A Differential Privacy Approach to Preserve GWAS Data Sharing based on A Game Theoretic Perspective

Jun Yan[1,2], Ziwei Han[1], Yihui Zhou[1*], Laifeng Lu[3]
[1] School of Computer Science, Shaanxi Normal University
Xi'an, 710119 - China
[e-mail: yanrongjunde@snnu.edu.cn, HanZiwei@snnu.edu.cn, zhouyihui@snnu.edu.cn]
[2]School of Mathematics and Computer Applications, Shangluo College,
Shangluo, 72600 - China
[3]School of mathematics and statistics, Shaanxi Normal University
Xi'an, 710119 - China
[e-mail: lulaifeng@snnu.edu.cn]
*Corresponding author: Yihui Zhou

Received January 11, 2022; accepted March 6, 2022; published March 31, 2022

Abstract

Genome-wide association studies (GWAS) aim to find the significant genetic variants for common complex disease. However, genotype data has privacy information such as disease status and identity, which make data sharing and research difficult. Differential privacy is widely used in the privacy protection of data sharing. The current differential privacy approach in GWAS pays no attention to raw data but to statistical data, and doesn't achieve equilibrium between utility and privacy, so that data sharing is hindered and it hampers the development of genomics. To share data more securely, we propose a differential privacy preserving approach of data sharing for GWAS, and achieve the equilibrium between privacy and data utility. Firstly, a reasonable disturbance interval for the genotype is calculated based on the expected utility. Secondly, based on the interval, we get the Nash equilibrium point between utility and privacy. Finally, based on the equilibrium point, the original genotype matrix is perturbed with differential privacy, and the corresponding random genotype matrix is obtained. We theoretically and experimentally show that the method satisfies expected privacy protection and utility. This method provides engineering guidance for protecting GWAS data privacy.

Keywords: Genome-wide association studies, differential privacy, Nash equilibrium, chi-squared test

A preliminary version of this paper appeared in IEEE, International Conference on Networking and Network Applications (NaNA), 2018, October 12-15, Xi'an, China. This version includes a system model, a threat model, a concrete analysis, and comparison with other method. This research was supported by the Fundamental Research Funds for the Central Universities (No.GK201704017 and No.GK201903091), and the Scientific and Technological Project of Shangluo (No. 2021-C-0004).

http://doi.org/10.3837/tiis.2022.03.015 **ISSN : 1976-7277**

1. Introduction

In the genomic era, people desire to reveal the genetic basis of diseases so that we can have genetic therapy for some diseases. By statistical comparison of large number of samples across the genome, bioinformaticians discover the effects of genetic variations on certain diseases. This type of research is called a genome-wide association study (GWAS). The genetic variation mentioned here mainly refers to single nucleotide polymorphism (SNP), which is the point mutations in one nucleotide of a genetic sequence [1]. Then geneticists perform the case-control study to explore the top-*k* most relevant SNPs to a disease [2-7]. It is undeniable that sharing these genomic data is critical to the progress of GWAS as it could result in new scientific discoveries. However, the privacy risk is a big challenge during genomic data sharing. A survey demonstrated over half of participants would not publicly share their genome on the web even for research purposes [8]. Malicious attackers can infer individual's identity, genotypes in some special loci, disease status and gene information of one's family member [10-12]. Therefore, how to share the genomic data without compromising privacy is an immediate and pressing issue.

Due to privacy concerns, to overcome the privacy bottle-neck in genome-wide association studies, five kinds of approaches are mainly used. (i) Controlled-access. Access to raw data is limited to a few trusted individuals. Only if the researchers go through a time consuming and burdensome application process, would they get raw genotype data of a GWAS. It is worth noting that the strict controlled-access has hampered the research practices [8]. The antenna systems as main part of communication systems can be useful for protect privacy in GWAS, and relevant antenna system security design can be used for reference[13-15] (ii) Encryption. Some ways of using homomorphic encryption to protect data privacy have been proposed in GWAS [16-18]. However, the untrusted third party may leak encrypted information or decrypt data privately. (iii) De-identification. The name of participants will be removed to protect privacy, but other meta data such as birthdate, hometown and gender will not be deleted because these are useful as experimental variable. Actually these meta data can be used to perform link-attack to identify the participant [8]. (iv) Federated learning. For the large-scale distributed data analysis of GWAS, the federated learning method does not directly collect the data from user terminals, instead, it collects the latest model training updates on each terminals to avoid user privacy issues, which caused by local data uploads[9]. (v) Differential privacy (DP). It is a rigorous privacy preserving method which can quantify the privacy risk by the privacy budget, ignore the background knowledge of the attacker and avoid the inference attack [19-22]. For two adjacent datasets that differ in at most one entry, differential privacy add noises to raw dataset and output a disturbed dataset which has almost the same statistical result as the raw dataset, so that the two datasets are indistinguishable to any adversary. Therefore, the adversary cannot judge whether the individual exists in the database by only the statistical value difference.

