RESEARCH Open Access

Privacy-preserving decision tree for epistasis detection

Qingfeng Chen[1,2*], Xu Zhang[1] and Ruchang Zhang[1]

Abstract

The interaction between gene loci, namely epistasis, is a widespread biological genetic phenomenon. In genome-wide association studies(GWAS), epistasis detection of complex diseases is a major challenge. Although many approaches using statistics, machine learning, and information entropy were proposed for epistasis detection, the privacy preserving for single nucleotide polymorphism(SNP) data has been largely ignored. Thus, this paper proposes a novel two-stage approach. A fusion strategy assists in combining and sorting the SNPs importance scores obtained by the relief and mutual information, thereby obtaining a candidate set of SNPs. This avoids missing some SNPs with strong interaction. Furthermore, differentially private decision tree is applied to search for SNPs. This achieves the efficient epistasis detection of complex diseases on the basis of privacy preserving compared with heuristic methods. The recognition rate on simulation data set is more than 90%. Also, several susceptible loci including rs380390 and rs1329428 are found in the real data set for Age-related Macular Degeneration (AMD). This demonstrates that our method is promising in epistasis detection.

Keywords: Epistasis, Relief, Mutual information, Decision tree, Differential privacy

Introduction

The search for genetic markers significantly associated with diseases within the genome-wide has become a hot topic of life science in recent years (Chen et al. 2017). Researchers usually understand the pathogenesis of disease through epistasis detection of complex diseases, and thus make decision for prevention, diagnosis and treatment. Many studies on GWAS show that common human diseases (also known as complex diseases such as hypertension, diabetes, rheumatoid arthritis, etc.) are mostly caused by gene-gene interaction and gene-environment interaction. The former is called epistasis (Guo et al. 2011), a widespread biological genetic phenomenon. It is essential to explore the epistasis detection by the interaction between SNP loci. Unfortunately, a number of studies have shown that it possible for attackers to breach genetic privacy based on SNP data (Nils et al. 2008; Yaniv and Arvind 2014; Naveed et al. 2015). Therefore, it is important to take data privacy into account when analyzing genetic data.

The application of advanced experiment techniques leads to a rapid growth of gene data. At present, there have been considerable efforts for epistasis detection, generally including four categories: *statistics*, *machine learning*, *information entropy and two-stage*. They are successful in exploiting the epistasis of weak marginal effects, and improving the efficiency of epistasis detection. However, they often overlook the potential privacy issue while analyzing the SNP data. Since the SNP data for epistasis detection are based on the genome-wide case-control data sets. They contain much individual sensitive information, such as skin color, health status. If the personal data are improperly used, it may lead to the privacy disclosure. Therefore, privacy preserving has become a critical issue in epistasis detection of complex diseases. The differential privacy method was first proposed by Dwork et al. (2006) in the cryptography community. It is based on random algorithm to perturb the query output. Compared with the early anonymization preserving technology, differential privacy defines a very strict attack model. In general, it has a strict theoretical basis, a high degree of privacy preserving in the case of less noise, and a low

*Correspondence: qingfeng@gxu.edu.cn
[1] School of Computer Electronics and Information, Guangxi University, Nanning, People's Republic of China
[2] Department of Computer Science and Information Technology, La Trobe University, Melbourne, Victoria 3086, Australia

risk of privacy leakage (Dwork 2011). Recently, several privacy-preserving methods based on differential privacy have been applied to real GWAS data (Uhlerop et al. 2012; Johnson and Shmatikov 2013; Yu et al. 2014; Simmons and Berger 2016; Simmons et al. 2016).

A novel two-stage approach is proposed in this paper by combing differential privacy technology with the decision tree algorithm for epistasis detection of complex disease. Two different strategies are used to score the importance of SNPs (Chen et al. 2017). The derived scores are then fused and sorted to extract the SNPs candidate set. The fusion process guarantees SNP loci with weak marginal effects but strong interaction are reserved. In contrast to traditional two-stage method, the decision tree does not need to find suboptimal solutions like greedy algorithms, which is easy to fall into local optimum (Chen et al. 2016). Unlike heuristic search algorithm (Wang et al. 2010), the decision tree generated by the candidate set of SNPs is applied to search for pathogenic SNPs. In particular, the differential privacy is considered while constructing the decision tree by adding the noise of the laplace distribution to the sample count at the non-leaf node and the class count at the leaf node. The experimental results demonstrate our method is prospective for epistasis detection with privacy preserving.

The remaining of the paper is organized as follows. "Related work" section is a brief introduction of previous work and our contributions. "Preliminaries" section offers the theoretical basis of decision trees and differential privacy. The detailed explanation of the research methods is described in "Epistasis detection by decision tree" section. "Experimental results" section shows the data sources and a detailed analysis of the experimental results. "Conclusion" section is a summary of the paper.

Related work

In recent years, researchers have studied epistasis detection mainly in four ways. Statistical method calculates the pathogenic SNP loci based on the statistical characteristics of disease data, and is appropriate for small-scale data. They include logistic regression (Marchini et al. 2005), multi-factor dimensionality reduction (MDR) (Ritchie et al. 2003), SNPRuler (Wan et al. 2010). Machine learning method views the epistasis detection as a feature selection problem, and selects the SNP set with the strongest correlation as the final results (Chen et al. 2016). For example, genetic programming optimization neural network algorithm (GPNN) (Motsinger-Reif et al. 2008) and TEAM algorithm (Zhang et al. 2010). Nevertheless, the results are hard to interpret. Information entropy method aims to describe the relationship between SNP combination and disease. The SNPs that are significantly associated with disease are found by amplifying the frequency differences between the SNP combinations in the case and control data sets. For example, ESNP2 by Dong et al. (2008), the contingency table of phenotype and genotype by Yee et al. (2013), and the two-order epistasis by Anunciacao et al. (2013). Although these methods promote the study of epistasis detection to some extent, the large computational overhead owing to massive genomic data makes the model too complicated to realize in many cases. Thus, the two-stage method was developed. Those irrelevant and redundant features are eliminated by screening out the important SNP loci, by which to determine a significantly correlated SNP combination. Representatives of such method are SNPHarvester (Yang et al. 2009), AntEpiSeeker (Wang et al. 2010), and BOOST (Wan et al. 2010).

