MMGKG (G), SEE (E) and SHEIB-AGM (A) with the DME and DNME 100 dataset. The bars represent powers of the algorithms.

file contains 100 SNPs, 800 cases, and 800 controls. The models were generated by GAMETES_2.1 based on a variety of MAFs [0.2,0.4] and Heritabilities [0.025,0.05,0.1,0.2]. The penetrance tables of the 40 models are shown in Table S2 in the Supplementary Appendix.

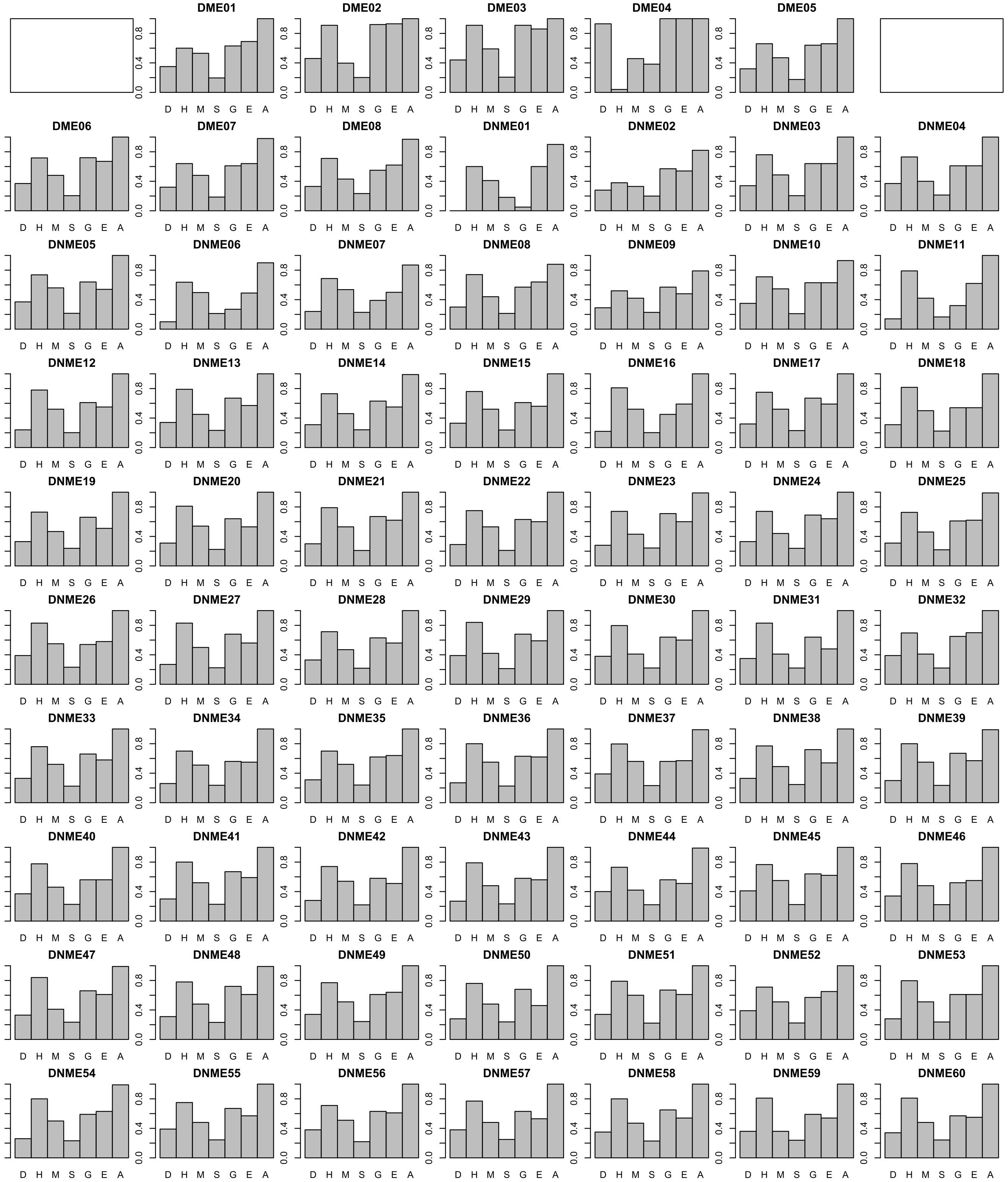


FIGURE 5. F-measure comparisons between DECMDR (D), SNPHarvester (H), MACOED (M), AntEpiSeeker (S), HS-MMGKG (G), SEE (E) and SHEIB-AGM (A) with the DME and DNME 100 dataset. The bars represent F-measures of the algorithms.

All three simulated datasets were generated by GAMETES_2.1 based on their penetrance tables.

2) EVALUATION CRITERIA

In this work, we utilize the F-measure [25], [30] and power [22], [39] to evaluate the performance of the methods.