Unfortunately, most DP approaches for protecting privacy in GWAS mainly focus on output statistic data, such as allele frequency, *p*-values, not the genotype data, so the raw format genotype data privacy still can't be resolved. Moreover, the equilibrium between utility and privacy of processed data is still difficult to achieve. If the disturbance is so much, the utility will be affected a lot, and the statistical result will not be within the reasonable error range. If the disturbance is not enough, the disturbed data may be very close to the original data, so that the privacy protection is not enough, which may lead to fewer volunteers donating their genotype data. The equilibrium between utility and privacy is what we focus on. Obviously, Nash Equilibrium can help us to find the equilibrium point.

For GWAS data sharing, we propose a differential privacy preserving GWAS data sharing approach which is satisfied with Nash equilibrium. We call the approach as NEDP (Nash equilibrium and differential privacy). The NEDP approach disturbs the raw genotype data using Laplace mechanism or discrete Laplace mechanism and finally outputs disturbed genotype data with the expected utility and privacy. We compare our NEDP with the approach in Simmons's paper [29]. Simmons add noises to the independent variables of the statistical value, which we call "independent variable perturbation" (IVP) for the sake of convenience. IVP was proved to achieve good data utility. In the comparison between IVP and NEDP, the utility of two approaches are very close, while the privacy of our approach is far better than IVP. Our major contributions are as follows:

(1) We construct a differential privacy preserving model of genotype data sharing for GWAS, and propose NEDP approach, which uses not only Laplace mechanism but also discrete Laplace mechanism to find an equilibrium point between privacy and utility and then disturbs the raw genotype data based on the equilibrium point.

(2) We theoretically prove that the NEDP method satisfies differential privacy and achieves the equilibrium between privacy and utility. The NEDP method satisfies differential privacy, which can resist statistical inference, so the method is robust when the sharing data is subjected to such inference. The scalability and the engineering feasibility of NEDP method is also proved by experiments to be suitable for dominant model, recessive model, and multiplicative model in GWAS, provided the model can be represented as a 2X2 contingency table.

The remainder of the paper is organized as follows. In Section 2, we introduce the current privacy protection methods in GWAS. Section 3 introduces the related preliminaries of our approach. Section 4 shows our system model, threat model and the privacy preserving model. Section 5 demonstrates our NEDP approach. In Section 6, we have experimental analysis. Section 7 draws the conclusion of this paper.

2. Related Work

In Section 1, we introduce four kinds of techniques to protect privacy in GWAS. However, in this section, we introduce privacy protection angles based on the different sources of the data. The main data sources are the following three: (i) the access, storage and query of the genomic data, (ii) the statistical data such as *p*-values, the top-*k* most relevant SNPs and the MAFs and (iii) the input raw data such as genotypes or phenotypes.

The first kind of method mainly protects the access [23] , the storage [24] and the query [25] of the data. To solve the secret sharing among several independent data centers, Kamm L et al. [24] proposed a secure multi-party computation algorithm for a host center to construct case and control groups, without privacy leakage among three or more biobanks. Shringarpure and Bustamante [25] developed a likelihood-ratio test to avoid the re-identification attacks on beacons which are web servers that answer allele-presence queries.

The second kind of method is widely used for privacy preserving in GWAS, it mainly protect the statistical data. Johnson and Shmatikov [26] put forward a differential privacy distance-score mechanism to protect the output of the chi-squared test in GWAS. Uhler et al. [27] proposed a differential privacy method for the release of summary statistics in GWAS. Then Yu. et al. [28]extend the methods of Uhler et al. by proposing a new algorithm based on the exponential mechanism. Wang et al. [22] provided a software "dpTDT" to protect the kinship privacy in GWAS.

The third kind of method is only studied in cryptography, but as analyzed in Section 1, the encryption method is not beneficial to data release. Protecting the input raw data with differential privacy should be considered so that the regulatory authorities can release the data with original format for researchers to perform a targeted study. Simmons and Berger [29] proved their "input perturbation" approach to overcome accuracy issues in the output perturbation, and we will prove that our approach overcomes the privacy issues in Simmons' approach. Here we show the approach to release the original format data for GWAS with the equilibrium between personal privacy and data utility.

3. Preliminaries

In this section, we introduce the preliminaries about GWAS, differential privacy, chi-squared test and Nash Equilibrium. In **Table 1**, we give the explanations of some notations used in this paper.

Table 1. Definition of notations.

