OXFORD

Briefings in Bioinformatics, 22(6), 2021, 1–12

https://doi.org/10.1093/bib/bbab276
Problem solving protocol

GESLM algorithm for detecting causal SNPs in GWAS with multiple phenotypes

Ruiqi Lyu, Jianle Sun, Dong Xu, Qianxue Jiang, Chaochun Wei and Yue Zhang

Corresponding authors: Yue Zhang, Shanghai Jiao Tong University, Department of Bioinformatics and Biostatistics, Shanghai, 200240, China. E-mail: yue.zhang@sjtu.edu.cn; Chaochun Wei, Shanghai Jiao Tong University, Department of Bioinformatics and Biostatistics, Shanghai, 200240, China. E-mail: ccwei@sjtu.edu.cn.

Abstract

With the development of genome-wide association studies, how to gain information from a large scale of data has become an issue of common concern, since traditional methods are not fully developed to solve problems such as identifying loci-to-loci interactions (also known as epistasis). Previous epistatic studies mainly focused on local information with a single outcome (phenotype), while in this paper, we developed a two-stage global search algorithm, Greedy Equivalence Search with Local Modification (GESLM), to implement a global search of directed acyclic graph in order to identify genome-wide epistatic interactions with multiple outcome variables (phenotypes) in a case–control design. GESLM integrates the advantages of score-based methods and constraint-based methods to learn the phenotype-related Bayesian network and is powerful and robust to find the interaction structures that display both genetic associations with phenotypes and gene interactions. We compared GESLM with some common phenotype-related loci detecting methods in simulation studies. The results showed that our method improved the accuracy and efficiency compared with others, especially in an unbalanced case–control study. Besides, its application on the UK Biobank dataset suggested that our algorithm has great performance when handling genome-wide association data with more than one phenotype.

Key words: GWAS; global search; multiple-phenotype analysis; DAG

INTRODUCTION

Nowadays, as high-throughput technology advances, genome-wide association studies (GWAS) have been rapidly developed and the investigation of associated single-nucleotide polymorphisms (SNPs) and phenotypes is becoming more and more common. The original method of GWAS was genotyping individuals from a case–control study, then comparing the SNP distributions between two groups to identify the SNPs associated with the phenotypes [1]. But this technique can only estimate one locus with one phenotype at a time, which is not applicable to the complex situation of epistasis and pleiotropy [2]. Especially for the epistasis issue, people usually discuss this problem for one outcome case, but the methods for investigating polymorphism and analyzing data with multiple phenotypes are underdeveloped. Therefore, the design of robust and efficient multi-loci and multi-phenotype analysis methods is regarded as a key to overcome the bottlenecks of genetic association studies.

Ruiqi Lyu is currently a junior student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University.
Jianle Sun is currently a senior student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University.
Dong Xu is currently a junior student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University.
Qianxue Jiang is currently a junior student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University.
Chaochun Wei is a full professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His main research areas include functional element finding in genomes, metagenomics and high-performance computing for bioinformatics.
Yue Zhang is an associate professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His main research interests focus on Bayesian survival analysis, Bayesian network and missing values.
Submitted: 6 April 2021; **Received (in revised form):** 5 June 2021

Traditional methods of GWAS analysis with multiple phenotypes commonly employed some simplified techniques. For example, some studies have determined SNPs associated with two phenotypes (diseases) by calculating conjunction false discovery rate, $FDR_{trait1\&trait2}$, as the posterior probability that a given SNP is null for both phenotypes simultaneously, which is estimated conservatively by taking the minimum of conditional false discovery rate, $FDR_{trait1|trait2}$ or $FDR_{trait2|trait1}$. After that, they applied a random pruning procedure to control the linkage disequilibrium (LD) [3, 4]. Moreover, some articles used Manhattan plots to find significant SNPs in 'disease 1 group', 'disease 1 plus disease 2 group', 'disease 1 not disease 2 group', and then identified the overlap SNPs of the two diseases [5]. Obviously, those methods were not sufficient to detect epistasis efficiently and were vulnerable to false-positives, thus they were not feasible enough when confronted with complicated situations.

Single-phenotype multi-loci analysis methods have been developed for years, and they include statistical methods (such as penalized regression approaches [6, 7]), machine learning methods (such as support vector machine (SVM) [8, 9]) and some advanced modifications to recognize epistasis (such as some heuristic and step-wise search methods [10–14]). However, when applied to the multiple-phenotype analysis, they have to be performed on one phenotype after another. Since these phenotypes are not independent of each other, the loss of information of phenotypic interactions may lead to biased results. Thus, in order to detect multiple genes and phenotypes interactions at the same time, we used a structure learning method to approximate the directed acyclic graphs (DAGs) that model the causal structures of the dataset. DAG can be used to represent a probability distribution over a set of random variables, where the variables can be SNPs together with phenotypes. In such models, parents of some vertexes in the graph are understood as causes, while the edges have the meaning of causal influences [15]. We do not care too much about the orientation of edges in GWAS because SNPs are naturally the cause. We concentrate on the skeleton of the DAG, that is, the graph that has the same edges as the DAG but no directions. We proposed a two-step causal structure learning method that combines a score-based approach and a constraint-based approach to detect the possible skeleton of DAG that describes the interaction patterns of genes and phenotypes. The flowchart of this procedure is depicted in Figure 1.

In this paper, we first introduced the Greedy Equivalence Search with Local Modification (GESLM) algorithm for the application in multiple-phenotype GWAS based on Bayesian structure learning. Then the simulation process was given and the responses of different parameters were compared and tested and simulation results were displayed after that. Moreover, we introduced applications in the real-world UK Biobank dataset of our method for reference. Finally, we discussed the properties of our algorithm and existing challenges that need to be further improved.

METHODS

Greedy Equivalence Search with Local Modification

We proposed GESLM to search the underlying DAG of genetic associations with phenotypes and gene interactions in this research. GESLM is performed with a score-based step and a constraint-based step sequentially due to the different properties of the two methods in application. Generally, in the score-based approaches, Bayesian networks were treated as probabilistic models that use a score function to evaluate the fitting result, thus transforming a structural learning problem into a model selection problem [15]. As for the constraint-based approaches that verify the constraints from data, independence relationships between variables are employed to infer the network skeleton and then other Bayesian network properties are used to evaluate the direction of edges [16].

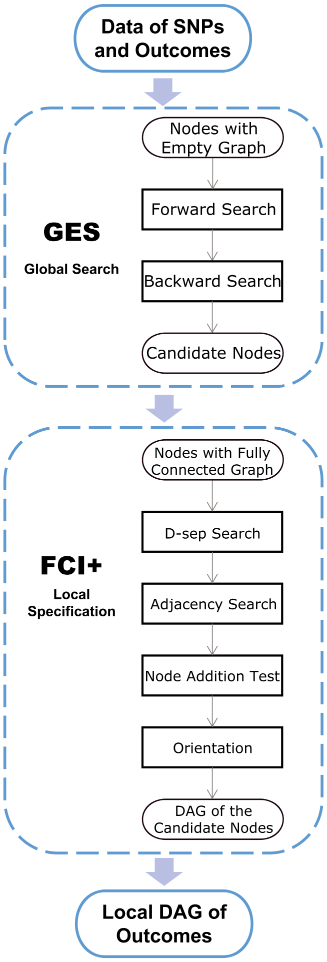

Figure 1. The figure shows the full workflow of our method that combines two stages of DAG search.

Comparatively, for network structure learning, score-based methods are usually found to be more effective [17]. Constraint-based approaches, particularly in the case of limited sample size, are prone to failure of conditional independence tests, resulting in unstable results [18]. Thus, score-based methods may produce more reliable results in the first stage. However, how to execute model correction is a challenging topic in the view of scoring functions since an inadequate search methodology can easily lead to a local optimum [19]. In this case, we added a constraint-based step to detect the local skeleton of DAG.