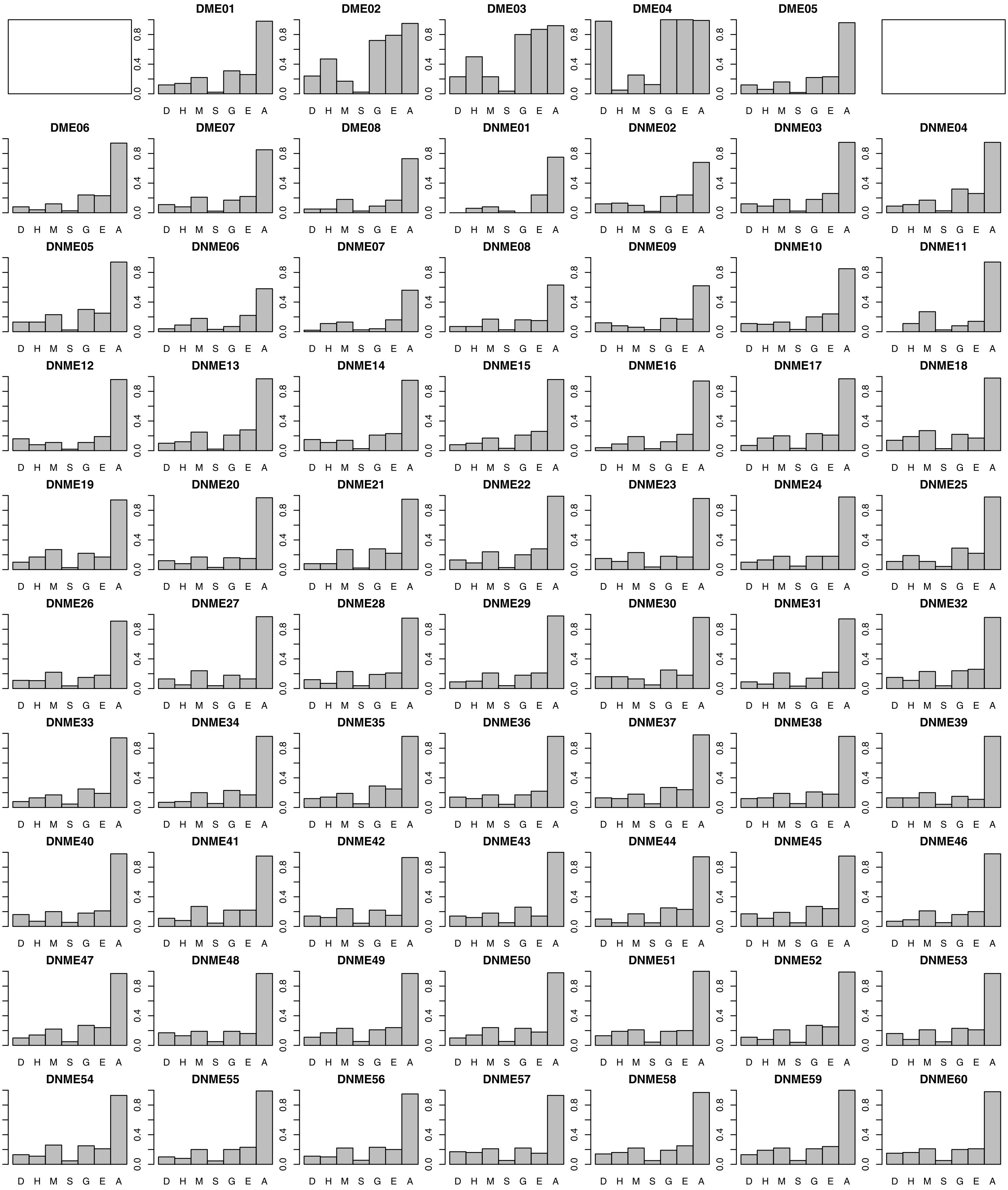


FIGURE 6. F-measure comparisons between DECMDR (D), SNPHarvester (H), MACOED (M), AntEpiSeeker (S), HS-MMGKG (G), SEE (E) and SHEIB-AGM (A) with the DME and DNME 1000 dataset. The bars represent F-measures of the algorithms.

They are both widely used criteria to evaluate the ability to detect epistatic interactions. The F-measure and power are calculated as shown in (6). For each disease model, the algorithm detects epistatic interaction in 100 GWAS files. The power represents the rate at which we have detected the true epistatic interaction in the files. For each model and each