Notations	Definitions
χ^2	Chi-squared value of orginal genotype dataset
$\chi^2_{\alpha,v}$	The critical chi-squared value within the significance level α and the degree of freedom v
χ^2_{dp}	Chi-squared value of disturbed genotype dataset after using differential privacy
M	Original genotype matrix
M'	Disturbed genotype matrix
a	The original number of genotype 0 in cases
a'	The disturbed number of genotype 0 in cases
$a*$	The final equilibrium number of genotype 0 in cases
N	The total number of participants in GWAS
m	The original number of genotype 0
n	The original number of genotype 1
α	The significance level in Chi-squared test
EEE	Expected estimation error

3.1 SNP and Genotype in GWAS

SNP. A gene is a sequence of DNA or RNA which codes for a functional protein molecule. We know that there are four kind of bases in DNA molecules: adenine (A), thymine (T), cytosine (C) and guanine (G). The DNA sequence is composed of A, T, C, G. The raw GWAS data often refers to the single nucleotide polymorphisms (SNPs) data on the DNA. A single SNP is a variation in a single nucleotide that occurs at a specific position in the genome. For example, two sequenced DNA fragments from the same positions of two individuals are ATCGCAA and ATTGCAG, at the 3-th and 7-th base positions, the two differences C-T and A-G appear. We refer to those as 2 SNPs. There are two alleles for one SNP. If C nucleotide appears in most individuals while T in a minority of individuals, we call C a major allele and T a minor allele. It's worth noting that many diseases of the human body are caused by SNPs.

Genotype and Genetic Model. For one SNP with major C and minor T, the genotype of a SNP position may be CC, CT and TT. When the genetic effects of CC and CT are the same, we call that a dominant model and regard the CC+CT as one genotype. When the genetic effects of TT and CT are the same, it is in the recessive model. When the genetic effects of allele C and allele T are different, it is in the multiplicative model. These three genetic models are the main models investigated in GWAS [1]. For each genetic model, there are two genetic effects, and we code the pooled genotype (CC+CT or TT+CT) as 0 and the remaining genotype (TT or CC) as 1 [12]. Our approach applies to all three genetic models. It is generally assumed there are N participants in the GWAS and both of the number of cases and controls are equal to $N/2$. The number of genotype 0 is *m* and the number of genotype 1 is *n*.

3.2 Pearson's Chi-squared Statistic Test

Suppose we want to use chi-squared test to determine if a SNP has significant effect on a disease, according to the assumption in Section 3.1, for the SNP with major allele C and minor allele T, we get observed genotype counts for the SNP as shown in **Table 2**. Now, let's take **Table 2** as an example to illustrate the steps of the chi-squared test.

(1) Establish the test hypotheses and determine the significance level

H_0 : a null hypothesis that the SNP has no significant effect on the disease.

H_1 : a hypothesis that the SNP has significant effect on the disease.

(2) Calculate the χ^2 test statistic

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \tag{1}$$

In a $R \times C$ contingency table, each cell records the observed value denoted as $O_{ij} \in (a,b,c,d)$ where *i* is the number of rows, and *j* is the number of columns. The expected theoretical value for each cell denoted as $E_{ij} = N_i \times N_j / N$ where N_i is the total number of *i*-th row, N_j is the total number of *j*-th column and *N* is the total number of rows and columns in the table.

(3) Determine the *p*-value and draw the conclusion

For a 2×2 contingency table, the degree of freedom is $v = (R-1)\times(C-1) = (2-1)\times(2-1) = 1$. Given the significance level α and the degree of freedom v, compare the chi-squared value χ^2 and the critical chi-squared value $\chi^2_{\alpha,v}$. When $\chi^2 \geq \chi^2_{\alpha,v}$, $p \leq \alpha$, the H_0 hypothesis can be rejected, and H_1 can be accepted, so the SNP is significant for the disease. When $\chi^2 \leq \chi^2_{\alpha,v}$, $p \geq \alpha$, the H_0 hypothesis can be accepted, and H_1 can be rejected, so the SNP is not significant for the disease.

Table 2. Observed genotype counts for one SNP

genotype	cases	controls	total
0	a	c	m
1	b	d	n
total	$N/2$	$N/2$	N

3.3 Differential Privacy

Differential privacy (DP) aims to provide methods to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. The definition was first proposed by Cynthia Dwork [19]. Then the concentrated differential privacy which is a relaxation of differential privacy without compromising on cumulative privacy loss over multiple computations was proposed [30].