The first step of GESLM is a score-based method, Greedy Equivalent Search (GES), since it is more objective and accurate in high dimension datasets. We regard the first step as a dimensionality reduction procedure to select possible directly or indirectly related vertexes. Though GES performs well in the first step, a constraint-based approach, FCI+, should be employed as the second step to test for latent variables that may not be observed or removed in the previous step as well as other confounding factors (such as selection bias [19]). We select the outcome-related variables from the result of the first step, then perform the second step on a smaller dataset to find interactions more precisely. Due to the existence of epistasis interactions, which can be much stronger than the relationships of loci and phenotypes, a tuned significance threshold, α, can help us identify a number of relationships while reducing the risk of false-positives. The codes are available at https://github.com/Rachel-Lyu/GESLM and the R package pcalg [15, 20] is required.

Greedy Equivalent Search

GES is a prominent example of score-based learning [21]. It performs better than constraint-based approaches when tackling with high-dimensional interaction analysis [17]. Firstly, the score-based methods evaluate the model using a scoring criterion on a larger scale, so it can exclude the variables possible to pass the local conditional independence test, which may be confusing for constraint-based approaches [18]. Secondly, score-based methods do not need us to point out a significance level, α, since the criteria for the algorithm is to maximize the score rather than meet the threshold. It has been proven that as the sample size approaches infinity, the estimate of relationships can achieve consistency [22]. Hence, it is relatively easy to optimize the results by increasing the sample size.

GES uses a score-equivalent and decomposable score, such as a BIC score, to assess the causal structure [23]. Chickering [21] preferred the BIC score due to consistency; theoretically, any score equivalent and decomposable function is sufficient. GES is performed in two steps and the candidate with the highest score is selected in each step, or the step is terminated if no candidate has a score greater than the current graph G_{i+1} [21, 24]. A variant of the breadth-first search algorithm, lexicographic breadth-first search (LexBFS) that visits edges in lexicographic order, is used here to produce perfect elimination orderings of the DAGs, which plays an important role in the characterization of Markov equivalence classes [15]. In the forward step, the algorithm starts with an empty graph, $G_0 := ([p], \emptyset)$. It then sequentially goes from one graph G_i to a larger one, G_{i+1}, step by step, for which there are representatives $D_i \in D(G_i)$and $D_{i+1} \in D(G_{i+1})$ such that D_{i+1} has exactly one arrow more than D_i. In the backward step, the sequence $(G_i)_i$ continues by gradually stepping from one graph G_i to a smaller one, G_{i+1}, for which there are representatives $D_i \in D(G_i)$ and $D_{i+1} \in D(G_{i+1})$ such that D_{i+1} has exactly one arrow less than D_i. The algorithms are described in Algorithm 1, 2, 3.

Algorithm 1: AGES($\mathcal{T}, X$). Greedy Equivalence Search

```
Require: $(\mathcal{T}, X)$: data for targets $\mathcal{I}$
Ensure: $\mathcal{I}$-graph
G ← $([p], \emptyset)$
do
    DoContiue ← FALSE
    do
        $G_{old}$ ← G
        G ← ForwardStep(G; $\mathcal{T}, X$);          ▷ See Algorithm 2
    while $G_{old} \neq G$
    do
        $G_{old}$ ← G
        G ← BackwardStep(G; $\mathcal{T}, X$);         ▷ See Algorithm 3
        if $G_{old} \neq G$ then
            DoContinue ← TRUE
        end if
    while $G_{old} \neq G$
while DoContinue
```

Algorithm 2: ForwardStep(G; $\mathcal{T}, X$): $\mathcal{I}$-graph

```
Require: $G = ([p], E)$: $\mathcal{I}$-graph; $(\mathcal{T}, X)$: data for $\mathcal{I}$
Ensure: $G' \in \mathcal{E}^{+}_{\mathcal{I}}(G)$, or G
$\Delta S_{\max} \leftarrow 0$;
for each $v \in [p]$ do
    for each $u \in [p] \setminus \mathrm{ad}_G(v)$ do
        $N \leftarrow \mathrm{ne}_G(v) \cap \mathrm{ad}_G(u)$
        for each clique $C \subset ne_G(v)$ with $N \subset C$ do
            if $\nexists$ path from $v$ to $u$ in $G[[p] \setminus C]$ then
                $\Delta S \leftarrow s\left(\mathrm{pa}_G(v) \cup C \cup \{u\}\right) - s\left(\mathrm{pa}_G(v) \cup C\right)$
                if $\Delta S > \Delta S_{\max}$ then
                    $\Delta S_{\max} \leftarrow \Delta S$
                    $(u_{\max}, v_{\max}, C_{\max}) \leftarrow (u, v, C)$
                end if
            end if
        end for
    end for
end for
if $\Delta S_{\max} > 0$ then
    $\sigma \leftarrow$ LexBFS $\left((C_{\max}, v_{\max}, \ldots), E\left[T_G(v_{\max})\right]\right)$
    Orient edges of $G[T_G(v_{\max})]$ according to $\sigma$
    Insert edge $(u_{\max}, v_{\max})$ into G
    while $\exists\, a \rightarrow b \in G$ s.t. $a \rightarrow b$ ¬strongly $\mathcal{I} - protected$ do
        $G \leftarrow G + (b, a)$
    end while return G
else return G
end if
```

Fast Causal Inference+

Though GES performs well at the first stage, a constraint-based method allowing for local specification is needed as the second step. A typical example of constraint-based learning is PC algorithm, which is the basis of almost all constraint-based algorithms for estimating the completed partially directed acyclic graphs (CPDAGs [25], or essential graphs [26]) of the true causal structure. It consists of two steps including the adjacency search step and the orientation step [19]. However, PC algorithm may run into trouble when extended to causal models that do not guarantee causal sufficiency, which means not all the related variables are measured and recorded [27]. In this case, some separating sets, which may require nodes not to be adjacent to any of the separated nodes, are possible to be missed. To tackle this problem, Spirtes et al. [28] developed the so-called Fast Causal Inference (FCI) algorithm that introduces an additional step for adjacency search, which makes the algorithm allow for hidden variables, but it suffers from exponential running time in the worst case even if the underlying graph is sparse. Claassen et al. [19] improved the efficiency of FCI and proposed the FCI+ algorithm. FCI+ yields the true skeleton of DAG, with running time in the worst case polynomial in the number of nodes for sparse graphs. It assumes faithfulness and an underlying causal DAG, but allows for latent variables and selection bias.