A number of studies have been applied to the privacy and security of SNP data (Johnson and Shmatikov 2013; Simmons and Berger 2016; Uhlerop et al. 2012). Johnson and Shmatikov (2013) proved that differential privacy is a suitable basis for privacy-preserving query mechanisms in GWAS. Simmons and Berger (2016) proposed a convex analysis algorithm satisfying differential privacy to realize SNP data privacy protection. Uhler et al. (2012) deepened the application of differential privacy for SNPs data in GWAS. These technical methods guaranteed genetic privacy in common ways. However, their privacy research on SNP data is based on data publishing, instead of data mining. And they mainly discuss the effects of independent SNP loci on the disease. The effects of interactions between multiple SNP loci on complex diseases are not fully considered. Fei Yu et al. (2014) using penalized logistic regression with elastic-net regularization satisfying differential privacy to identify disease-causing gene combination. This method filters out a large number of SNP loci with weak main effect, so that the recognition rate of epistasis is not high. Moreover, it is only applied to the multiplicative model.

Recently, there are also many cross-researched algorithms in machine learning with differential privacy techniques, such as logistic regression, SVM, Bayesian, and decision trees, such as SuLQ-based ID3 (Blum et al. 2005), DiffP-C4.5 method (Friedman and Schuster 2010), and DiffGen method (Mohammed et al. 2011). These methods mainly consider how to select the splitting attribute of each node of the decision tree, and are similar to ID3 method in the construction of classifier. SuLQ-based ID3 method suffers from a large number of attributes and needs to divide the privacy budget into several parts, and then calculate the information gain value of each attribute. This results in extra privacy budget. However, DiffP-C4.5 uses the exponential mechanism to select splitting attributes for the disadvantages of SuLQ-based ID3 and effectively reduces noise. The disadvantage of the DiffP-C4.5 algorithm is that in each iteration, the exponential mechanism has to be called twice.

Unlike SuLQ-based ID3 and DiffP-C4.5 algorithms, DiffGen uses information gain and max operator as scoring function to construct decision tree from top to bottom, and finally adds Laplace noise to the calculation value of released leaf nodes. But, for each recursion of DiffGen algorithm, it is necessary to assign a certain privacy budget to the continuous attribute. It uses the exponential mechanism to select a subdivision scheme from the continuous attribute, and then invokes the exponential mechanism together with the discrete attribute. Zhu et al. (2013) improved DiffGen algorithm. In each iteration of subdivision, all continuous attribute subdivision schemes were multiplied by corresponding weights and then combined with discrete attribute subdivision schemes to form a candidate scheme set. This algorithm reduces the number of calls to the exponential mechanism, thus increasing the utilization rate of privacy budget and improving the accuracy of classification model.

Our method is partially adapted from DiffGen algorithm, but there is no continuous attribute allocation privacy budget. Furthermore, our algorithm applies the data mining process and does not need to generalize the original data. Therefore, data privacy preserving can achieve better results in epistasis detection. There are three main contributions in this paper:

- A fusion strategy is designed to select the features of SNP data. It avoids removing SNP loci with weak main effect but strong interaction.
- Applying decision tree and differential privacy to identify association between SNP loci and disease to achieve privacy preserving of SNP data.
- Using decision tree to search for pathogenic SNP loci to achieve epistasis detection with low time consumption.

Preliminaries

In this section, the background knowledge of decision trees and differential privacy preserving are presented. The main symbols are used in this paper and their interpretations are shown in Table 1 below.

Decision tree

Decision tree is a tree structure consisting of root node, internal node (decision node), and leaf node. Each of its non-leaf nodes represents an attribute. Each branch indidates the output of the attribute over a range of values, and each leaf node stores a category. The decision tree is constructed to obtain a classification model. Furthermore, the model is validated by using the test data and then pruned until the desired classification accuracy is reached.

The decision tree is generated based on splitting attribute nodes. The commonly used splitting criteria includes *Information Gain* and *Gini Index*. They select the splitting attribute by making each splitting subset as "pure" as possible, so that a splitting subgroup is classified into the same category.

Table 1 Main symbols and their meanings

Symbols	Meanings
D	Data set
A	Attribute set
Δq	Sensitivity of the scoring function q
ε	The privacy budget
H	Information entropy
I	Mutual information
W	Attribute weights
Score	The score of SNP loci
Power	The recognition rate of epistasis detection

Differential privacy

Differential privacy is an emerging privacy preserving technology (Dwork 2006). Unlike traditional privacy preserving relying on anonymized concealment processing of raw data, differential privacy uses a random algorithm to interfere with the query output. It is a privacy preserving model with strict mathematical proof that has been widely used (Dwork 2006).

Definition 1 (ε-Differential Privacy) *A random algorithm M satisfies ε-differential privacy. If there is only one different record between the datasets D and D'(called neighboring dataset). And for all testable sets $S \in Range(M)$, we have:*

$$\Pr(M(D) \in S) \leq \exp(\varepsilon) \Pr\left(M\left(D'\right) \in S\right) \tag{1}$$

where ε is the privacy budget.

In Definition 1, ε is used to control the probability ratio of algorithm M to obtain the same output on two neighborhood datasets. It reflects the level of privacy preserving that M can provide. The closer ε is to 0, the higher the privacy is, but the lower the data availability. Obviously, in terms of privacy preserving, we hope to set ε as small as possible. Unfortunately, this is at the expense of useful information for the data. Thus, the selection of a suitable ε is important (Zhu et al. 2017).

According to Definition 1, we know that differential privacy implements privacy preserving by adding an appropriate amount of perturbation noise to the return value of the query function. However, the sensitivity of the algorithm is a key parameter to determine the noise size. It expresses the maximal possible change in its value due to the addition or removal of a single record (Nissim and Raskhodnikova 2007).

Definition 2 (Sensitivity) *Given a function $f : D \rightarrow R$ in any neighboring datasets D and D', its sensitivity can be defined as:*

$$\Delta f = \max_{D,D'} \left\| f(D) - f\left(D'\right) \right\|_1 \tag{2}$$

Suppose the sensitivity of the function f on the dataset D is known. We only need to add the noise obeying the Laplace distribution to the calculation result of the function f. The Laplace mechanism (Dwork et al. 2006) that satisfies the differential privacy preserving is achieved. The probability density function of the Laplace distribution is:

$$P(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right) \tag{3}$$

Definition 3 (Laplace Mechanism) *Given a function $f : D \rightarrow R$, the mechanism F provides the ε-differential privacy if the following equation is true:*

$$F(D) = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right) \tag{4}$$

From the above observation, the Laplace mechanism is only suitable for numerical data. However, most of the data in real life is stored in non-numeric form. Thus, researchers have proposed an exponential mechanism (Mcsherry and Talwar 2007) for differential privacy preserving.

Definition 4 (Exponential Mechanism) *Let $q(D, \psi)$ be a scoring function of dataset D. The exponential mechanism F is ε-differential privacy if*

$$F(D) = \left\{ \psi : | \Pr(\psi \in \Psi) \propto \exp(\frac{\varepsilon q(D, \psi)}{2\Delta q)} \right\} \tag{5}$$

where Δq is the sensitivity of the function q.