We will think of a database D_1 as a multiset of rows. Assuming a database D_2 that differs in one entry from the database D_1, which means D_1 and D_2 are adjacent databases. Then the definition of DP is as follows:

Definition 1 (ε-Differential Privacy [30]). Given $\varepsilon \geq 0$, a randomized algorithm M is ε-differentially private if for the adjacent databases D_1 and D_2, and for all sets $S \subseteq Range\left(M\left(D_1\right)\right) \cup Range\left(M\left(D_2\right)\right)$

$$\Pr\left[M\left(D_1\right) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[M\left(D_2\right) \in S\right] \tag{2}$$

That means when the mechanism M is applied to two adjacent databases D and D', the probability that two results $M\left(D_1\right)$ and $M\left(D_2\right)$ belonging to the same range is very close. This mechanism protects the privacy information of any individual in database. Whether the individual is present in the database, the outcome of the mechanism is not changed much. We can assume the randomized algorithm M is a query mechanism. The ε is called “privacy budget” which represents the degree of privacy protection. The smaller the ε, the higher the degree of privacy protection.

Differential privacy is immune to post-processing:

Lemma 1 (Post-Processing [31]). Let $F: D \to D'$ be a random mechanism that is ε-differentially private on dataset D. Let $f: D' \to D''$ be an arbitrary randomized mapping. Then $f \circ F: D \to D''$ is ε- differentially private.

Before we introduce Laplace Mechanism (LM) and discrete Laplace mechanism (DLM) , we first show sensitivity, which is an important parameter that will determine the max difference under a same query function for two adjacent databases.

Definition 2 (Sensitivity). Let $f: D^n \to R^k$, for two adjacent databases $D_1, D_2 \in D^n$, the sensitivity of the query function f is

$$\Delta f = \max \left\| f\left(D_1\right) - f\left(D_2\right) \right\|_1 \tag{3}$$

The sensitivity Δf captures the max difference caused by an individual’s data between two adjacent databases. With the max difference, we can be able to know the upper bound of how much we must perturb the function’s output to preserve privacy.

Definition 3 (Laplace Mechanism). Given any function $f : D^n \to R^k$, for a database D as a multiset of rows , the Laplace mechanism is defined as

$$f(D)+(Y_1,\ldots,Y_k) \tag{4}$$

Where Y_i are i.i.d random variables drawn from Laplace distribution $Lap(\Delta f/\varepsilon)$.

Definition 4 (Discrete Laplace Distribution [32]). A random variable Y has the discrete Laplace distribution with parameter $p \in (0,1)$, denoted by DL(p), if

$$f_p(k) = \mathrm{P}(Y=k) = \frac{1-p}{1+p}p^{|k|}, k \in R = 0, \pm 1, \pm 2, \ldots \tag{5}$$

Definition 5 (Discrete Laplace Mechanism). Given any function $f : D^n \to R^k$, for a database D as a multiset of rows , the discrete Laplace mechanism is defined as

$$f(D)+(Y_1,\ldots,Y_k) \tag{6}$$

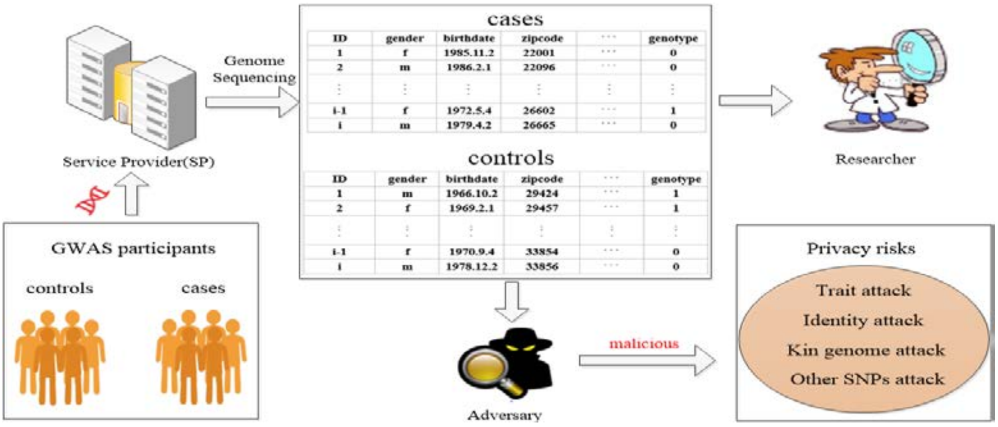
Where Y_i are i.i.d random variables drawn from discrete Laplace distribution DL(p), $p = e^{-1/b}$, and $b = \Delta f/\varepsilon$.