The FCI+ algorithm in Algorithm 4 starts with the PC adjacency search where an initial skeleton is found from a fully connected undirected graph G: for each pair of nodes that are still connected in G, it searches for a subset of adjacent nodes Z to separate the pair of nodes; if found, the edge is deleted. By

Algorithm 3: BackwardStep($G; \mathcal{T}, X$): $\mathcal{I}$-Phenotypes

```
Require: $G = ([p], E)$: $\mathcal{I}$-Phenotypes; $(\mathcal{T}, X)$: data for $\mathcal{I}$
Ensure: $G' \in \mathcal{E}_{\mathcal{I}}^{-}(G)$, or G
$\Delta S_{\max} \leftarrow 0$;
for each $v \in [p]$ do
    for each $u \in \mathrm{ne}_G(v) \cup \mathrm{pa}_G(v)$ do
        $N \leftarrow \mathrm{ne}_G(v) \cap \mathrm{ad}_G(u)$
        for clique $C \subset N$ do
            $\Delta S \leftarrow s\left(\mathrm{pa}_G(v) \cup C \backslash \{u\}\right) - s\left(\mathrm{pa}_G(v) \cup C \cup \{u\}\right)$
            if $\Delta S > \Delta S_{\max}$ then
                $\Delta S_{\max} \leftarrow \Delta S$
                $(u_{\max}, v_{\max}, C_{\max}) \leftarrow (u, v, C)$
            end if
        end for
    end for
end for
if $\Delta S_{\max} > 0$ then
    if $u_{\max} \in \mathrm{ne}_G(v_{\max})$ then
        $\sigma \leftarrow \mathrm{LexBFS}\left((C_{\max}, u_{\max}, v_{\max}, \ldots), E\left[T_G(v_{\max})\right]\right)$
    else
        $\sigma \leftarrow \mathrm{LexBFS}\left((C_{\max}, v_{\max}, \ldots), E\left[T_G(v_{\max})\right]\right)$
    end if
    Orient edges of $G[T_G(v_{\max})]$ according to $\sigma$
    Remove edge $(u_{\max}, v_{\max})$ from G
    while $\exists\, a \rightarrow b \in G$ s.t. $a \rightarrow b\neg$ strongly $\mathcal{I}$-protected do
        $G \leftarrow G + (b, a)$
    end while return G
else return G
end if
```

Algorithm 4: FCI+ Algorithm

```
Require: independence oracle $\mathcal{O}$ for variables $\mathbf{V}$, sparsity $k$
Ensure: causal model G over $\mathbf{V}$
$G \leftarrow$ fully connected undirected graph over $\mathbf{V}$
$n = 0$
while $\exists$ X with $|\mathrm{Adj}_G(X)| >$ n do
    while $\exists$ edge $X - Y$ in G have not been checked do
        select X with $|Adj_G(X)| > n$, select Y$\in \mathrm{Adj}_G(X)$
        while $\exists$ subsets size n have not been tested do
            select subset $\mathbf{Z}$ size $n$ from $Adj_G(X) \backslash Y$
            if $X \perp Y | Z$ then
                $Sepset(X, Y) = Sepset(Y, X) = \mathbf{Z}$, Remove edgeX − YfromG
            end if
        end while
    end while $n = n + 1$
end while                                          ▷ Finish Adjacency Search
$G, \mathcal{I} \leftarrow$ causal model $G$, minimal $Sepset\ \mathcal{I}$
$G^{+} \leftarrow AugmentGraph(G, \mathcal{I}, \mathcal{O})$      ▷ Test for single node additions destroying independence
$PosDsepLinks \leftarrow GetPDseps\left(G^{+}\right)$          ▷ Identify edges corresponding to possible D-sep
while $PosDsepLinks \neq \varnothing$ do
    $X, Y \leftarrow Pop(PosDsepLinks)$, $BaseX \leftarrow Adj(X)\backslash_{Y}$, $BaseY \leftarrow Adj(Y)\backslash_{X}$
    for $n = 1 \ldots k$ do
        for $m = 1 \ldots k$ do
            get subset $\mathbf{Z}_X \subseteq BaseX$, size $n$, get subset $\mathbf{Z}_Y \subseteq BaseY$, size $m$
            $\mathbf{Z}^{*} \leftarrow HIE\left(\{X, Y\} \cup \mathbf{Z}_X \cup \mathbf{Z}_Y, \mathcal{I}\right) \backslash_{\{X,Y\}}$
            if $X \perp Y \mid \mathbf{Z}^{*}$ then
                $\mathbf{Z} \leftarrow MinimalDsep\left(X, Y, \mathbf{Z}^{*}\right)$
                $\mathcal{I} \leftarrow UpdateSepsets(\mathcal{I}, X, Y, \mathbf{Z})$
                $G^{+} \leftarrow AugmentGraph\left(G^{+}, \mathcal{I}, \mathcal{O}\right)$
                $PosDsepLinks \leftarrow GetPDseps$
                (continue while)
            end if
        end for
    end for
end while                                          ▷ Finish D-sep Search
for all unshielded triples $X - Z - Y$ in G do
    if $Z \notin Sepset(X, Y)$ then
        orient $v$-structure $X \rightarrow Z \leftarrow Y$
    end if
end for
Run other orientation rules until no more new
    return causal model G                          ▷ Finish Orientation Step
```

checking all adjacent node pairs in G for possible separating sets of increasing size, the algorithm ensures that it finds separating sets as small as possible [19, 25]. Then test for single-node additions that destroy the independence, which is the basis for identifying the edges corresponding to possible D-sep links. This list is processed and updated along the way until no more unchecked possible D-sep links remain. For a pair of nodes X-Y on a possible D-sep edge in G^{+} the 'Base' of adjacent nodes (possible ancestors) is determined. For each combination of possible nodes from this base around X and Y (not restricted to adjacency sets of X and Y), the corresponding hierarchy is computed and tested for independence. If found, it is converted into a minimal separating set and stored in the separating set list. This is used to update the augmented skeleton G^{+} and to update the set of possible D-sep links. Finally, unshielded colliders in the updated skeleton are oriented based on the updated list of separating sets, then further orientation rules are applied.

SIMULATION

Datasets generation

We evaluated and compared the performance of our method and several other approaches using the simulated datasets generated from a common two-loci disease model [29, 30], whose disease odds for every genotype are displayed in Table 1. Two directly related loci of each disease have an independent multiplicative genotype effect. On the appropriate scale, this model is additive and has marginal effects that should be 'detectable' independent of other loci.

The model specified that the odds of disease increase in a multiplicative fashion both within and between two loci. In this model, an individual heterozygous at locus A has increased odds of $(1+\theta)$ relative to those of an individual who is homozygous aa; the AA homozygote has further multiplicative odds of $(1+\theta)^2$. The similar effects of locus B are also reflected in θ, and the odds of disease for each combination of genotypes at loci A and B are the product of the two within-locus effects.

We used α and θ to denote the baseline effect and genotype effect respectively. For the sake of simplicity, we introduced a few parameters to reflect the dataset's characteristics: a marginal

Table 1. The odds of the two-loci disease model

	bb	Bb	BB
aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
Aa	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$
AA	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	$\alpha(1+\theta)^4$

α, base-line effect; θ, genotype effect

Table 2. Four sets of parameters that investigate the properties of the algorithm

λ	r^2	MAF	size
0.3	0.7	0.05, 0.1, 0.2, 0.5	500:5000, by=500
0.3	0.7:0.99, by=0.01	0.05, 0.1, 0.2, 0.5	5000
0.3	0.7,0.9	0.05, 0.1, 0.2, 0.5	5000
0.3	0.7,0.9	0.05, 0.1, 0.2, 0.5	5000

Four sets of parameters were used to generate datasets that examine the influence of sample size and linkage imbalance and compare GESLM with other algorithms in both balanced and unbalanced datasets.