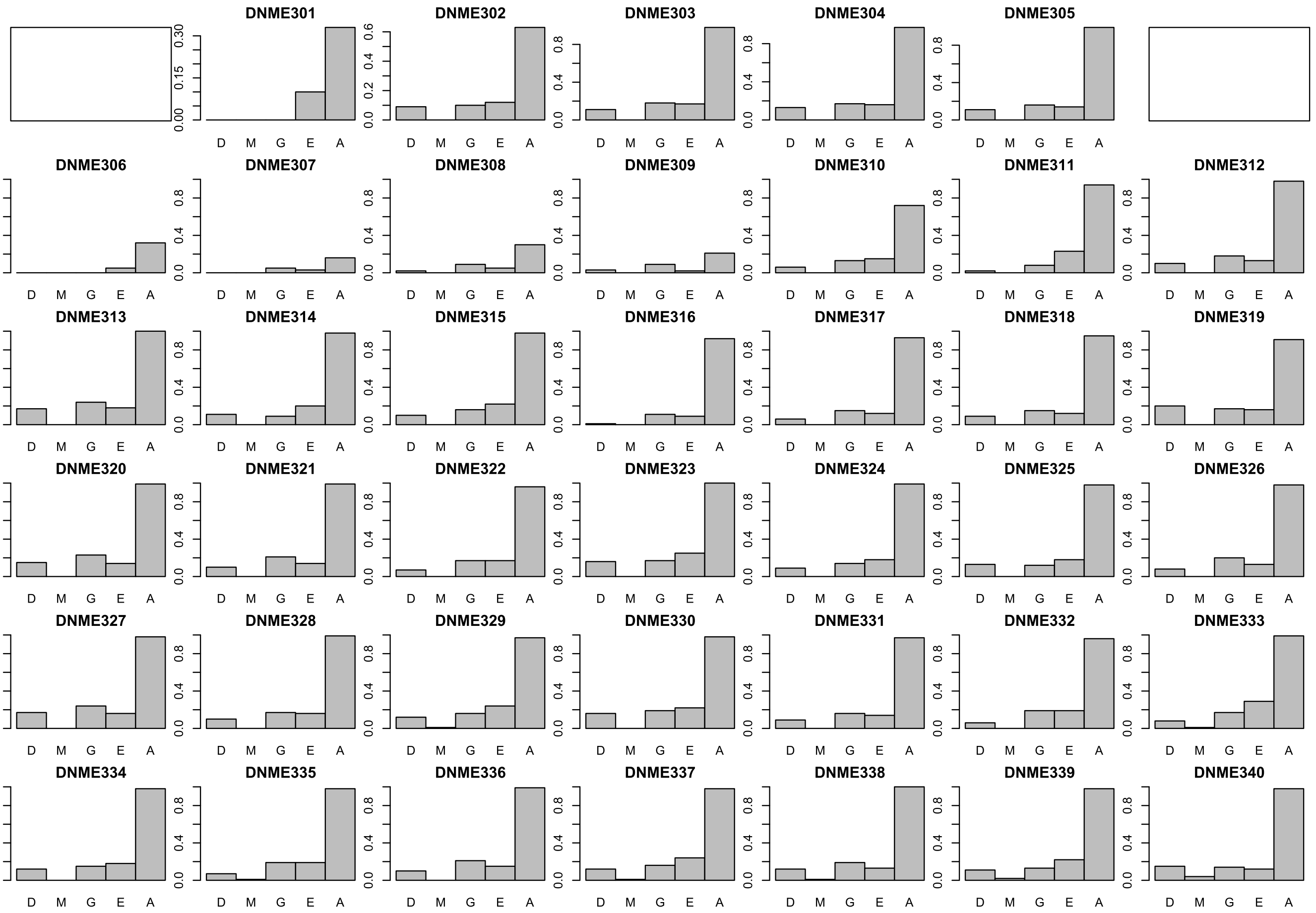


FIGURE 7. F-measure comparisons between DECMDR (D), MACOED (M), HS-MMGKG (G), SEE (E) and SHEIB-AGM (A) with the DNME3 100 dataset. The bars represent F-measures of the algorithms.

algorithm, 100 F-measures are generated from 100 GWAS files, and the F-measure of the algorithm on the model is the average of the 100 values.

$$
\begin{aligned}
power &= \frac{\#(S)}{100} \\
recall &= \frac{TP}{TP+FN} \\
precision &= \frac{TP}{TP+FP} \\
F-measure &= \frac{2}{\frac{1}{recall}+\frac{1}{precision}}
\end{aligned}
\tag{6}
$$

In (6), *power* and $F - measure$ are the two evaluation criteria used in this work. $\#(S)$ means the number of files (100 files in total) in which the algorithm has detected the true epistatic interaction. *TP* (true positive) is the number of true epistatic interactions found by the algorithm. *FN* (false negative) is the number of true epistatic interactions not found by the algorithm. *FP* is the number of SNP combinations which are not epistatic interactions and not found by the algorithm.

3) COMPARISON OF SEHIB-AGM WITH EXISTING METHODS ON SIMULATED DATA

On the DME and DNME 100 dataset, we compared SHEIB-AGM with the other methods. The parameters are given in Table 1. The average powers of DECMDR, SNPHarvester, MACOED, AntEpiSeeker, HS-MMGKG and SEE are 0.328088235, 0.741029412, 0.483823529, 0.918970588, 0.615735294 and 0.5975, respectively. Their average F-measures are 0.328088235, 0.740343191, 0.483350868, 0.224458529, 0.615735294, and 0.5975, respectively. The F-measure and power of SHEIB-AGM are both 0.984558824. Fig. 4 and Fig. 5 show the comparisons between the seven methods with the DME and DNME 100 dataset. It is found that SHEIB-AGM outperforms the other six methods in terms of power and F-measure with this simulated dataset. The detailed experiment results are shown in Table S3 and Table S4 in the Supplementary Appendix.

On the DME and DNME 1000 dataset, there are ten times as many SNPs in the simulated data, and detecting epistasis is more difficult. The parameters were set as in Table 1. The average powers of DECMDR, SNPHarvester, MACOED, AntEpiSeeker, HS-MMGKG and SEE

TABLE 2. The final seven GWAS data constructed from the WTCCC dataset.

data	disease	number of SNPs	number of cases	number of controls	number of samples
bd_gwas	Bipolar Disorder	458922	1868	2938	4806
cad_gwas	Coronary Artery Disease	458743	1926	2938	4864
cd_gwas	Crohn's Disease	459472	1748	2938	4686
ht_gwas	Hypertension	458851	1952	2938	4890
ra_gwas	Rheumatoid Arthritis	458854	1860	2938	4798
t1d_gwas	Type 1 Diabetes	459244	1963	2938	4901
t2d_gwas	Type 2 Diabetes	459112	1924	2938	4862

TABLE 3. The number of epistatic interactions detected on the seven GWAS data using SHEIB-AGM without bioinformation.

epistatic interactions	bd_gwas	cad_gwas	cd_gwas	ht_gwas	ra_gwas	t1d_gwas	t2d_gwas
2-order	28	4510	37	35	5867	46	23
3-order	2	624	0	1	757	11	2
4-order	0	31	0	0	23	0	0
5-order	0	3	0	0	0	0	0
total	30	5168	37	36	6647	57	25