Epistasis detection by decision tree

In this paper, decision tree is used to search for pathogenic SNP loci. Differential privacy and decision tree are combined to realize the privacy preserving of SNP data in the process of epistatsis detection. This section consists of dimensionality reduction, selection of a few important features (SNP loci), and epistasis detection in combination with differential privacy preserving.

Candidate feature selection by fusion strategy

Feature selection is prevalent in two-stage method to remove redundant and unrelated features. However, the previous filtering criteria are based on a single main effect. Some features (SNP loci) that have weak main effect but strong interaction might be pruned. Thus, the relief and mutual information are applied to score and sort the SNP loci, respectively. It tries to reserve the features with weak main effect but obvious interaction effect as much as possible. The importance scores of SNPs are merged to generate the candidate set of SNPs.

Relief algorithm (Kira and Rendell 1992) was first proposed by Kira et al. The features are assigned different weights $W = \{w_1, w_2, ..., w_n\}$ according to the correlation between the corresponding features and categories. A threshold δ can be specified by the data characteristics. If $\delta < w_k$, the feature is removed. The correlation is based on the ability of features to distinguish between the nearest distance samples. A random sample R is chosen from the training set *D-train*. It is used to find the nearest neighbor sample NH(called Near Hit) of the same class as R and the nearest neighbor sample NM(called Near Miss) of a different class from R. The weight of each feature is then updated according to the Eqs. 6 and 7. Note, only the discrete features are considered here.

$$w(k)_{i+1} = w(k)_i - \frac{diff(k, R, NH)}{m} + \frac{diff(k, R, NM)}{m} \tag{6}$$

$$\mathrm{diff}(k, R_1, R_2) = \begin{cases} 0 \ \ if \ \ R_1(k) = R_2(k) \\ 1 \ \ if \ \ R_1(k) \neq R_2(k) \end{cases} \tag{7}$$

If the distance between R and NH is less than the distance between R and NM on the *k-th* feature, it indicates that the feature is useful to distinguish the nearest neighborhood samples of different categories. Thus, a higher weight should be assigned to the feature. The above process is repeated m times, and eventually the average weight of each feature can be obtained. The greater the weight is, the stronger the classification ability of the feature.

Mutual information (Li et al. 2013) is used to measure the degree of association between two variables by scoring the correlation between random variables SNPs and disease. Mutual information is defined as:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \tag{8}$$

where $H(X)$ is the information entropy of X, and $H(X, Y)$ is the joint entropy of random variables X and Y. X and Y represent different SNP locus and a disease state(i.e., *control or case*), respectively.

Let $X = \{x_1, x_2, ..., x_n\}$. $p(X = x_i)$ denote the frequency of x_i appearing in X. Thus, $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$ is to measure the degree of uncertainty of the random variable X. $p(x_i)$ indicates the distribution frequency of different alleles at a SNP locus. $I(X, Y)$ is the degree of association between the SNP locus X and the disease state Y. The larger the value, the higher the degree of association between X and Y.

Two feature importance scores W and $I(I = I(X, Y))$ are obtained by the aforementioned relief algorithm and mutual information. The initial scores are normalized to

W' and I'. They are summed by weights to obtain the final merged feature ranking score. The fusion is defined as

$$Score = p_1 \cdot W' + p_2 \cdot I' \tag{9}$$

where p_1 and p_2 are the weights of the two methods, and *Score* represents the final feature ranking score. The SNPs candidate set is decided by *Score*.

Decision tree based on differential privacy

The decision tree is constructed by selecting the combination of features (SNPs) from the derived candidate feature set. However, the counting information of the SNP data may lead to a risk of personal privacy breach. The differential privacy is thus merged into the construction of decision tree as below:

- Add noise obeying the laplace distribution to the sample count of the data set;
- Use the exponential mechanism to select splitting attributes from the attribute set;
- Add the noise of the laplace distribution to the sample count of the split node. If the node satisfies the splitting termination condition, the noise is added to the sample count of the leaf node in the same way, and the class with the largest leaf node class count is retured. Otherwise go to the second step.

Algorithm 1 offers the pseudo-code of the decision tree algorithm based on differential privacy. $D(i)$ and D_c represent samples at non-leaf nodes and leaf nodes, respectively. *STC* is the splitting termination condition, as follows:

- The classification attribute of all records of the node are consistent;
- Or reaches the depth h of the decision tree;
- Or the allocated privacy budget ε is exhausted.

Algorithm 1 Decision Tree Based on Differential Privacy.

```
Require: Data set $D$, attribute set $A = \{a_1, a_2, ..., a_t\}$,
    privacy budget $\varepsilon$, decision tree depth $h$;
Ensure: Decision tree that satisfies differential
    privacy:DP-DTree;
$\varepsilon' = \varepsilon/2h$;
repeat
    $N_{(D(i))} = |D(i)| + lap\left(\Delta f/\varepsilon'\right)$;
    if the attribute node satisfies $STC$ then
        $N_c = |D_c| + lap\left(\Delta f/\varepsilon'\right)$. Return $c = argmax(N_c)$;
    else
        Using the exponential mechanism to pick $a$ by $p(a)$;
            $p(a) = \frac{\exp\left(\frac{\varepsilon' q(D_i,a)}{2\Delta q}\right)}{\sum_{a \in A} \exp\left(\frac{\varepsilon' q(D_i,a)}{2\Delta q}\right)}$
        Splitting $a$ into two child nodes, $A = A - a$;
    end if
    $h = h - 1$;
until $h = 0$ or $A = \{\}$;
return DP-DTree.
```

Pruning decision tree

To classify the training samples as accurately as possible in decision tree, some features unique to the training set are considered as general attributes of the data set, thereby over-fitting. In addition, it is no longer possible to identify a leaf with pure class values due to the application of extra noise in this paper. The splitting attribute will continue to split until the instances are insufficient and the depth constraint is not reached. It is thus important to trim the decision tree.

Some methods implement pruning by a validation set (mutually exclusive with the training set and the test data), such as the minimal cost complexity pruning and reduced error pruning. However, the validation set reduces the size of the training set. This would increase the size of relative noise in this paper. Therefore, the following formula 10 is used to trim the tree.

$$H(D_i) \geq \sum_{v \in a} \frac{|D_{i,a_v}|}{|D_i|} H\left(D_{i,a_v}\right) \tag{10}$$

where H is the information entropy and D_{i,a_v} is the leaf node. Information entropy calculates the average purity of all leaf nodes. They are compared to their parent nodes. If the above formula is satisfied, all leaf nodes of D_i is deleted and D_i become a new leaf node (Fletcher and Islam 2015).