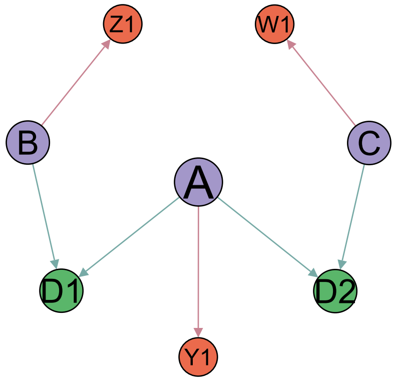

Figure 2. DAG that describes the interactions simulated datasets.

parameter, λ; a disease prevalence, p; the minor allele frequency, MAF; and the LD (measured by the parameter r^2). First, we specified the MAF of the disease loci, λ, p and r^2. Under the assumption of Hardy–Weinberg law, we could deduce the value of α and θ from the expressions of λ and p, where D represents an individual has the disease, $\bar{D}$ represents an individual who doesn't have the diseases and g_A, g_B are genotypes. The expressions of λ and p can be written as Equations 1 and 2:

$$\lambda = \frac{p\left(D|1_A\right)}{p\left(\bar{D}|1_A\right)} \Big/ \frac{p\left(D|0_A\right)}{p\left(\bar{D}|0_A\right)} - 1 \tag{1}$$

$$p = p(D) = \sum_{g_A, g_B} p\left(D|g_A, g_B\right) p\left(g_A, g_B\right) = 0.1 \tag{2}$$

We could also calculate the conditional probability of the locus having LD with disease locus given the allele of disease locus using r^2. In population genetics, LD describes a phenomenon that the existence of non-random associations between different genetic markers in a given population. The allele frequencies were written as $\pi_A, \pi_B, \pi_a, \pi_b$, and the haplotype frequencies were written as $\pi_{AB}, \pi_{Ab}, \pi_{aB}, \pi_{ab}$. Then, the expression of r^2 can be written as Equation 3:

$$r^2 = \frac{\left(\pi_{AB} - \pi_A \pi_B\right)^2}{\pi_A \pi_B \pi_a \pi_b} \tag{3}$$

Once the parameters were prepared, we could generate the disease status at predetermined proportions, and genotypes of the disease loci could be generated afterward. Then the disease loci genotypes could be used to generate genotypes of their associated loci.

In this study, we chose different sets of parameters as shown in Table 2 to guarantee the generality of the experiment, whose parameter combination was similar to the study of Han et al. [31]. For each parameter setting, we generated 50 datasets, each of which contains 106 SNPs. There were two diseases, $D1$ and $D2$, involved where SNP A and B were directly related to $D1$, A and C were directly related to $D2$, while Y1, Z1, W1 were directly related to A, B, C, respectively, but have no significant association with the two diseases. The MAFs of each non-disease marker were randomly generated from a uniform distribution between (0, 0.5]. The relations could be described by Figure 2.

Comparison with other algorithms

We compared our algorithm with Chi-squared test, elastic net, BOOST and bNEAT, then histograms in Figures 3 and 4 were used to show the simulation results. Since the methods have different types of output such as p-values in Chi-squared test and BOOST, the ranks of coefficient value in elastic net, and the candidate node lists connected to the outcomes in bNEAT and GESLM. To assess the results, we defined power as the proportion of datasets that accurately recorded diseases and associated loci without false-positives. We selected the first two variables in p-value and coefficient rank as the result of Chi-squared test, elastic net regression and BOOST, while the candidate node list itself as the result of GESLM and bNEAT. The criteria seem to be unfair to the latter two algorithms since it is hard to determine the number of variables in the real world, (here the number is two), but the result could reflect their characteristics in performance to some extent. The powers of different methods were calculated and compared under the parameter in line 3 of Table 2. We compared the performance of these approaches under different r^2 and MAF values to estimate their stability.

A straightforward way to do multi-loci analysis is Chi-squared test, which is performed on one SNP at a time, then set a threshold or order the p-value to select the significant ones. However, the Chi-squared test assumes that the SNPs are independent of each other. Additionally, the tuned p-value threshold or the number of selected variables can be tricky in application. What is more, because of the existence of epistatic interactions in the real world, interacting SNPs often have too close χ^2-statistics in Chi-squared tests. In this case, we cannot distinguish the directly and indirectly related SNPs, thus some false-positive loci can be introduced and weak interactions can also easily be ignored.

Penalized regression approaches, also called shrinkage or regularization methods, offer an attractive alternative to SNP detection in GWAS [32]. Penalized logistic regression methods shrink down to zero the coefficient of markers that have no significant effect on the phenotypes of interest, resulting in a parsimonious subset of what we would expect to be truly pertinent predictors. Wan et al. [33] showed that penalized methods outperform single marker analysis, with the main difference that penalized methods allow the simultaneous inclusion of

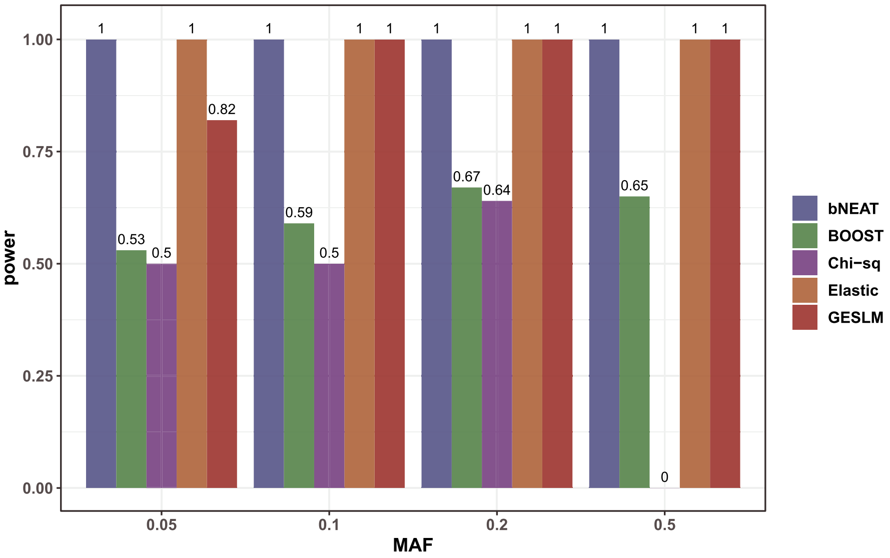

Figure 3. Comparison of different algorithms when $r^2 = 0.7$. Under the parameter in line 3 of Table 2.

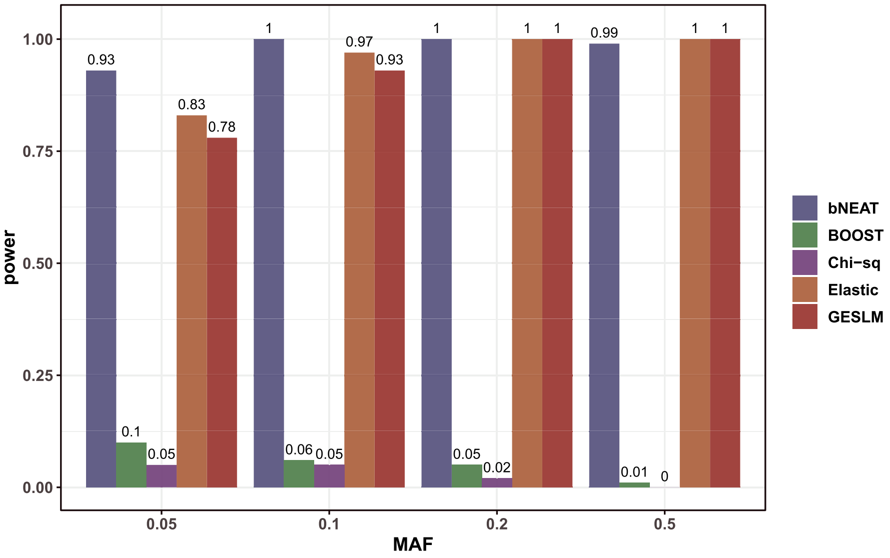

Figure 4. Comparison of different algorithms when $r^2 = 0.9$. Under the parameter in line 3 of Table 2.

a number of markers and generally do not allow correlated variables to enter the model, producing a sparse model in which most of the identified explanatory markers are accounted for [7]. The loss function of elastic-net is defined as $L(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2 + \lambda(\alpha\|\beta\|_1 + \frac{(1-\alpha)}{2}\|\beta\|_2^2)$, which creates a useful a compromise between the ridge-regression penalty whose $\alpha = 0$ and the LASSO penalty whose $\alpha = 1$. The elastic net with $\alpha = 1 - \epsilon$ for some small $\epsilon > 0$ performs much like LASSO, but is robust to extreme correlations among predictor variables [34]. The choice of α and the regularization parameter λ is critical to selecting important variables with accurate estimation and tuning parameter are usually selected to minimize mean-squared prediction error based on cross-validations, which can be time and space consuming. Also, more samples are required to ensure the stability of the estimation with the exponential increase in possible combinations [35]. Moreover, the shrinkage procedures allow for variable selection, and only important predictors remain in the model [36], whereas non-causal factors may perform well in the prediction, which may introduce false-positives. We implement the elastic net penalized regression in R.