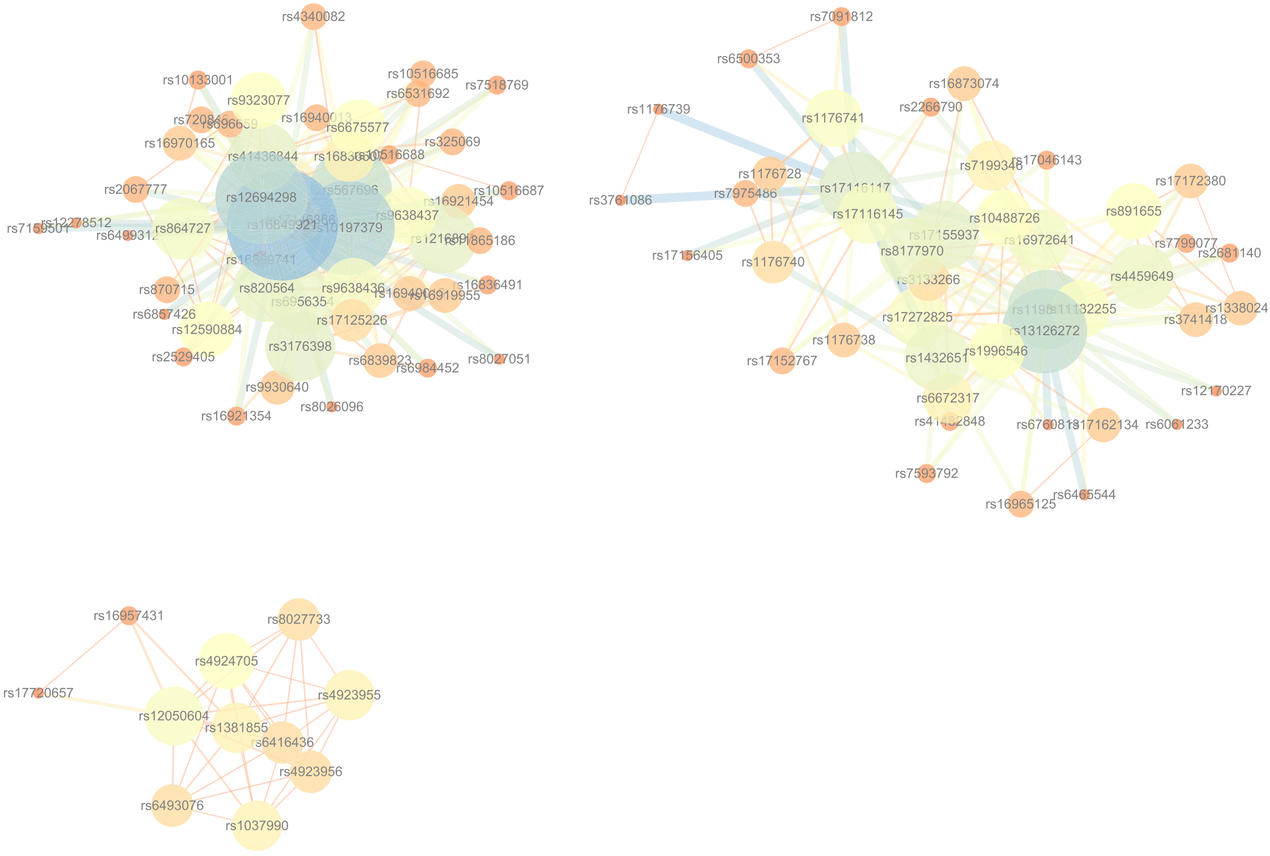


FIGURE 8. The SNP network of the epistatic interactions detected for Bipolar Disorder (only 400 SNP pairs with minimum p-values of G-test). The SNP networks for the other six diseases are shown in Fig. S3-S8 in the Supplementary Appendix.

are 0.126764706, 0.120882353, 0.196029412, 0.608088235, 0.230441176 and 0.237647059, respectively. Their average F-measures are 0.126764706, 0.120833338, 0.195931368, 0.038724412, 0.230441176 and 0.237647059, respectively. The F-measure and power of SHEIB-AGM are both 0.926323529. Fig. 6 shows the F-measure comparisons between the seven methods on this dataset. Fig. S1 in the Supplementary Appendix shows the power comparisons between the seven methods with this dataset. Although there are many more SNPs in the GWAS data, SHEIB-AGM still outperforms the other six methods with respect to power and F-measure. The detailed experimental results are shown in Table S5 and Table S6 in the Supplementary Appendix.

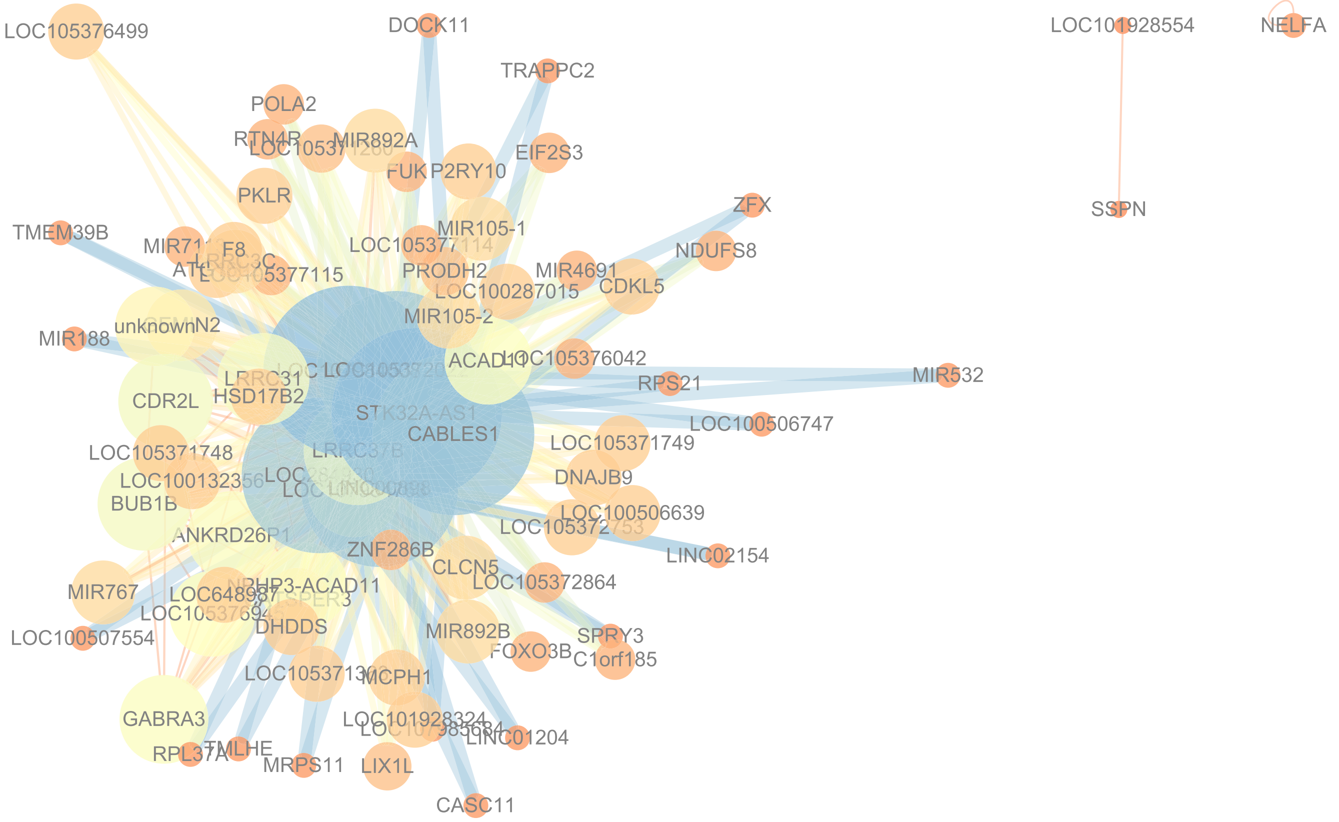


FIGURE 9. The gene network of the epistatic interactions detected for Bipolar Disorder (only 400 gene pairs with the highest frequency of occurrence). The gene networks for the other six diseases are shown in Fig. S9-S14 in the Supplementary Appendix.