Scoring Function

The scoring function of the exponential mechanism is also the splitting criterion of the decision tree. It directly determines the quality of the splitting attribute selection. In this paper, *Information Gain* and *Max operator* is chosen as the scoring function. $d = |D|$ is the number of records in the data set, r_a and r_C represent the values of attribute a and class C, respectively. $D_j^a = \{r \in D : r_a = j\}$, $d_j^a = \left|D_j^a\right|$, $d_c = |r \in D : r_C = c|$, $d_c = |r \in D : r_C = c|$.

Information Gain. The greater the information gain, the simpler the decision tree and the higher the classification accuracy. The information entropy of the class attribute C is defined as $H_C(D) = -\sum_{c \in C} \frac{d_c}{d} log \frac{d_c}{d}$, where d_c and d are the number of records belonging to class c and the total number of records, respectively. If the sample set D is divided by using the attribute a, the obtained information gain is:

$$\mathrm{InfoGain}\,(D,a) = H_C\,(D) - H_{C|a}\,(D) \tag{11}$$

where $H_{C|a}\,(D) = \sum_{j\in a} \frac{d_j^a}{d} \cdot H_C\left(D_j^a\right)$ is the weighted sum of the information entropy of all subsets. Since the maximum of $H_C(D)$ is $log|C|$ and the minimum of $H_{C|a}(D)$ is 0, the sensitivity Δq of $q(D,a) = \mathrm{InfoGain}\,(D,a)$ is equal to $log|C|$. Due to $C = \{control, case\}$ in the SNP data, so $|C| = 2$, and $\Delta q = log2 = 1$.

Max Operator. Max operator (Breiman et al. 1984) is used to select the class with the highest frequency as the score value of the corresponding node:

$$\mathrm{Max}\,(D,a) = \sum_{j\in a}\left(\max_c\left(d_{j,c}^a\right)\right) \tag{12}$$

According to the formula 12, the sensitivity Δq of $q(D,a) = \mathrm{Max}\,(D,a)$ is equal to 1.

Privacy analysis

We apply two composite properties of privacy budget:the sequential and the parallel composition (Mcsherry and Talwar 2007) to analyze privacy. The two lemmas are as follows:

Lemma 1 (Sequential Composition) *Suppose each G_i provide ε-differential privacy. A sequence of $G = \{G_1, G_2..., G_n\}$ over the data set D privides $(n\cdot\varepsilon)$-differential privacy.*

Lemma 2 (Parallel Composition) *Suppose each G_i provide ε_i-differential privacy. The parallel of $G = \{G_1, G_2..., G_n\}$ over a set of disjoint data sets D_i will provides $max\{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$-differential privacy.*

Each layer of the decision tree is the same data set. According to the Lemma 1, the privacy budget assigned to each layer is $E = \varepsilon/h$. The splitting of nodes at each level is on disjoint data sets. According to Lemma 2, each node is assigned a privacy budget that is less than or equal to this layer's privacy budget. Here, we assume that the privacy budget of each node is equal to the privacy budget of this layer. Then half of the privacy budget assigned to each node, $\varepsilon' = E/2 = \varepsilon/2h$, is used to estimate the instance count of the node (adding Laplacian noise), and the other half of the privacy budget$\left(\varepsilon' = \varepsilon/2h\right)$ is used by the exponential mechanism to select the optimal splitting node or added Laplacian noise to the leaf node instance count. Consequently, the total privacy budget consumed by the algorithm is not greater than $h * (\varepsilon/2h + \varepsilon/2h) = \varepsilon$. It satisfies ε-differential privacy.

Time complexity analysis

In order to generate a decision tree, we need to scan the entire data set D. Then use the exponential mechanism to select an attribute to split, the time complexity of this process is $O(t|D|log|D|)$ (t is the number of attribute set). After the exponential mechanism selects the splitting attribute, the data set needs to be divided once. In the worst case, the entire data set needs to be scanned, and the time complexity is $O(|D|)$. Since the decision tree depth is h, the time complexity of the algorithm is $O(h|D|log|D|)$ under a certain number of attributes.

Experimental results

In this section, the generation process of the simulation data sets, the source and preprocessing of real data are introduced. In addition, the performance evaluation and parameter configuration of our algorithm are explained. Finally, we analyzed and summarized the experimental results.

Experiment data

Two kinds of data are applied to evaluate the performance of our method, including the simulation data sets and a real disease data.

Simulation data

The simulation data sets are generated by three common disease models, namely *additive model*, *multiplicative model* and *threshold model* as shown in Table 2. The disease models contains both the marginal effect (main effect) of the single locus and the interaction of multiple loci. Each model generates 100 data sets according to different parameters. Each data set contains 2000 samples, of which includes 1000 case samples and 1000 control samples. Each sample has 1000 SNP loci, including 2 pathogenic SNP loci (SNP11 and SNP21) and 998 nonpathogenic SNP loci. The values of the *gene effect* θ and

Table 2 Three common disease models

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

the *baseline effect* α can be calculated by the minor allelic frequency *MAF*, linkage disequilibrium (LD) r^2, main effect λ and disease penetrance $p(D)$. The two values are to generate the corresponding simulation data set.

In this paper, the simulation data (Wang et al. 2012) (http://compbio.ddns.comp.nus.edu.sg/~wangyue/public_data/) used three disease models. We set $MAF \in \{0.2, 0.5\}$, $\lambda \in \{0.3, 0.5\}$, $r^2 = 1$ and $p(D) = 0.1$. Therefore, there are $3*2*2 = 12$ sets of data sets, including 020301_*i*, 020501_*i*, 050301_*i*, 050501_*i*. $i \in \{1, 2, 3\}$ represents the model 1, 2, 3.

Real data

Age-related Macular Degeneration (AMD) (Klein et al. 2005) is real disease data containing 116,204 SNP loci genotyped with 96 cases and 50 controls. This data set has proved that two SNP loci, rs380390 and rs1329428, are significantly associated with this disease(AMD).

Data preprocessing of the AMD data is performed:

- SNP loci that do not satisfy the polymorphism and Hardy-Warmbert equilibrium conditions are excluded.
- The genotypes *aa, Aa, AA* in the SNP data are coded as 0, 1, 2 (*AA* is dominant homozygote; *Aa* is heterozygote; *aa* is recessive homozygote), respectively. The case and control labels in the attribute class are coded as 0, 1.
- The missing values in the processed data: SNP loci with data missing rates above 10% in SNPs are deleted; otherwise the three alleles at the SNP loci are counted and the missing values are filled with the most count allele.