3.4 Nash Equilibrium

Let (S,U) be a game with n players, where S_i is the strategy set for player i, $S = S_1 \times S_2 \times \ldots \times S_n$ is the set of strategy profiles and $U(x) = (U_1(x),\ldots,U_n(x))$ is its payoff function evaluated at $x \in S$. Let x_i be a strategy of player i and x_{-i} be a strategy profile of all players except for player i. When each player $i \in \{1,\ldots,n\}$ chooses strategy x_i resulting in strategy profile $(x = x_1,\ldots,x_n)$ then player i obtains payoff $U_i(x)$. A strategy profile $x^* \in S$ is a Nash equilibrium if no unilateral deviation in strategy by any single player is profitable for that player, that is

$$\forall i, x_i \in S_i : U_i(x_i^*, x_{-i}^*) \geq U_i(x_i, x_{-i}^*) \tag{7}$$

Informally, a strategy profile is a Nash equilibrium if no player can do better by unilaterally changing his or her strategy.



cases

ID	gender	birthdate	zipcode	···	genotype
1	f	1985.11.2	22001	···	0
2	m	1986.2.1	22096	···	0
⋮	⋮	⋮	⋮	⋮	⋮
i-1	f	1972.5.4	26602	···	1
i	m	1979.4.2	26665	···	0

controls

ID	gender	birthdate	zipcode	···	genotype
1	m	1966.10.2	29424	···	1
2	f	1969.2.1	29457	···	1
⋮	⋮	⋮	⋮	⋮	⋮
i-1	f	1970.9.4	33854	···	0
i	m	1978.12.2	33856	···	0

Fig. 1. System and threat models for sharing GWAS original genotype data

4. Models and Design Goal

4.1 System Model

We consider a system which contains GWAS participants (the cases and the controls), a service provider (SP), the GWAS researcher and malicious adversary as shown in **Fig. 1**. The SP can be a direct-to-customer genomic company, a hospital or a genome research institute which collect genome data from customers or volunteers. The cases and the controls group send their biological sample such as blood and saliva to the SP, then their chemical DNA will be extracted and be used to perform genome sequencing. We assume the study is about the significance of one SNP for one disease, and the allele of the study SNP is T and C. After sequencing, a document with raw genotype data information of participants is produced, along with an analysis report. The genotype is digitalized as 0 or 1 (see in Section 3.1) and stored in the database of the SP. For the scientific progress, most of SPs will share the genomic data to researchers. For the protection of personal identity, the name will be replaced by a serial number (individual ID). The publicly accessible databases attract more researchers such as geneticists, bioinformaticians or demographers, which use these data for genetics and genomics studies.

4.2 Threat Model

Fig. 1 also shows the threat model for sharing GWAS original genotype data. Here, we assume that all GWAS participants and researchers are rational. The rational participants wish to achieve expected utility and the rational researchers wish to achieve expected privacy. Threats mainly come from malicious adversaries or some honest but curious researchers. **Fig. 2** shows four kinds of attack. Once the attack is successful, serious consequences such as genetic discrimination and blackmails occur.

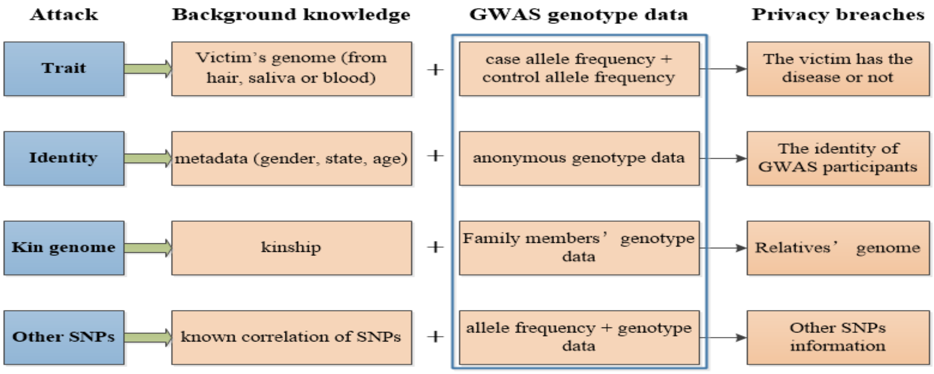
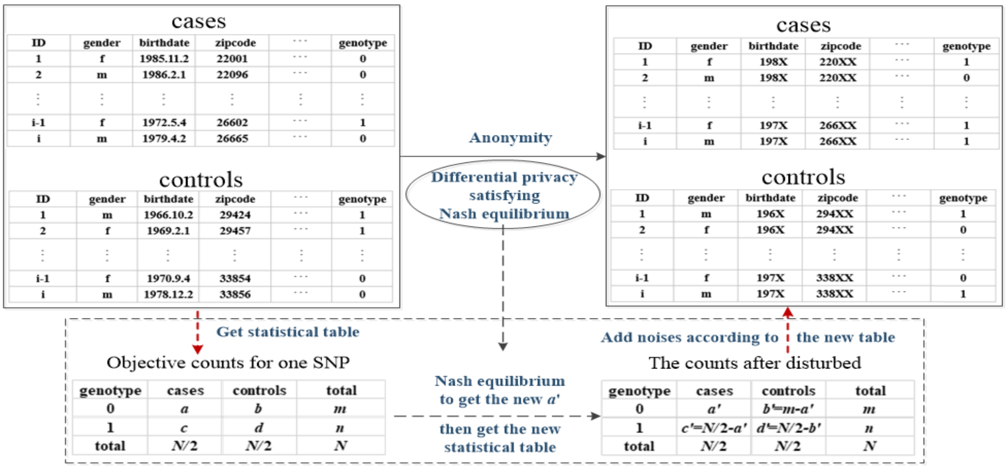