BOOST (BOolean Operation-based Screening and Testing) is a computationally efficient two-stage statistical method applied to analyze all pairwise interactions in genome-wide data [33]. BOOST designed a Boolean representation of genotype data that not only improves space efficiency but also increases CPU efficiency as it only contains Boolean values and can be used to perform fast logistic regression calculations on contingency tables. BOOST uses a two-stage search method: In the filtering stage, a non-iterative method is used to calculate the approximate likelihood ratio to evaluate all site pairs, and SNP pairs that are greater than the specified threshold are selected; in the inspection stage, a classical likelihood test is used to measure the interaction of selected SNP pairs. The highlight of BOOST is the ability to analyze SNP pairs in the whole genome. During filtration, all SNP pairs are analyzed, so the single locus with weak main effects but strong epistasis effects are not filtered

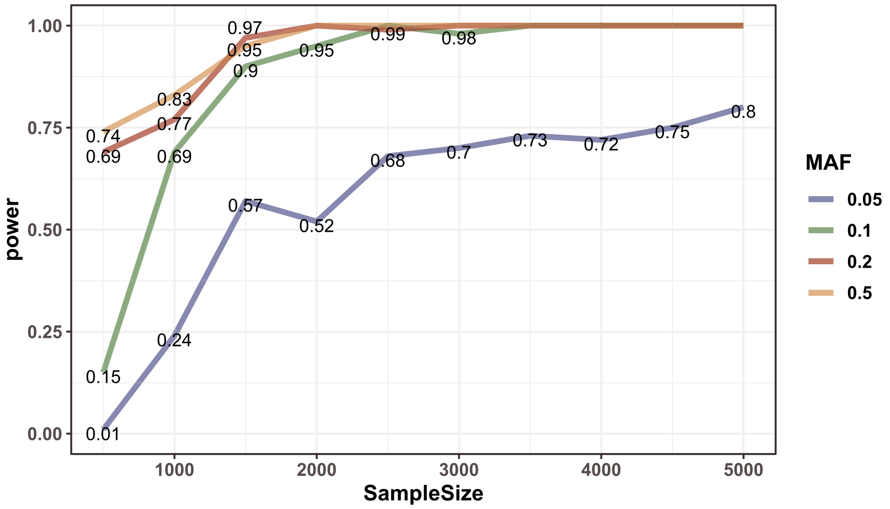

Figure 5. Power under different sample size. Under the parameter in line 1 of Table 2.

out. However, in the absence of marginal effects, BOOST has been shown to produce many false-positive results [37]. BOOST software we used here is downloaded from http://bioinformatics.ust.hk/BOOST.html.

bNEAT is a Bayesian network method that is based on a score-based approach and is suitable for data with small sample size. Though performing better than DASSO-MB [38], it is based on a greedy search program and is sensitive to improper input orders. Unfortunately, although designers attempt to deliberately reduce the complexity of the calculations, it is still difficult to apply this method directly to the GWAS data. The pseudocode is shown by Han et al. [38] and it is implemented in R.

We can see the results displayed in Figures 3 and 4. In most cases, the performance of GESLM was closed to that of bNEAT and elastic net, but higher than the other two methods. Part of the reason why other approaches may not work well is that the idea of using only p-value as the inclusion criteria can introduce many false-positives.

Under balanced simulation, GESLM, bNEAT and elastic net are all good in accuracy, though GESLM runs faster than the other two (see Section 3.6). When MAF is small, GESLM is greatly influenced, which could also be noticed in unbalanced settings according to Figures 7 and 8. In other situations, three algorithms had the inspiring power close to 100%. Although our algorithm did not work well when MAF is small, loci with small MAF are often deleted in quality control procedures. Thus, the disadvantage of our algorithm is not fatal.

Sample size

It is said that when confronted with a large sample size, GES algorithm provably identifies a perfect DAG of the generative distribution [21]. In order to investigate the asymptotic property of GES, we changed the sample size from 500 to 5000, and the results are displayed in Figure 5.

As the sample size increases, the power of GESLM also increases. Under the parameter combination of our dataset, only when the sample size exceeds about 1500 will GESLM have satisfactory results. Too few samples may lead to false-positive results, so we can enhance the effectiveness of the search algorithm by enlarging the sample size. What is more, when MAF is as small as 0.05, it is difficult to improve the accuracy by enlarging the sample size and the quality control of MAF needs to be relatively strict.

Linkage disequilibrium

LD (measured by the parameter r^2) can also be a vital factor that influences the accuracy of the results. To examine the changes of power as LD varies, we changed r^2 from 0.7 to 0.99. It can be found that with the increase of LD effect, the search accuracy will be inhibited and the influence on power in the cases of low MAF could be greater than that in the cases of high MAF. Therefore, if there is too strong LD, especially when the MAF is small, we should pretreat the data beforehand. The quality control process often discards SNPs with MAF less than a certain threshold, where 0.05 and 0.1 are commonly used in GWAS, since it is hard to detect associations with rare variants and people usually select against low MAF values [39]. This process may help to improve the performance of our algorithm. The influence of different LD and MAF are described in Figure 6.

Unbalanced sample

In the real world, the investigation into SNP-phenotype interaction may suffer from the following troubles: if we want to study some rare diseases, the sample size of the case group may be small, and the overlap sample of two diseases may be difficult to find. Thus, we generated a dataset for an unbalanced sample where the outcome variables consist of common disease and rare disease. The sample size of each group in the simulated sample was $\bar{D}_1\bar{D}_2 : D_1\bar{D}_2 : \bar{D}_1D_2 : D_1D_2 = 2000 : 2000 : 900 : 100$, which is close to the real data we applied our algorithm into, and the sample size in each group was $\bar{D}_1\bar{D}_2 : D_1\bar{D}_2 : \bar{D}_1D_2 : D_1D_2 = 2000 : 2000 : 840 : 103$. Then we used the unbalanced data to compare the performance of each algorithm. The result is displayed in Figures 7 and 8.

In order to compare GESLM with other algorithms, we noted that an unbalanced sample may reduce the power of bNEAT, BOOST and Chi-squared test, while GESLM and elastic net were not significantly affected. With the increase of MAF, the inhibition effect is more and more obvious, which is contrary to GESLM and elastic net that have better performance at high MAF. Under

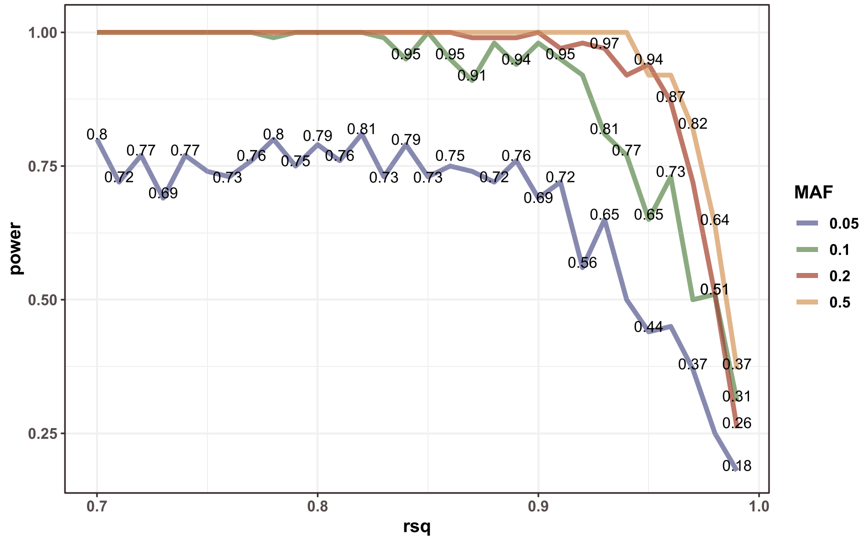

Figure 6. Comparison of different algorithms when sample size is unbalanced with $r^2 = 0.9$. Under the parameter in line 4 of Table 2.