Search strategy based on decision tree

The decision tree obtained in "Decision tree based on differential privacy" section is used to search for pathogenic SNP loci. All non-leaf nodes from the root node to the *L-th* layer of the decision tree represent SNP loci that may be pathogenic, namely *SL*. Here, two scenarios are considered. Scenario A: if SNP11 and SNP21 exist in *SL* simultaneously, this disease model is a disease-causing model. Considering that both SNP11 and SNP21 may be loci with weak main effects and strong interactions, some disease models cannot simultaneously detect them. Thus, we have Scenario B: if SNP11 or SNP21 exists in *SL*, this disease model is also a disease-causing model. The experiments in "Results on simulation data" section compare the performance of our algorithm on each simulation data set.

Performance evaluation

To evaluate the performance of the proposed algorithm, the power of an epistasis detection method is evaluated based on Eq. 13.

$$Power = \#TP/\#D \tag{13}$$

Power is a measure of the capability of all data sets to detect disease-causing models, also known as the the recognition rate. Where *#TP* is the number of disease-causing models from all *#D* datasets (there are 100 data matrices for each disease model). Here, the disease model that can search for SNP11 and SNP21(labeled as a disease-causing locus) in *SL* is called a disease-causing model.

Parameter configuration

In the simulation experiments of the Figs. 1, 2, 3 and 4, the parameters is adjusted through multiple experiments to obtain the depth of the tree $h = 10$, privacy budget $\varepsilon = 0.5$. The scoring function used by the exponential mechanism in this experiments are based on information gain. At the same time, we set layer *L* of the decision tree is 2,3,4, respectively. In the simulation experiments of the Figs. 5, and 6, according to the results of the parameter adjustment, $h = 10$. The disease model 050301_2 and

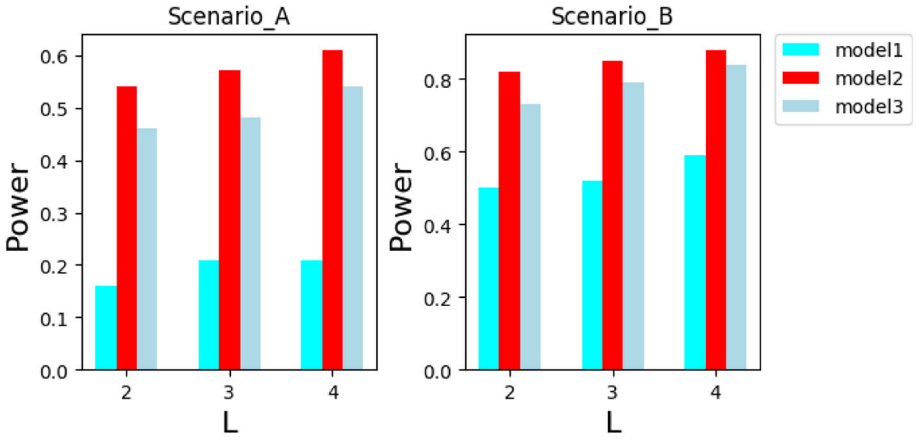

Fig. 1 020301($MAF = 0.2, \lambda = 0.3, r^2 = 1$)

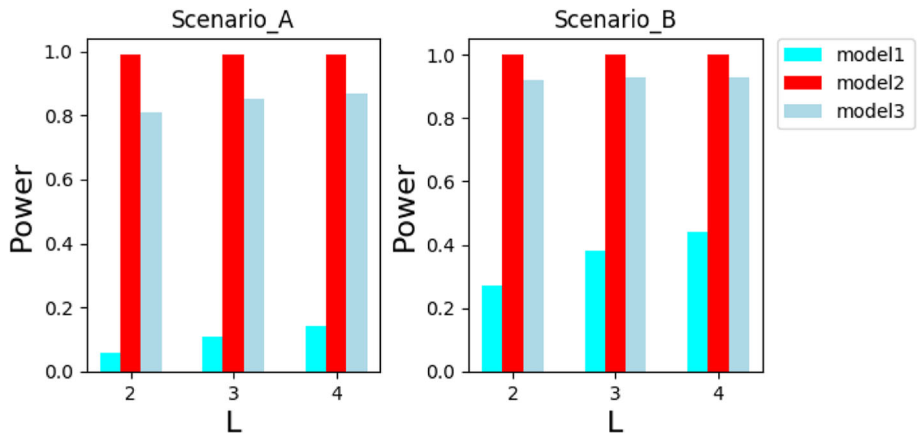

Fig. 2 020501($MAF = 0.2, \lambda = 0.5, r^2 = 1$)

020501_3 are used to verify the impact of the privacy budget on the recognition rate, and in 050301_2 and 020501_3 the level L is seted to 3,2 respectively. *Information Gain* and the *Max Operator* are applied as the scoring function of the exponential mechanism respectively. In the simulation experiment of Fig. 7, the disease model 020501_3 and 050301_2 were selected to perform experiments in Scenario A, and the level $L = 2$. The privacy budget ε considered in this experiment is 0.01,0.05,0.1,0.5 and 1, and the *information gain* as the splitting criterion of the decision tree. In the AMD data experiments, the parameters were seted as $h = 10$, $\varepsilon = 0.5$, $L = 3$.

Results on simulation data

In this section, the effects of MAF and MainEffect on the detection of pathogenic SNP loci are compared in the same disease model. With the same MAF and MainEffect, the performance of the proposed algorithm in detecting pathogenic SNP loci in different disease models is compared. And in the case of the same disease model, the impact of the privacy budget on the epistasis detection is verified.