TABLE 4. The number of epistatic interactions detected on the seven GWAS data using SHEIB-AGM with bioinformation.

epistatic interactions	bd_gwas	cad_gwas	cd_gwas	ht_gwas	ra_gwas	t1d_gwas	t2d_gwas
2-order	698	45579	174	404	106671	5643	349
3-order	6315	273362	450	1800	358274	60713	2078
4-order	23474	315054	537	4142	219963	71442	4839
5-order	16687	135112	91	3217	55539	31479	3258
6-order	3461	21750	4	856	6057	5499	658
7-order	84	686	0	23	146	180	16
total	50719	791543	1256	10442	746650	174956	11198

On the DNME3 100 dataset, we compared SHEIB-AGM with the other methods in detecting third-order epistatic interactions. The parameters were as listed in Table 1. The average powers of DECMDR, MACOED, HS-MMGKG and SEE are 0.094, 0.00275, 0.14975 and 0.1565, respectively. Their average F-measures are 0.094, 0.00275, 0.14975 and 0.1565, respectively. The F-measure and power of SHEIB-AGM are both 0.8705. Fig. 7 shows the F-measure comparisons between the five methods with this dataset. Fig. S2 in the Supplementary Appendix shows the power comparisons between the five methods with this dataset. It is found that SHEIB-AGM outperforms DECMDR, MACOED, HS-MMGKG and SEE with respect to power and F-measure on this simulated dataset. The detailed experimental results are shown in Table S7 and Table S8 in the Supplementary Appendix.

B. EXPERIMENTS ON REAL DATA

1) WTCCC DATASET

In this work, we used a real dataset from WTCCC (the Wellcome Trust Case Control Consortium). In the dataset, there are approximately 14,000 cases of seven complex diseases and 3000 controls. The seven diseases are Bipolar Disorder, Coronary Artery Disease, Crohn's Disease, Hypertension, Rheumatoid Arthritis, Type 1 Diabetes, and Type 2 Diabetes. For each disease, there are approximately 2,000 samples. For each sample, the genotype of approximately 500,000 SNPs has been measured. Table S9 in the Supplementary Appendix gives a description to the WTCCC dataset. We combined the cases of each disease and the controls to construct seven GWAS data. Following the WTCCC's recommendation, we removed some SNPs and samples. For each GWAS data, we also removed the SNPs whose genotype

TABLE 5. Some of the epistastic interactions detected on the seven GWAS data using SHEIB-AGM with bioinformation. The complete table is shown in Table S11 in the Supplementary Appendix.

g	gc	snp1	snp2	snp3	snp4	snp5	snp6	snp7
Bipolar Disorder								
0	0	rs1381855	rs12050604	rs1037990				
0	0	rs16849921	rs10197379					
0	0	rs17116145	rs17116117	rs1176728	rs1176740	rs1176741		
0	0	rs4924705	rs4923955	rs12050604	rs1037990			
0	0	rs16849921	rs10197379	rs9638436	rs9638437	rs567696	rs3176398	
0	0	rs6675577	rs16849921	rs12694298	rs820564	rs6956354	rs41436844	rs567696
Coronary Artery Disease								
0	0	rs2266829	rs523773					
0	0	rs228337	rs17330041	rs5927017				
0	0	rs41435147	rs34106226	rs115571	rs10521972			
0	0	rs2167594	rs3176406	rs3176398	rs1933428	rs3027898		
0	0	rs11249085	rs16829083	rs3176406	rs3176398	rs1933428	rs3027898	
0	0	rs2032749	rs5971434	rs5925268	rs1933428	rs5955612	rs916313	rs7878756
Crohn’s Disease								
0	0	rs17116117	rs1176728	rs1176738				
0	0	rs17116145	rs17116117					
0	0	rs11921179	rs7154773	rs10142834	rs17097262			
0	0	rs11921179	rs8011227	rs1887103	rs7154773	rs10142834	rs8012816	
0	0	rs11921179	rs11064445	rs1887103	rs7154773	rs17097262		
Hypertension								
0	0	rs8137391	rs16986990					
0	0	rs10016497	rs6840033	rs868653				
0	0	rs1887104	rs7154773	rs10142834	rs8012816	rs17097243	rs7158657	
0	0	rs6840033	rs883455	rs10519535	rs7678137			
0	0	rs7154773	rs10142834	rs17097262	rs2972437	rs2972438		
0	0	rs7867394	rs7045602	rs8011227	rs7154773	rs10142834	rs8012816	rs7158657
Rheumatoid Arthritis								
0	0	rs7154773	rs10142834	rs10130695				
0	0	rs5972125	rs5985895					
0	0	rs7154773	rs10142834	rs10130695	rs17097243			
0	0	rs16836194	rs16876800	rs5980711	rs727562	rs5987569		
0	0	rs11125352	rs2075800	rs2722496	rs17114865	rs6521112	rs5987569	
0	0	rs707974	rs10244032	rs16886500	rs16987067	rs5964260	rs2236153	rs41371346
Type 1 Diabetes								
0	0	rs41417553	rs2904776					
0	0	rs8011227	rs7154773	rs10142834	rs7158657			
0	0	rs3016013	rs2523485	rs16905827				
0	0	rs4861558	rs3177928	rs3135392	rs7194	rs1051336		
0	0	rs3016013	rs2507976	rs4081552	rs9266775	rs17154559	rs16958762	
0	0	rs34717730	rs3130284	rs408359	rs3130348	rs9266774	rs7067635	rs16958762
Type 2 Diabetes								
0	0	rs16849921	rs10197379					
0	0	rs16849921	rs10197379	rs12694298				
0	0	rs8011227	rs7154773	rs17097243	rs7158657			
0	0	rs6940205	rs10499044	rs4255065	rs6958533	rs6135716		
0	0	rs17037861	rs6940205	rs10499044	rs4255065	rs6958533	rs7256304	
0	0	rs6940205	rs10499044	rs11033219	rs7195033	rs16958762	rs16998352	rs17004654

is unchanged in all samples. Table 2 shows the final seven GWAS data.