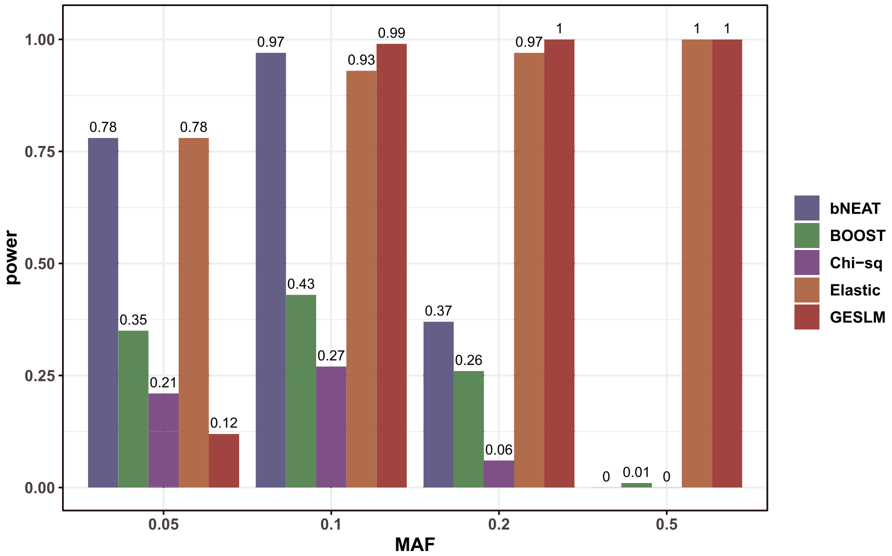

Figure 7. Comparison of different algorithms when sample size is unbalanced with $r^2 = 0.7$. Under the parameter in line 4 of Table 2.

this unbalanced scenario, the power of GESLM surpasses that of elastic net when MAF is no less than 0.1, suggesting that our algorithm is more resilient to the shift in case-control group settings. Thus, if we perform a strict quality control where the threshold of MAF is relatively high, GESLM can be suitable even if the sample is unbalanced.

Running time

We ran all of the algorithms on the CPU (Intel Xeon Gold 6150 Processor with 24.75M Cache, 2.70 GHz) and evaluated their time efficiency. The algorithms were performed in 800 files on the balanced settings (under the parameter in line 3 of Table 2), and the average time consumption of each file is displayed in Figure 9.

In this situation, bNEAT spent the longest time (6.620s on each file), while elastic net (with tuned λ and α, did not include the time of tuning) and BOOST spent 1.145s and 1.060s on each file respectively. GESLM consumed 0.251s on each file, and had similar calculation efficiency to the Chi-squared test (0.191s on each file), and outperformed other algorithms. For some algorithms that need to learn Bayesian networks, due to the subtlety of Markov blanket search, they perform well in terms of accuracy, but are not satisfactory in terms of time efficiency. As for penalized regression such as LASSO and elastic net, the tuning procedure is important but takes time. However, the greedy procedure of GES step avoids the time-consuming step of DAG search and strikes a good balance between efficiency and accuracy.

APPLICATION

We applied our algorithm to the real-world dataset, UK Biobank, to evaluate its adaptability to the real-world situation. Parkinson's disease (PD) is a common kind of neurodegenerative disease and there is considerable evidence supporting that the anormal gather of α-synuclein proteins, encoded by SNCA [40]. Salbutamol has been used quite widely for the treatment of

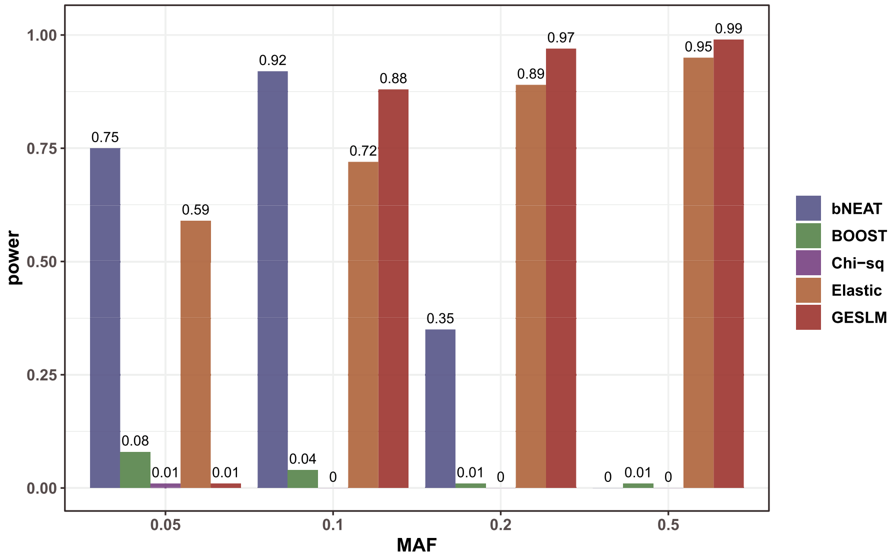

Figure 8. Power under different LD and MAF. Under the parameter in line 2 of Table 2.

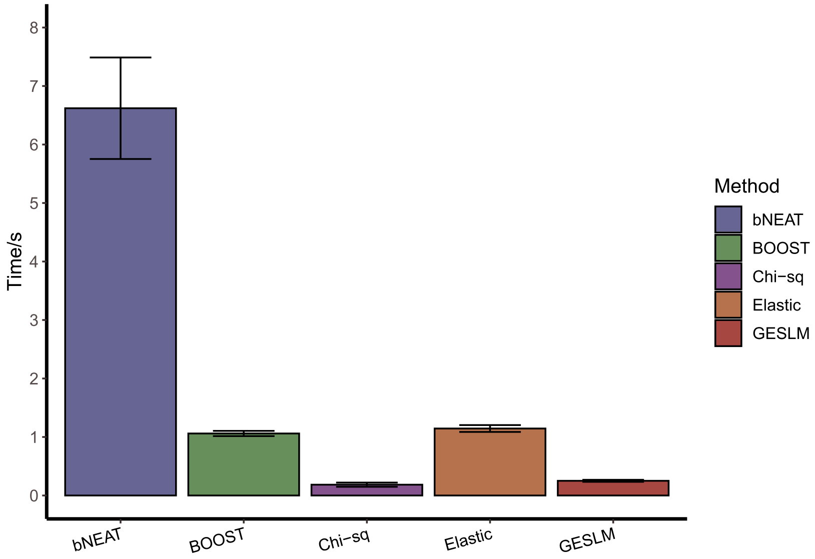

Figure 9. Average time consumption of each algorithm on one file under the parameter in line 3 of Table 2.

asthma since it shows the ability to bronchiectasis by simulating the β_2-adrenoceptors on tracheal smooth muscle cells and then changing the intracellular concentration of cAMP [41]. A previous study [42] showed that β_2-adrenergic agonist can also perform as a regulator of *SNCA*, and salbutamol has already been observed to be associated with reduced risk of PD through a cohort study, which indicates the potential relationship between PD and asthma, and that is why we were encouraged to choose those two diseases as a trial. We chose the case group by selecting the persons who were infected with asthma (D_1) or PD (D_2). $\bar{D}_i$ represents an individual who does not have the diseases. The sample size of each group in the simulated sample was $\bar{D}_1\bar{D}_2 : D_1\bar{D}_2 : \bar{D}_1D_2 : D_1D_2 = 2000 : 2000 : 840 : 103$. The control group included 2,000 people who did not have any mental illness or asthma. After the quality control procedure that excluded core genes and rare genes, GESLM was performed on each chromosome. The final results are shown in Table 3.