As you can see in Figs. 1, 2, 3 and 4, our algorithm is suitable for model 3, more suitable for model 2, and unsuitable for model 1. This is because the two pathogenic SNP loci in the disease model 1 independently contribute to the disease risk, making the main effect account for a high proportion. However, when using the relief and mutual information to select the candidate sets of SNPs from different angles, the independent effect of single locus is weakened. This makes the importance scores of disease-causing SNPs in model 1 lower. Therefore, it is difficult to screen disease-causing SNP loci into the candidate sets. Model 3 is the opposite of model 1, with a

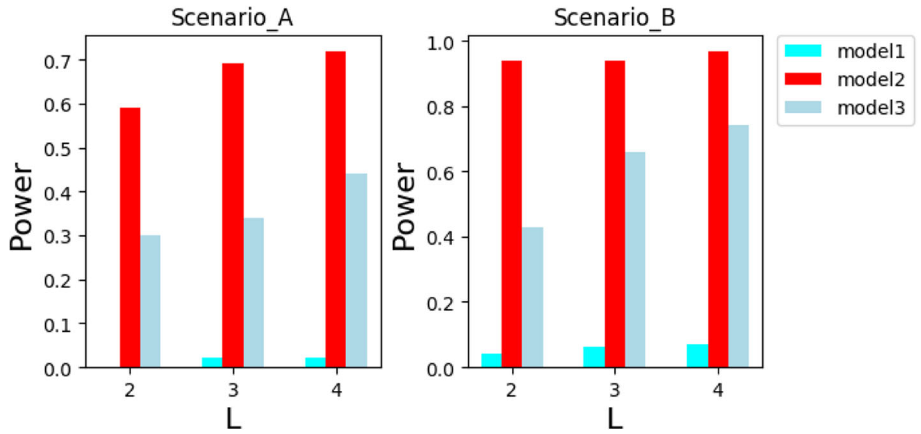

Fig. 3 050301($MAF = 0.5, \lambda = 0.3, r^2 = 1$)

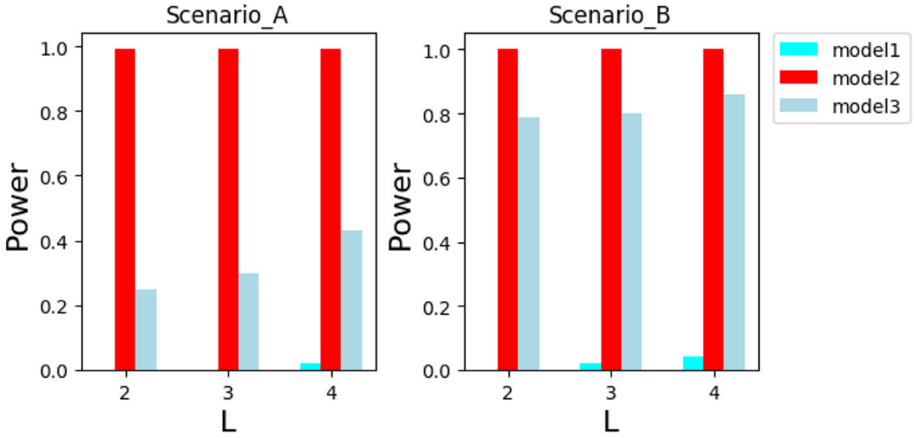

Fig. 4 050501($MAF = 0.5, \lambda = 0.5, r^2 = 1$)

larger proportion of interaction. Only the proportion of interaction to main effect in model 2 is more suitable for the screening rules of the relief and mutual information methods. From Figs. 1 and 2 or Figs. 3 and 4, we can see that the model 1 has a lower recognition rate when the main effect is increased, while the models 2 and 3 are correspondingly increased. This exactly confirms that the disease-causing SNP loci were not screened to the candidate sets of SNPs. Moreover, it can be seen from Figs. 1 and 3 or Figs. 2 and 4 that only model 2, in the case of increased MAF, the recognition rate increases, and models 1, 3 are reduced. This can be a secondary evidece for the advantages of our method in detecting pathogenic loci on model 2.

In Figs. 5 and 6, the privacy budget ε is larger and the recognition rate is higher. This is because in the interval of [0.01, 1], when the privacy budget ε is larger, the probability that the better attribute is selected as the splitting attribute is higher. When the privacy budget ε is small, the *information gain* is more suitable as a scoring function for the exponential mechanism in this experiment than *Max Operator*. However, if the privacy budget ε reaches 0.5, the recognition rate of both cases are the highest. This shows that there is a good trade-off between data privacy preserving and the performance of epistasis detection. Although the performance of data privacy preserving is better when ε is smaller, the damage degree of epistasis detection is also higher.

From the above experimental results, we can see that our method can complete the data privacy preserving when detecting the epistasis on the simulation dataset. In order to further illustrate the

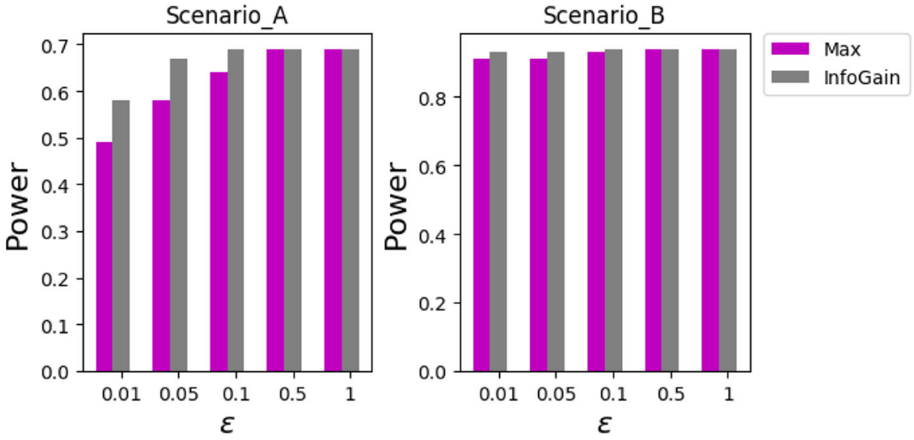

Fig. 5 050301_2($MAF = 0.5, \lambda = 0.3, r^2 = 1$),$L = 3$

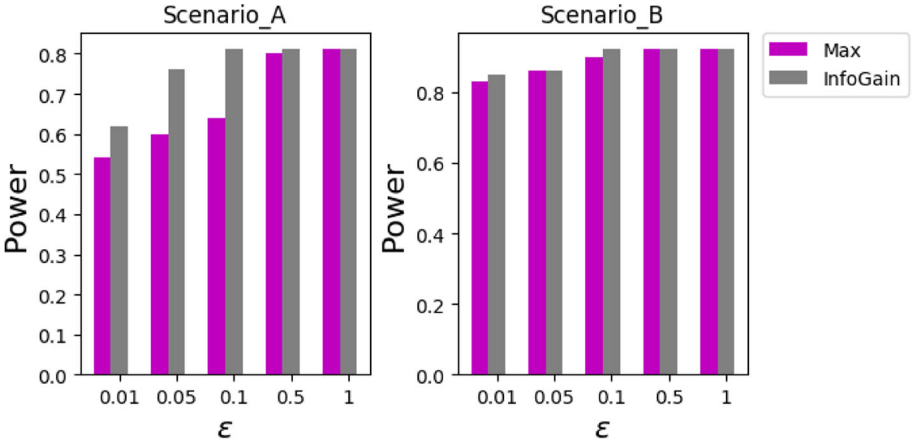

Fig. 6 020501_3($MAF = 0.2, \lambda = 0.5, r^2 = 1$),$L = 2$

effectiveness of the proposed method, we compare the non-privacy decision tree(DTree) with our method(DP-DTree). The experimental results are shown in Fig. 7 below:

As the privacy budget is varied from 0.01 to 1, the degree of data perturbation becomes smaller. According to Definition 1 data availability becomes higher and higher. At this point, in Fig. 7, Power is increasingly higher, indicating that data availability and Power are positively correlated. As a result, the data availability is increasingly stronger, and the experimental results obtained by DP-DTree are getting better. If the privacy budget ε is close to 0.5 and is increased to 1, the performance of the DP-DTree algorithm is close to DTree. This shows that our algorithm can achieve a good tradeoff between data privacy preserving and epistasis detection.