2) RESULTS ON THE SEVEN WTCCC GWAS DATA USING SHEIB-AGM WITHOUT BIOINFORMATION

According to the previous analysis of the simulation experiments, compared to other methods, the proposed algorithm achieves a good performance on the three simulated datasets. We applied SHEIB-AGM without bioinformation to analyze the seven GWAS data from WTCCC. We set $pb = 0.9$, $decRate = 0.01$, $cG = 0.05$, $cGc = 0.05$, $o = -1$, $maxGen = 4 \times 10^7$, $seed = 0$, $rn = -1$, and other parameters as default values. We have found many epistatic interactions with varying orders, as shown in Table 3. It is more difficult to detect higher order epistatic interactions than lower order interactions. More detailed results can be found in Table S10 in the Supplementary Appendix (Table S10-S13 can be obtained at https://github.com/sunliyan0000/sheib-agm).

3) RESULTS ON THE SEVEN WTCCC GWAS DATA USING SHEIB-AGM WITH BIOINFORMATION

To verify whether the introduction of bioinformation improves the detection ability of SHEIB-AGM, we applied SHEIB-AGM with bioinformation to analyze the seven GWAS data from WTCCC. We set $pb = 0.9$, $decRate = 0.01$, $cG = 0.05$,

TABLE 6. Some of the gene pairs of the epistastic interactions detected on the seven GWAS data using SHEIB-AGM with bioinformation. The complete table is shown in Table S12 in the Supplementary Appendix.

gene1	ctd1	gene2	ctd2	number of occurrences
Bipolar Disorder				
CABLES1	NDE	LOC107984008	NF	5362
CABLES1	NDE	STK32A-AS1	NF	5362
LOC105372022	NF	LOC107984008	NF	5319
LOC105372022	NF	STK32A-AS1	NF	5319
Coronary Artery Disease				
LOC105371749	NF	RPL10	NDE	4549
LOC105371748	NF	LOC107987332	NF	4549
LOC105371749	NF	LOC107987332	NF	4549
LOC105371748	NF	RPL10	NDE	4549
Crohn's Disease				
FAM155B	NF	MTRNR2L10	NF	357
PPM1A	NDE	PPM1A	NDE	334
SNHG11	NDE	SNHG14	NF	159
JPT2	NDE	SNHG14	NF	159
Hypertension				
NCK1-AS1	NF	SLC35G2	NDE	2552
SGSH	NDE	SLC26A11	NDE	2262
CCAR1	NDE	MIR1254-1	NF	1594
CCAR1	NDE	CCAR1	NDE	1594
Rheumatoid Arthritis				
GABRA3	NDE	MAGEA12	NDE	2199
CSAG4	NF	GABRA3	NDE	2199
GLRA4	NDE	LINC00630	NDE	1668
GLRA4	NDE	TMEM27	NDE	1628
Type 1 Diabetes				
LOC105379656	NF	LOC105379664	NF	17226
LOC105379664	NF	LOC107987429	NF	13917
LOC105379656	NF	LOC107987429	NF	13917
LOC105379664	NF	LOC105379664	NF	8613
Type 2 Diabetes				
LOC105377926	NF	LOC105377926	NF	3262
MAPKAPK5	NDE	TMEM116	NDE	2134
ADAM1A	NDE	MAPKAPK5	NDE	1875
CCAR1	NDE	MIR1254-1	NF	1531

TABLE 7. Some of the genes of the epistastic interactions detected on the seven GWAS data using SHEIB-AGM with bioinformation. The complete table is shown in Table S13 in the Supplementary Appendix.

gene	ctd	number of occurrences
Bipolar Disorder		
STK32A-AS1	NF	18328
LOC107984008	NF	18328
CABLES1	NDE	17394
LOC105372022	NF	17303
Coronary Artery Disease		
GABRA3	NDE	57583
SNORA70	NDE	49888
LOC107987332	NF	49888
RPL10	NDE	49888
Crohn's Disease		
PPM1A	NDE	392
FAM155B	NF	357
MTRNR2L10	NF	357
MEIOB	NF	213
Hypertension		
CCAR1	NDE	3189
LOC105377926	NF	3170
NCK1-AS1	NF	2552
SLC35G2	NDE	2552
Rheumatoid Arthritis		
GABRA3	NDE	33622
GLRA4	NDE	25130
unknown	NF	24566
BAG6	NDE	23587
Type 1 Diabetes		
LOC105379664	NF	48774
LOC105379656	NF	48774
LOC107987429	NF	38809
AGPAT1	NDE	31948
Type 2 Diabetes		
LOC105377926	NF	6668
CCAR1	NDE	3070
MAPKAPK5	NDE	2467
TMEM116	NDE	2159

$cGc = 0.05$, $o = -1$, $maxGen = 4 \times 10^7$, $seed = 0$, $rn = -1$, and other parameters as default values. The detected epistatic interactions are shown in Table 4. The detailed results can be found in Table S11 in the Supplementary Appendix. Compared to Table 3, with bioinformation, SHEIB-AGM can detect 33.94∼3069.40-times more epistatic interactions. This represents a good performance in detecting higher order epistatic interactions. The results demonstrate that Assumption 1 and Assumption 2 are reasonable. SHEIB-AGM can use bioinformation to greatly improve the detection ability.