Attack
Background knowledge
GWAS genotype data
Privacy breaches
Trait
Victim's genome (from hair, saliva or blood)
+
case allele frequency + control allele frequency
The victim has the disease or not
Identity
metadata (gender, state, age)
+
anonymous genotype data
The identity of GWAS participants
Kin genome
kinship
+
Family members' genotype data
Relatives' genome
Other SNPs
known correlation of SNPs
+
allele frequency + genotype data
Other SNPs information

Fig. 2. Overview of threat in GWAS original genotype data sharing

As shown in **Fig. 2**, the GWAS genotype data can be used to perform four privacy attacks. Thus, when genotype data is shared by volunteers or service providers, the data should be pre-processed for privacy. With the goal of wide-spread data sharing and privacy preserving, we apply the differential privacy on genotype data to release protected disturbed data with expected data utility and privacy.


Anonymity
Differential privacy satisfying Nash equilibrium
Get statistical table
Nash equilibrium to get the new a'
then get the new statistical table
Add noises according to the new table

cases

ID	gender	birthdate	zipcode	···	genotype
1	f	1985.11.2	22001	···	0
2	m	1986.2.1	22096	···	0
⋮	⋮	⋮	⋮	⋮	⋮
i-1	f	1972.5.4	26602	···	1
i	m	1979.4.2	26665	···	0

controls

ID	gender	birthdate	zipcode	···	genotype
1	m	1966.10.2	29424	···	1
2	f	1969.2.1	29457	···	1
⋮	⋮	⋮	⋮	⋮	⋮
i-1	f	1970.9.4	33854	···	0
i	m	1978.12.2	33856	···	0

cases

ID	gender	birthdate	zipcode	···	genotype
1	f	198X	220XX	···	1
2	m	198X	220XX	···	0
⋮	⋮	⋮	⋮	⋮	⋮
i-1	f	197X	266XX	···	1
i	m	197X	266XX	···	1

controls

ID	gender	birthdate	zipcode	···	genotype
1	m	196X	294XX	···	1
2	f	196X	294XX	···	0
⋮	⋮	⋮	⋮	⋮	⋮
i-1	f	197X	338XX	···	0
i	m	197X	338XX	···	1

Objective counts for one SNP

genotype	cases	controls	total
0	a	b	m
1	c	d	n
total	$N/2$	$N/2$	N

The counts after disturbed

genotype	cases	controls	total
0	a'	$b'=m-a'$	m
1	$c'=N/2-a'$	$d'=N/2-b'$	n
total	$N/2$	$N/2$	N

Fig. 3. Privacy preserving model of NEDP

4.3 Design Goal

Now we want to release disturbed genotype data by our approach to promote the genetics and genomics studies. Firstly, we will use anonymity tecnology in metadata, as shown in **Fig. 3**. After data generalization, the birthdate and zipcode is protected so that it's hard to use inference attack. Now we consider how to protect the genotype data. We know that GWAS participants expect that attackers cannot infer their identity from disturbed genotype data, and they want privacy protection to be the largest. The researchers hope that the data utility of the disturbed genotype data will be guaranteed, and the GWAS statistics will remain the original statistical significance. The problem is to find the Nash equilibrium point between utility and privacy.

As shown in **Fig. 3**, the raw genotype data correspond to the observed genotype counts table. In a GWAS, to ensure the consistency of the total amount of data, the total number and the total cases or controls number can't be changed, so if we disturb the raw data, we will get a new counts table with *a*', *b*', *c*' and *d*'. Note that this means after using the differential privacy mechanisms, we should make sure the number of genotype 0 in cases is *a*' while in controls is *b*', the number of genotype 1 in cases is *c*' while in controls is *d*' . Therefore, the disturbance to the original data is equivalent to finding the right *a*', *b*', *c*' and *d*'. We know that if the "*a*'" is determined, then other three cells "*b*', *c*', *d*' " in the table can also be calculated, so the problem boils down to finding an appropriate Nash equilibrium point $a' = a^*$ -- the new number of genotype 0 in cases. After we find the appropriate Nash equilibrium point $a' = a^*$, we can add noises to change the number of genotype 0 in cases from *a* to *a*'.