Results on AMD data

Our method is also used to analyze AMD data. Through experiments, we detected disease-causing SNP loci of the AMD data in the top 3($L = 3$) layer non-leaf nodes of the decision tree. They are shown in Fig. 8 and Table 3.

In Table 3, the top 10 pathogenic SNP loci on the real disease data AMD are detected. Among them, rs380390 and rs1329428 are in the first and second layers of the tree, respectively, and have been shown to be associated with AMD disease. In the third layer, rs10507949 (Tang et al. 2009) and rs786358 (Jiang et al. 2009) are also

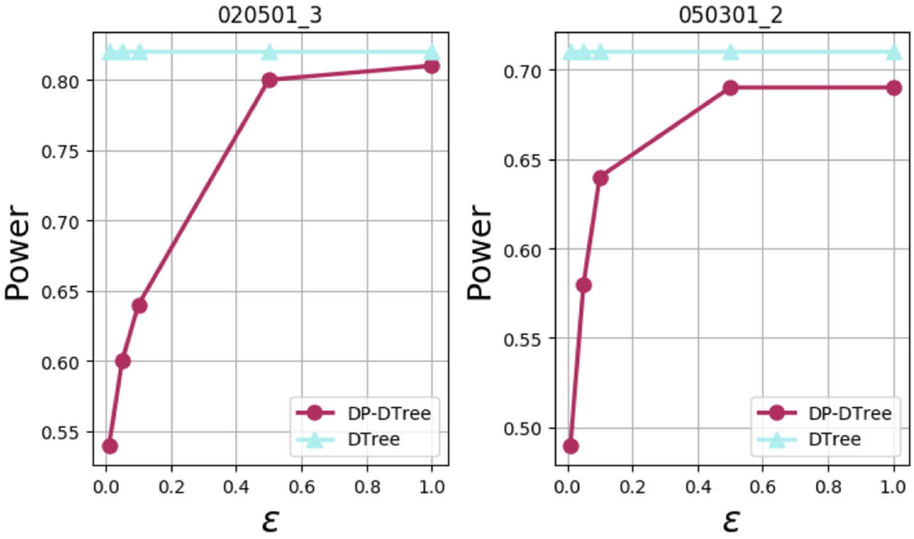

Fig. 7 Experimental results on DTree and DP-DTree(ε=0.01,0.05,0.1,0.5,1)

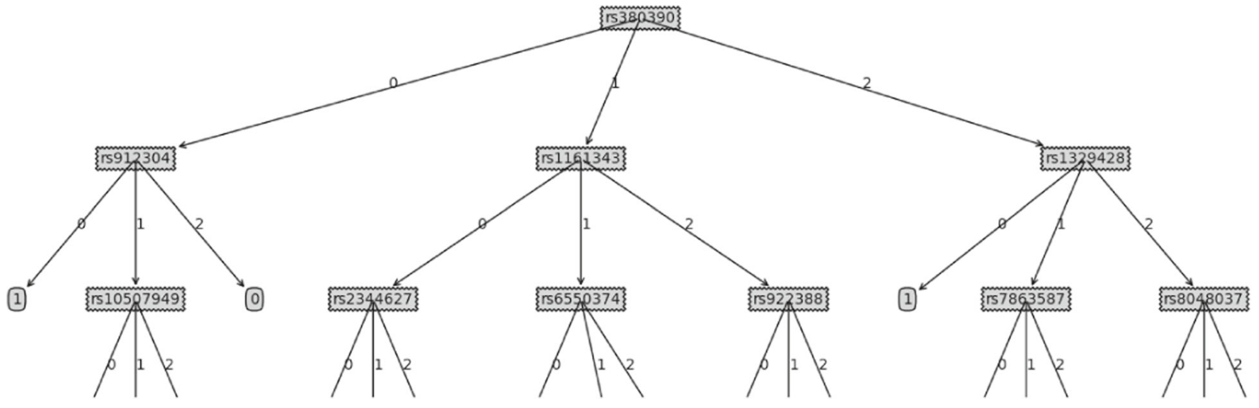

Fig. 8 Epistasis results on the dataset AMD

detected in other literatures as having a strong association with this disease. In addition, other loci are detected in this paper are rs912304, rs1161343. Although they have not been found to be associated with the disease in related works. According to the experimental results, they has a high possibility to be associated with the disease. Therefore, the method of this paper is also effective in real data.

Conclusion

In this paper, we present a novel two-stage method for the epistasis detection of complex disease. A fusion strategy was proposed to select a small number of important SNP loci. SNPs with weak main effects but significant interaction effects are reserved. Furthermore, a decision tree was used to search for pathogenic SNP loci. In particular, differential privacy technology is applied in decision tree to ensure that the SNP data privacy information is not leaked during the epistasis detection. The experimental results from both simulation data and real data demonstarte our methods is able to perform epistasis detection of complex diseases with high accuracy and privacy preserving.

Table 3 Epistasis results on the dataset AMD

Rank(DTree layer)	SNP locus	Chromosome location
1 (1)	rs380390	1
2 (2)	rs912304	14
3 (2)	rs1161343	5
4 (2)	rs1329428	1
5 (3)	rs10507949	13
6 (3)	rs2344627	4
7 (3)	rs2344627	3
8 (3)	rs922388	4
9 (3)	rs7863587	9
10 (3)	rs8048037	16

Acknowledgements
We thank for the reviewers for their valuable comments.

Funding
The work reported in this paper was partially supported by two National Natural Science Foundation of China projects 61363025, 61751314, a key project of Natural Science Foundation of Guangxi 2017GXNSFDA198033 and a key research and development plan of Guangxi AB17195055.

Availability of data and materials
The dataset supporting the conclusions of this article is available at http://compbio.ddns.comp.nus.edu.sg/~wangyue/public_data/.