We then conducted a search of related articles to verify our results. According to the articles, with regards to asthma-related SNPs, rs10782001 may have overlapping effects on asthma and psoriasis [5]. rs6926374 is located in the HLA-DR/DQ region that is highly associated with asthma, and the region possibly contributes to changes in gene expression levels and antigen recognition procedures related to asthma [43]. rs755023315 was also identified in a large-scale genetic analysis for asthma [44].

In terms of PD SNPs, rs3813020 can be of high risk when using ATAC-seq to identify SNPs located in regions of open chromatin in iPSC-derived microglia [45]. rs13227860 is contained in the generated genotype database for PD patients and controls [46]. Mutations in the gene encoding leucine-rich repeat kinase 2 (*LRRK2*), where rs73277531, rs73277533 are located, have been linked with autosomal-dominant parkinsonism that is clinically indistinguishable from typical, idiopathic, late-onset PD [47]. And rs11158026 can affects early PD risk through altered dopamine uptake [48]. In addition, from a pathway-based asso-

Table 3. Result of the analysis using UK Biobank Dataset

Disease	SNP
asthma (D_1)	rs9820645_C, rs9274552_A, rs9486932_A, rs10225384_T, rs10244830_T, rs1429648_C, rs6474018_C, rs10733512_T, rs7485486_G, rs707555_G, rs9661394_G, rs2660312_T, rs987267_G, rs16835030_G, rs59657202_T, rs1446553_A, rs78153984_C, rs73069896_T, rs76917570_G, rs301961_G, rs4912540_A, rs35454701_T, rs12646270_A, rs72709216_A, rs76617725_C, rs10493161_C, rs11768648_A, rs7785272_G, rs10081610_C, rs11998723_C, rs10109493_G, rs10086947_C, rs77007376_G, rs77203213_A, rs7916384_C, rs3762086_A, rs11003050_A, rs11050523_C, rs2129140_T, rs1039302_T, rs9318053_T, rs1028531_C, rs9319588_C, rs11150600_C, rs34873012_T, rs10782001_G, rs4889514_A, rs12928852_C, rs12599631_A, rs750408_G, rs9938050_G, rs897986_T, rs8058320_G, rs755023315_G, rs12944467_A, rs4090488_G, rs112365422_A, rs62125146_G, rs4806093_A, rs74777463_C, rs2024564_A, rs551438_C, rs9977638_T
PD (D_2)	rs13303010_G, rs9820645_C, rs115354364_G, rs78153984_C, rs79719770_C, rs10244830_T, rs35735067_T, rs16922295_C, rs4921739_C, rs963475_T, rs4361809_T, rs7916384_C, rs10774568_G, rs11158026_T, rs200746_T, rs4845528_C, rs1514681_C, rs792068_C, rs74432250_T, rs17588199_T, rs11678541_T, rs4561907_A, rs3130286_T, rs3763309_A, rs3763312_A, rs4348358_A, rs9268605_A, rs9268606_A, rs9268607_G, rs9268608_T, rs9268609_A, rs111756805_C, rs10946101_G, rs4243839_C, rs35570345_T, rs73277533_A, rs12944467_A, rs4090488_G
Overlap	rs6926374_G, rs3813020_G, rs116531886_G, rs2152750_T, rs13227860_A, rs73102553_A, X9_29725362_AG_A_A, rs4748900_T, rs12930545_A, rs9938550_A, rs73277531_G

According to the result, 63 SNPs only relates to asthma and 38 SNPs only relates to PD. 11 overlap SNPs associated with both diseases.

ciation study, rs9938550 is in a pathway related to the bile acid metabolic process and steroid metabolic process to contribute to PD susceptibility [49]. And rs3763312 can be an independent pleiotropic loci in PD [4]. rs4921739 was identified as a novel PD risk loci as a lead SNP of *ZDHHC2* [50].

DISCUSSION

Compared with other algorithms we used, the GESLM algorithm is robust and efficient in detecting epistatic interactions of multiple phenotypes on both balanced and unbalanced datasets. For penalized regression methods, the unrelated SNPs may perform well in predicting the output variables if LD exists, thus the bias can be significant. When MAF is relatively large, Chi-squared test and BOOST are problematic, because corresponding statistics of SNPs in LD cases can become too close. And the interactions between SNPs are much stronger than the SNP-phenotype relationship, introducing large confounders. When the positive and negative samples are unbalanced, in bNEAT, the scoring function of different phenotypes may be distorted and the score-and-search procedure may be affected, resulting in a poor performance in the unbalanced sample dataset. However, in terms of the GESLM algorithm, for one thing, in both GES and FCI+ process, the dependent variables and the independent variables are not distinguished, thus the relationship between the phenotypes and SNPs can be found, as well as the relationship between SNPs. Moreover, borrowing global structure information to specify local relations can be more objective and the second step FCI+, which considers latent variables and selection bias [19], can be a double-robust process [24].

The GESLM algorithm has good properties in improving the recognition efficiency and reducing false-positives, and has great application value. Under different sample sizes and parameter settings, our method searched for SNPs associated with two phenotypes with high accuracy and efficiency. Compared with other tests, the advantages of the GESLM algorithm are (i) it achieves a balance between effectiveness and time-complexity; (ii) there can be fewer false-positive results; and (iii) it can present search results in the form of graphs instead of trees or sets. Certainly, GESLM also has its shortcomings. For example, for the first step global search process in high-dimensional data, the local graph is difficult to be completely correct due to various interference and latent variables, and the time complexity remains a concern. Moreover, in the second step of FCI+ local search, the accuracy is affected when MAF is small and LD is relatively large. Therefore, if GESLM is used in GWAS, the quality control requirements are supposed to be more stringent, which also limits the scope of application in some aspects. To ameliorate the time complexity, a Markov blanket procedure can be added before the GES step to construct a set of possible adjacencies to search among, serving as a variable selection process. It will generally have less running time of applying GESLM directly to the full dataset [51]. We hope to improve our algorithm in future studies, so that greater tools can be used in the field of GWAS.

Key Points

- Genome-wide association study with multiple phenotypes is very important. It is necessary to consider epistatic interactions with global information and local specification.
- Directed acyclic graph is good at displaying genetic associations together with phenotypes and gene interactions. Greedy Equivalence Search with Local Modification (GESLM) outperforms some other phenotype-related loci detecting methods in simulation studies in accuracy and efficiency.
- GESLM does well in handling genome-wide association data with multiple phenotypes, especially in an unbalanced case–control study.

Acknowledgments

We would like to thank the anonymous reviewers and the editor for their helpful comments and valuable suggestions, which improved this paper in various aspects including language use and content enrichment.