How to find the Nash equilibrium point a^* ? We know individuals are concerned about their privacy and the researchers care about the utility after disturbed. As we can see, this is a game between participants and researchers while their payoff is privacy and utility, respectively, so we should define the expected privacy and utility (see in Section 4.4), then based on that, we give the strategy, the Nash equilibrium point a^*, for participants and researchers.

4.4 Utility and Privacy Metrics

Utility metric. In GWAS, researchers care about the *p*-value of the significant SNP (see in Section 3.2), so we take the *p*-value as the utility metric. We know each χ^2 corresponding to a *p*-value, so we control the range of *p*-values by controlling the range of disturbed chi-squared values. In **Table 2** , we can get the original test statistic of the SNP is:

$$\chi^2 = \frac{(2a-m)^2}{m} + \frac{(2b-n)^2}{n} = \frac{(2a-m)^2}{m} + \frac{(N-2a-n)^2}{n} \tag{8}$$

Privacy metric. To demonstrate the extent of privacy protection, we use the expected estimation error (*EEE*) [33] as privacy metric. In our mechanism, we have the noise *Y*, for raw database D and the disturbed database D' , the error measures the deviation between D and D'. It is defined as follows:

$$EEE = \sum_{i=1}^{|D|} P(Y_i)\left|D_i - D_i^{'}\right| \tag{9}$$

Where $P(Y_i)$ is the probability of noise Y_i which is added to D_i, $|D|$ is the size of genotype database D, $\left|D_i - D_i^{'}\right|$ is the absolute value of the corresponding element difference. The *EEE* intuitively shows the degree of disturbance. The larger the *EEE* is, the more disturbance we get, the higher the privacy is.

5. Proposed NEDP Approach

In this section, we propose the NEDP approach to disturb the raw genotype matrix to a new genotype matrix with the expected utility and privacy. We first use the χ^2 to constraint the interval of *a*′, then use *EEE* and *p*-value to define which *a*′ should be chosen as the Nash equilibrium point. Now we show how to transform the genotype data into a genotype matrix, then we illustrate the processing steps for the algorithm.

5.1 Genotype matrix in NEDP

In Section 3.2, we show the **Table 2** which is a statistical table for GWAS with *N* participants. We abstract the raw genotype matrix. We define a $N/2\times 2$ matrix *M*. Two columns represent cases and controls, respectively. Actually, *M* is composed of four block matrices differed in size. The raw genotype matrix is

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \tag{10}$$

Where $M_{11} = O_{a\times 1}$, $M_{12} = O_{c\times 1}$, $M_{21} = I_{b\times 1}$ and $M_{22} = I_{d\times 1}$. And O are zero matrix while all elements of I are 1, respectively. The M_{11} corresponding to the *a* cases of genotype 0. The M_{12} corresponding to the *c* controls of genotype 0. The M_{21} corresponding to the *b* cases of genotype 1. The M_{22} corresponding to the *d* controls of genotype 1. If we give the raw genotype matrix *M*, we can get the *N*, *a*, *m*, *n*. If we give the *N*, a' , *m*, *n*, the new genotype matrix is obtained by using the differential privacy protection approach satisfying Nash equilibrium.

5.2 Find the equilibrium point a^*

In order not to affect the data utility after the disturbance, we should guarantee the disturbed test statistic will remain the original statistical significance. It’s important to note that when the SNP is significant, after disturbing, its chi-squared value χ^2_{dp} may be much larger than the original χ^2, so we control the upper bound as $\chi^2_{\alpha,v}$ to avoid too much perturbance and to ensure data controllability. Thus, we have

$$\chi^2_{dp} = \frac{(2a'-m)^2}{m} + \frac{(N-2a'-n)^2}{n} \tag{11}$$

$$\chi^2 > \chi^2_{dp} > \chi^2_{\alpha,v} \tag{12}$$

From Equation (11) and Inequality (12), we can solve $a' \in (a_1, a_2) \cup (a_3, a_4)$. We have

$$a_1 = \frac{m-\sqrt{m^2-C_1}}{2},\ a_2 = \frac{m-\sqrt{m^2-C_2}}{2},\ a_3 = \frac{m+\sqrt{m^2-C_2}}{2},\ a_4 = \frac{m+\sqrt{m^2-C_1}}{2}$$

where $C_1 = m^2 - nm\chi^2/N$, $C_2 = m^2 - nm\chi^2_{\alpha,v}/N$.