Authors' contributions
The authors have equally contributed. They have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 25 October 2018 Accepted: 24 January 2019
Published online: 18 February 2019

References

Anunciação O, Vinga S, Oliveira AL (2013) Using information interaction to discover epistatic effects in complex diseases. PLoS ONE 8(10):e76300
Blum A, Dwork C, Mcsherry F, Nissim K (2005) Practical privacy:the sulq framework. In: Proceedings of the Twenty-fourth ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. ACM, New York. pp 128–138
Breiman LI, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees (cart). Encycl Ecol 40(3):582–588
Chen Q, Chen YP, Zhang C (2016) Interval-based similarity for classifying conserved rna secondary structures. IEEE Intell Syst 31(3):78–85. https://doi.org/10.1109/MIS.2015.2
Chen Q, Lan C, Chen B, Wang L, Li J, Zhang C (2016) Exploring consensus rna substructural patterns using subgraph mining. IEEE/ACM Trans Comput Biol Bioinforma 14(5):1134–1146
Chen Q, Lan C, Zhao L, Wang J, Chen B, Chen YP (2017) Recent advances in sequence assembly: principles and applications. Brief Funct Genomics 16(6):361–378. https://doi.org/10.1109/MIS.2015.2
Chen Q, Wang Y, Chen B, Zhang C, Wang L, Li J (2017) Using propensity scores to predict the kinases of unannotated phosphopeptides. Knowl-Based Syst 135:60–76
Dong C, Chu X, Wang Y, Wang Y, Jin L, Shi T, Huang W, Li Y (2008) Exploration of gene-gene interaction effects using entropy-based methods. Eur J Hum Genet 16(2):229–235

Dwork C (2006) Differential privacy. Lect Notes Comput Sci 26(2):1–12

Dwork C (2011) Differential Privacy. Springer, Berlin Heidelberg

Dwork C, Mcsherry F, Nissim K (2006) Calibrating noise to sensitivity in private data analysis. In: Proceedings of the Third Conference on Theory of Cryptography. Springer-Verlag, Berlin. pp 265–284

Fletcher S, Islam MZ (2015) A Differentially Private Decision Forest. In: Ong K. L., Zhao Y., Stone M. G., Islam M. Z. (eds). Thirteenth Australasian Data Mining Conference (AusDM 2015). ACS, Sydney Vol. 168. pp 99–108

Friedman A, Schuster A (2010) Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, NewYork. pp 493–502

Guo H, Li FG, Wang ZP, Hui L (2011) Current status of snps interaction in genome-wide association study. Hereditas 33(9):901

Jiang R, Tang W, Wu X, Fu W (2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10(Suppl 1):1–12

Johnson A, Shmatikov V (2013) Privacy-preserving data exploration in genome-wide association studies. KDD Proc Int Conf Knowl Disc Data Min 2013(1):1079–1087

Kira K, Rendell LA (1992) A practical approach to feature selection. In: Proceedings of the Ninth International Workshop on Machine Learning (ML 1992). Morgan Kaufmann, San Fracisco. pp 249–256

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST (2005) Complement factor h polymorphism in age-related macular degeneration. Science 308(5720):385–389

Li X, Liao B, Cai L, Cao Z, Zhu W (2013) Informative snps selection based on two-locus and multilocus linkage disequilibrium: Criteria of max-correlation and min-redundancy. IEEE/ACM Trans Comput Biol Bioinforma 10(3):688–695

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37(4):413

Mcsherry F, Talwar K (2007) Mechanism design via differential privacy. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. IEEE Computer Society, Washington, DC. pp 94–103

Mohammed N, Chen R, Fung BCM, Yu PS (2011) Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York. pp 493–501

Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD (2008) Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol 32(4):325–340

Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, Malin BA, Wang X (2015) Privacy in the genomic era. ACM Comput Surv 48(1):1–44

Nils H, Szabolcs S, Margot R, David D, Waibhav T, Jill M, Pearson JV, Stephan DA, Nelson SF, Craig DW (2008) Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet 4(8):e1000167

Nissim K, Raskhodnikova S (2007) Smooth sensitivity and sampling in private data analysis. In: Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing. ACM, New York. pp 75–84

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24(2):150–7

Simmons S, Berger B (2016) Realizing privacy preserving genome-wide association studies. Bioinformatics 32(9):1293–1300

Simmons S, Sahinalp C, Berger B (2016) Enabling privacy-preserving gwass in heterogeneous human populations. Cell Syst 3(1):54–61

Tang W, Wu X, Jiang R, Li Y (2009) Epistatic module detection for case-control studies: a bayesian model with a gibbs sampling strategy. PLoS Genet 5(5):e1000464

Uhlerop C, Slavković A, Fienberg SE (2012) Privacy-preserving data sharing for genome-wide association studies. J Priv Confidentiality 5(1):137

Wan X, Yang C, Yang Q, Xue H, Fan X (2010) Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 87(3):325–340

Wan X, Yang C, Yang Q, Xue H, Tang NLS, Yu W (2010) Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics 26(1):30–37

Wang Y, Liu G, Feng M, Wong L (2012) Response: an empirical comparison of several recent epistatic interaction detection methods. Bioinformatics 28(1):145–146

Wang Y, Liu X, Robbins K, Rekaya R (2010) Antepiseeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. BMC Res Notes 3(1):1–8

Yang C, He Z, Wan X, Yang Q, Xue H, Yu W (2009) Snpharvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics 25(4):504

Yaniv E, Arvind N (2014) Routes for breaching and protecting genetic privacy. Nat Rev Genet 15(6):409–421

Yee J, Kwon MS, Park T, Park M (2013) A modified entropy-based approach for identifying gene-gene interactions in case-control study. PloS ONE 8(7):e69321

Yu F, Fienberg SE, Slavković AB, Uhler C (2014) Scalable privacy-preserving data sharing methodology for genome-wide association studies. J Biomed Inform 50(S1):133–141

Yu F, Rybar M, Uhler C, Fienberg SE (2014) Differentially-Private Logistic Regression for Detecting Multiple-SNP Association in GWAS Databases. In: Josep Domingo-Ferrer (ed). Privacy in Statistical Databases, PSD 2014. Springer International Publishing, Cham, Ibiza. pp 170–184

Zhang X, Huang S, Zou F, Wang W (2010) Team: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics 26(12):i217

Zhu T, Li G, Zhou W, Yu PS (2017) Differentially private data publishing and analysis: A survey. IEEE Trans Knowl Data Eng PP(99):1–1

Zhu T, Xiong P, Xiang Y, Zhou W (2013) An Effective Deferentially Private Data Releasing Algorithm for Decision Tree. In: Proceedings of the 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. IEEE Computer Society, Washington, DC. pp 388–395