Some of the detected epistatic interactions are shown in Table 5. The complete list of the interactions is given in Table S11 in the Supplementary Appendix. Based on the dbSNP database, many SNPs can be mapped to genes. We counted the number of occurrences for the genes and gene pairs. Table 6 and Table 7 show the occurrences of each gene and each gene pair, respectively. The genes and gene pairs with high numbers of occurrences may play a very important role in the corresponding disease. For each of the seven diseases, we searched for each detected gene on the CTD database (the Comparative Toxicogenomics Database). As shown in Table 6 and Table 7, some of the genes have DE (Direct Evidence) or NDE (Not Direct Evidence) on the CTD database. The genes that have NF (Not Found) on the CTD database may be helpful in further understanding the seven diseases. We have utilized Cytoscape [40] to generate SNP networks and gene networks for each disease. Fig. 8 and Fig. 9 are the SNP network and gene network for Bipolar Disorder. The networks for the other six diseases are shown in Fig. S3-S14 in the Supplementary Appendix.

IV. CONCLUSION

In this article, we propose a novel stochastic approach named SHEIB-AGM to detect epistatic interactions in GWAS. The approach maintains a gene matrix to manage the bioinformation. In each iteration, it randomly generates an SNP combination containing *mo* SNPs based on the gene matrix. The approach utilizes k2 to detect an epistatic interaction on the combination. According to the detection result, SHEIB-AGM updates the gene matrix. We have conducted extensive experiments on both simulated data and real GWAS data. The experimental results demonstrate that the proposed algorithm outperforms six existing methods: DECMDR, SNPHarvester, MACOED, AntEpiSeeker, HS-MMGKG and SEE. In addition, SHEIB-AGM can use bioinformation to greatly improve the detection ability. We believe that SHEIB-AGM is a powerful tool for helping us understand the pathogenesis of common and complex diseases.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

REFERENCES

[1] A. Collins, C. Lonjou, and N. E. Morton, "Genetic epidemiology of single-nucleotide polymorphisms," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 26, pp. 15173–15177, Dec. 1999, doi: 10.1073/pnas.96.26.15173.

[2] N. J. Schork, D. Fallin, and J. S. Lanchbury, "Single nucleotide polymorphisms and the future of genetic epidemiology," *Clin. Genet.*, vol. 58, no. 4, pp. 250–264, Mar. 2003, doi: 10.1034/j.1399-0004.2000.580402.x.

[3] F. Chen, G. K. Chen, D. O. Stram, R. C. Millikan, C. B. Ambrosone, E. M. John, L. Bernstein, W. Zheng, J. R. Palmer, J. J. Hu, and T. R. Rebbeck, "A genome-wide association study of breast cancer in women of African ancestry," *Hum. Genet.*, vol. 132, no. 1, pp. 39–48, Jan. 2013, doi: 10.1007/s00439-012-1214-y.

[4] S. L. Pulit, C. Stoneman, A. P. Morris, A. R. Wood, C. A. Glastonbury, J. Tyrrell, and J. Yang, "Meta-analysis of genome-wide association studies for body fat distribution in 694,649 individuals of European ancestry," *Hum. Mol. Genet.*, vol. 28, no. 1, pp. 166–174, Sep. 2018, doi: 10.1093/hmg/ddy327.

[5] L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, and P. M. Visscher, "Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry," *Hum. Mol. Genet.*, vol. 27, no. 20, pp. 3641–3649, Oct. 2018, doi: 10.1093/hmg/ddy271.

[6] L. T. Elliott, K. Sharp, F. Alfaro-Almagro, S. Shi, K. L. Miller, G. Douaud, J. Marchini, and S. M. Smith, "Genome-wide association studies of brain imaging phenotypes in UK Biobank," *Nature*, vol. 562, no. 7726, pp. 210–216, Oct. 2018, doi: 10.1038/s41586-018-0571-7.

[7] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Rev. Genet.*, vol. 6, no. 2, pp. 95–108, Feb. 2005, doi: 10.1038/nrg1521.

[8] J. Erdmann, T. Kessler, L. M. Venegas, and H. Schunkert, "A decade of genome-wide association studies for coronary artery disease: The challenges ahead," *Cardiovascular Res.*, vol. 114, no. 9, pp. 1241–1257, Mar. 2018, doi: 10.1093/cvr/cvy084.

[9] R. Misra and N. Arebi, "Re: Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease," *Gastroenterology*, vol. 152, no. 8, pp. 2082–2083, Jun. 2017, doi: 10.1053/j.gastro.2017.02.041.

[10] D. Chang, "A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci," *Nature Genet.*, vol. 49, pp. 1511–1516, Sep. 2017, doi: 10.1038/ng.3955.

[11] L. Fejerman, N. Ahmadiyeh, D. Hu, S. Huntsman, K. B. Beckman, J. L. Caswell, K. Tsung, E. M. John, and G. Torres-Mejia, "Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25k,," *Nature Commun.*, vol. 5, Oct. 2014, Art. no. 5260, doi: 10.1038/ncomms6260.

[12] L. H. Maguire, S. K. Handelman, X. Du, Y. Chen, T. H. Pers, and E. K. Speliotes, "Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease," *Nature Genet.*, vol. 50, no. 10, pp. 1359–1365, Oct. 2018, doi: 10.1038/s41588-018-0203-z.

[13] A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G. B. Chen, V. Emilsson, S. F. W. Meddens, and S. Oskarsson, "Genome-wide association study identifies 74 loci associated with educational attainment," *Nature*, vol. 533, pp. 539–542, May 2016, doi: 10.1038/nature17671.

[14] Z. Wang *et al.*, "Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor," *Nature Genet.*, vol. 49, no. 7, pp. 1141–1147, Jul. 2017, doi: 10.1038/ng.3879.

[15] Ö. Carlborg and C. S. Haley, "Epistasis: Too often neglected in complex trait studies?" *Nature Rev. Genet.*, vol. 5, no. 8, pp. 618–625, Aug. 2004, doi: 10.1038/nrg1407.

[16] H. J. Cordell, "Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans," *Hum. Mol. Genet.*, vol. 11, no. 20, pp. 2463–2468, Oct. 2002, doi: 10.1093/hmg/11.20.2463.