Funding

The work was supported by the National Natural Science Foundation of China (11901387), the Shanghai Philosophy and Social Sciences Planning Project (2018EJB006)

References

1. Ata SK, Min W, Fang Y, *et al*. Recent advances in network-based methods for disease gene prediction. *Brief Bioinform* 2020. https://doi.org/10.1093/bib/bbaa303.
2. Liu C, Tu Y, Liao S, *et al*. Genome-wide association study of flowering time reveals complex genetic heterogeneity and epistatic interactions in rice. *Gene* 2020.
3. Andreassen OA, Djurovic S, Thompson WK, *et al*. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet* 2013; **92**(2): 197–209.
4. Witoelar A, Jansen IE, Wang Y, *et al*. Genome-wide pleiotropy between parkinson disease and autoimmune diseases. *JAMA Neurol* 2017; **74**(7): 780–792.
5. Weidinger, S, Willis-Owen SAG, Kamatani Y, *et al*. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum Mol Genet* 2013; **22**(23): 4841–4856.
6. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008; **9**(1): 30–50.
7. Cherlin S, Howey RAJ, Cordell HJ. Using penalized regression to predict phenotype from snp data. *BMC Proceedings* 2018; **12**(9): 223–8.
8. Kang J, Rancati T, Lee S, *et al*. Machine learning and radiogenomics: lessons learned and future directions. *Front Oncol* 2018; **8**:228.
9. Piette ER. Jason H Moore. Improving machine learning reproducibility in genetic association studies with proportional instance cross validation (picv). *BioData Mining* 2018; **11**(1): 6.
10. Wang Y, Liu X, Robbins K, *et al*. Antepiseeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Res Notes* 2010; **3**(1): 117.
11. Wan X, Yang C, Yang Q, *et al*. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* 2010; **26**(1): 30–7.
12. Yang C, He Z, Wan X, *et al*. Snpharvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* 2009; **25**(4): 504–11.
13. Jünger D, Hundt C, Domínguez JG, *et al*. Speed and accuracy improvement of higher-order epistasis detection on cuda-enabled gpus. *Cluster Comput* 2017; **20**(3): 1899–908.
14. Tuo S. Fdhe-iw: a fast approach for detecting high-order epistasis in genome-wide case-control studies. *Genes* 2018; **9**(9): 435.
15. Hauser A, Bühlmann P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *J Mach Learn Res* 2012; **13**(1): 2409–64.
16. Nandy P, Hauser A, Maathuis MH, et al. High-dimensional consistency in score-based and hybrid structure learning. *Ann Stat*, **46**(6A): 3151–3183, 2018.
17. Neapolitan RE et al. *Learning Bayesian Networks*, vol. **38**. Upper Saddle River, NJ: Pearson Prentice Hall, 2004.
18. Dash D, Druzdzel MJ. A hybrid anytime algorithm for the constructiion of causal models from sparse data. *arXiv preprint*, 2013.
19. Claassen T, Mooij J, Heskes T. Learning sparse causal models is not np-hard. arXiv preprint, arXiv:1309.6824, 2013.
20. Kalisch M, Mächler M, Colombo D, *et al*. Causal inference using graphical models with the R package pcalg. *J Stat Softw* 2012; **47**(11): 1–26.
21. Chickering DM. Optimal structure identification with greedy search. *J Mach Learn Res* 2002; **3**(Nov): 507–54.
22. Chickering DM. Learning equivalence classes of Bayesian-network structures. *J Mach Learn Res* 2002; **2**(Feb): 445–98.
23. Schwarz G, *et al*. Estimating the dimension of a model. *Ann Stat* 1978; **6**(2): 461–4.
24. Kalisch M, Hauser A. MH Maathuis, Martin M An overview of the pcalg package for r.2020.
25. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. *Adaptive Computation and Machine Learning* 2000.
26. Andersson SA, Madigan D, Michael D. Perlman, et al. A characterization of markov equivalence classes for acyclic digraphs. *Ann Stat* 1997; **25**(2): 505–41.
27. Colombo D, Maathuis MH, Kalisch M, *et al*. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann Stat* 2012;294–321.
28. Spirtes P, Meek C, Richardson T. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, Causation, and Discovery* 1999; **21**:211–252.
29. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005; **37**(4): 413–7.
30. Li J, Chen Y. Generating samples for association studies based on hapmap data. *Bioinformatics* 2008; **9**(1): 44.
31. Han B, Park M, Chen X. A markov blanket-based method for detecting causal snps in gwas. *Bioinformatics* 2010; **11**(3): 1–8.
32. Waldmann P, Mészáros G, Gredler B, *et al*. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front Genet* 2013; **4**:270.

33. Wan X, Yang C, Yang Q, *et al*. Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010; **87**(3): 325–40.
34. Cho S, Kim H, Sohee O, *et al*. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proceedings* 2009; **3**(7): 1–6.
35. McKinney BA, Reif DM, Ritchie MD, *et al*. Machine learning for detecting gene-gene interactions. *Appl Bioinform* 2006; **5**(2): 77–88.
36. Szymczak S, Biernacka JM, Cordell HJ, *et al*. Machine learning in genome-wide association studies. *Genet Epidemiol* 2009; **33**(S1): S51–7.
37. Upstill-Goddard R, Eccles D, Fliege J, *et al*. Machine learning approaches for the discovery of gene–gene interactions in disease data. *Brief Bioinform* 2013; **14**(2): 251–60.
38. Han B, Chen X-w. bneat: a bayesian network method for detecting epistatic interactions in genome-wide association studies. *BMC Genomics* 2011; **12**(2): 1–8.
39. Gogarten SM, Bhangale T, Conomos MP, *et al*. Gwastools: an r/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 2012; **28**(24): 3329–3331.
40. Rocha EM, De Miranda B, Sanders LH. Alpha-synuclein: pathology, mitochondrial dysfunction and neuroinflammation in parkinson's disease. *Neurobiol Dis* 2018; **109**: 249–57.
41. Patel M, Thomson NC. (r)-salbutamol in the treatment of asthma and chronic obstructive airways disease. *Exp Opin Pharmacother* 2011; **12**(7): 1133–41.
42. Mittal S, Bjørnevik K, Im DS, *et al*. β2-adrenoreceptor is a regulator of the α-synuclein gene driving risk of parkinson's disease. *Science* 2017; **357**(6354): 891–898.
43. Xingnan Li, Timothy D, Howard, Siqun L Zheng, *et al*. Genome-wide association study of asthma identifies rad50-il13 and hla-dr/dq regions. *J Allergy Clin Immunol* 2010; **125**(2): 328–335.
44. Han Y, Jia Q, Jahani SP, *et al*. Large-scale genetic analysis identifies 66 novel loci for asthma. *bioRxiv*, 749598, 2019; 749598. doi: 10.1101/749598.
45. Booms A, Pierce SE, Coetzee GA. Parkinsons disease genetic risk evaluation in microglia highlights autophagy and lysosomal genes. bioRxiv, 2020; 2020.08.17.254276. doi: 10.1101/2020.08.17.254276.
46. Fung H-C, Scholz S, Matarin M, *et al*. Genome-wide genotyping in parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2006; **5**(11): 911–6.
47. Mata IF, Wedemeyer WJ, Farrer MJ, *et al*. Lrrk2 in parkinson's disease: protein domains and functional insights. *Trend Neurosci* 2006; **29**(5): 286–93.
48. Webb J, Willette AA. Aging modifies the effect of gch1 rs11158026 on dat uptake and parkinson's disease clinical severity. *Neurobiol Aging* 2017; **50**:39–46.
49. Song GG, Lee YH. Pathway analysis of genome-wide association studies for parkinson's disease. *Mol Biol Rep* 2013; **40**(3): 2599–607.
50. Smeland OB, Shadrin A, Bahrami S, *et al*. Genome-wide association analysis of parkinson's disease and schizophrenia reveals shared genetic architecture and identifies novel risk loci. *Biol Psychiatry*, 2021; **89**(3): 227–235.
51. Ramsey J, Glymour M, Sanchez-Romero R, *et al*. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int J Data Sci Anal* 2017; **3**(2): 121–9.