[17] H. J. Cordell, "Detecting gene–gene interactions that underlie human diseases," *Nature Rev. Genet.*, vol. 10, pp. 392–404, Jun. 2009, doi: 10.1038/nrg2579.

[18] T. F. Mackay and J. H. Moore, "Why epistasis is important for tackling complex human disease genetics," *Genome Med.*, vol. 6, no. 6, p. 125, 2014, doi: 10.1186/gm561.

[19] P. C. Phillips, "Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems," *Nature Rev. Genet.*, vol. 9, pp. 855–867, Nov. 2018, doi: 10.1038/nrg2452.

[20] W.-H. Wei, G. Hemani, and C. S. Haley, "Detecting epistasis in human complex traits," *Nature Rev. Genet.*, vol. 15, no. 11, pp. 722–733, Nov. 2014, doi: 10.1038/nrg3747.

[21] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Amer. J. Hum. Genet.*, vol. 69, no. 1, pp. 138–147, Jul. 2001, doi: 10.1086/321276.

[22] C.-H. Yang, L.-Y. Chuang, and Y.-D. Lin, "CMDR based differential evolution identifies the epistatic interaction in genome-wide association studies," *Bioinformatics*, vol. 33, no. 15, pp. 2354–2362, Aug. 2017, doi: 10.1093/bioinformatics/btx163.

[23] Z. Zhu, X. Tong, Z. Zhu, M. Liang, W. Cui, K. Su, M. D. Li, and J. Zhu, "Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes," *PLoS ONE*, vol. 8, no. 4, Apr. 2013, Art. no. e61943, doi: 10.1371/journal.pone.0061943.

[24] C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu, "SNPHarvester: A filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics*, vol. 25, no. 4, pp. 504–511, Feb. 2009, doi: 10.1093/bioinformatics/btn652.

[25] P.-J. Jing and H.-B. Shen, "MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies," *Bioinformatics*, vol. 31, no. 5, pp. 634–641, Mar. 2015, doi: 10.1093/bioinformatics/btu702.

[26] Y. Wang, "AntEpiSeeker: Detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Res. Notes*, vol. 3, Apr. 2010, Art. no. 117, doi: 10.1186/1756-0500-3-117.

[27] L. Yuan, C. Yuan, and D. Huang, "FAACOSE: A fast adaptive ant colony optimization algorithm for detecting SNP epistasis," *Complexity*, vol. 2017, Sep. 2017, Art. no. 5024867, doi: 10.1155/2017/5024867.

[28] Y. Sun, J. Shang, J.-X. Liu, S. Li, and C.-H. Zheng, "epiACO—A method for identifying epistasis based on ant Colony optimization algorithm," *BioData Mining*, vol. 10, Jul. 2017, Art. no. 23, doi: 10.1186/s13040-017-0143-7.

[29] L. Sun, G. Liu, L. Su, and R. Wang, "HS-MMGKG: A fast multi-objective harmony search algorithm for two-locus model detection in GWAS," *Current Bioinf.*, vol. 14, no. 8, pp. 749–761, Dec. 2019, doi: 10.2174/1574893614666190409110843.

[30] L. Sun, G. Liu, L. Su, and R. Wang, "SEE: A novel multi-objective evolutionary algorithm for identifying SNP epistasis in genome-wide association studies," *Biotechnol. Biotechnol. Equip.*, vol. 33, no. 1, pp. 529–547, Jan. 2019, doi: 10.1080/13102818.2019.1593052.

[31] P. Burton, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, Jun. 2007, doi: 10.1038/nature05911.

[32] C. J. Mattingly, G. T. Colby, J. N. Forrest, and J. L. Boyer, "The comparative toxicogenomics database (CTD)," *Environ. Health Perspect.*, vol. 111, no. 6, pp. 793–795, May 2003, doi: 10.1289/ehp.6028.

[33] 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," *Amer. J. Hum. Genet.*, vol. 87, no. 3, pp. 325–340, Sep. 2010, doi: 10.1016/j.ajhg.2010.07.021.

[34] S. T. Sherry, "DbSNP: The NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/nar/29.1.308.

[35] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp, "Database resources of the national center for biotechnology information," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 10–14, Jan. 2000, doi: 10.1093/nar/28.1.10.

[36] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, Oct. 1992, doi: 10.1007/bf00994110.

[37] D. H. Stamatis, *Essential Statistical Concepts for the Quality Professional*. Boca Raton, FL, USA: CRC Press, 2012, doi: 10.1201/b11909.

[38] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *BioData Mining*, vol. 5, Oct. 2012, Art. no. 16, doi: 10.1186/1756-0381-5-16.

[39] S. Tuo, J. Zhang, X. Yuan, Y. Zhang, and Z. Liu, "FHSA-SED: Two-locus model detection for genome-wide association study with harmony search algorithm," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, Art. no. e0150669, doi: 10.1371/journal.pone.0150669.

[40] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: Software for visualization and analysis of biological networks," in *Data Mining in Proteomics: From Standards to Applications*. Totowa, NJ, USA: Humana Press, 2011, pp. 291–303, doi: 10.1007/978-1-60761-987-1_18.

LIYAN SUN was born in Changchun, China, in 1987. He received the bachelor's degree in computer science from Jilin University, China, in 2010, where he is currently pursuing the joint master's and Ph.D. degrees in computer science. His research interests include machine learning, artificial intelligence, big data, evolution algorithm, and bioinformatics.

GUIXIA LIU received the bachelor's degree in computer science from Jilin University, in 1987, the master's degree in 1996, and the Ph.D. degree in 2007. She is currently a Professor with Jilin University. She has chaired three National Natural Science Foundation projects, participated in one science and technology development plan project of Jilin province and chaired one innovation fund project of Jilin University. For years, she has devoted herself to machine learning, computational intelligence, big data, cloud computing, and bioinformatics.

RONGQUAN WANG is currently pursuing the joint master's and Ph.D. degree in computer science with Jilin University. His researches focus on graph clustering algorithms, complex network analysis, machine learning, and bioinformatics.