Choose the equilibrium point for the disturbed matrix. When we calculate the range of a', there are two intervals $(a_1,a_2),(a_3,a_4)$ because of the inequation $\chi^2 > \chi^2_{dp} > \chi^2_{\alpha,v}$. According to Section 4.4, the greater the difference between a' and the original *a*, the more the perturbation and the better the privacy protection effect. Thus, we should choose the interval that is far from the original value *a*. Similarly, because the chi-squared value is a quadratic function of a', when a' is closer to the boundary point, the chi-squared value χ^2_{dp} is closer to the original boundary chi-square value χ^2. Thus, we get $a^* = a_1 + 1$ (when original *a* is far from (a_1,a_2)) or $a^* = a_4 - 1$ (when original *a* is far from (a_3,a_4)).

5.3 Disturb the raw genotype matrix

After we find the equilibrium point a^*, we add noises to the raw genotype matrix M by column and then get M'. We have

$$M' = round(M+Y) \bmod 2 = \begin{bmatrix} M'_{11} & M'_{12} \\ M'_{21} & M'_{22} \end{bmatrix} \tag{13}$$

Where $M'_{11} = O_{a^*\times 1}$, $M'_{12} = O_{c^*\times 1}$, $M'_{21} = I_{b^*\times 1}$ and $M'_{22} = I_{d^*\times 1}$. The M'_{11} corresponds to the a^* cases of genotype 0. The M'_{12} corresponds to the $c^* = m - a^*$ controls of genotype 0. The M'_{21} corresponds to the $b^* = N/2 - a^*$ cases of genotype 1. The M'_{22} corresponds to the $d^* = N/2 - m + a^*$ controls of genotype 1.

The disturbed matrix is not available unless we normalize the data into {0,1} because the genotype is 0 or 1. For each noise $y \in Y$, we need to make $M_{ij} + y$ be an integer and belong to {0,1}. Considering the discreteness of genotype, we also use discrete Laplace mechanism to get better utility. The noises are discrete integers so we don’t need the rounding operation, but need the modular arithmetic to let $M_{ij}+y$ belong to {0,1}. When we convert each result after adding noises to an integer, the method here can be ceil operation, floor operation, or rounding operation. However, noise is already a disturbance, so we want to reduce the disturbance as much as possible, only the result of the rounding operation is more close to the original value, the disturbance is smaller, and the data utility is relatively better, so we choose the rounding operation to limit the value to {0, 1}.

In Algorithm 1, we first calculate the original chi-square value χ^2, and we assume the disturbed chi-squared value χ^2_{dp}, which is discussed in Section 5.2. Then we calculate the range of a' and choose the equilibrium point a^* from a' based on Section 5.2. In Step 3, we generate noises Y which will be used on the orginal genotype metrix M. Then we add noises to M. When the first column M_{1*} is determined, the second column can be calculated. So in Step 4, we add the noises Y to $M_{1\mathrm{j}}$ -- the *j*-th value of the first column of M, and the goal is to make the number of genotype 0 in M_{1*} is a^*. We use Num*a* to count the number of genotype 0 in the disturbed $M_{1\mathrm{j}}'$, while Num*b* is to count the number of genotype 1 in the disturbed $M_{1\mathrm{j}}'$. To get Num*a* = a^*, we will judge if the noise Y_i can make $M_{1\mathrm{j}}'=0$ under the constraints of condition NUM*a* < a^*, if it can, we will add the Y_i to $M_{1\mathrm{j}}$, else we will judge if Y_i can make $M_{1\mathrm{j}}'=1$ under the constraints of condition NUM*b* < *N*/2- a^*, if it can, we will also add the Y_i to $M_{1\mathrm{j}}$. However, when the noise Y_i cannot satisfy the two judgements, we will discard the noise Y_i and choose the next noise Y_{i+1}. After selecting the noises and conducting post-processing operations such as rounding and modular arithmetic, we get the disturbed genotype matrix M', and then output it.

Algorithm 1. *The ε - differentially private algorithm for releasing the genomic data with expected utility and privacy in NEDP.*

Input: The raw genotype matrix M

Output: The disturbed genotype matrix M' which has the expected χ^2 、 p-value and EEE.

Global: Privacy budget ε, sensitivity Δf, and the significance critical value $\chi^2_{\alpha,v}$. Total number N of individuals, the number a of genotype 0 in cases, the total number m of genotype 0, the total number n of genotype 1.

1. Function Calculate_Chi (N, a, m, n):

$$\chi^2=\frac{(2a-m)^2}{m}+\frac{(2b-n)^2}{n}=\frac{(2a-m)^2}{m}+\frac{(N-2a-n)^2}{n},\ \chi^2_{dp}=\frac{(2a'-m)^2}{m}+\frac{(N-2a'-n)^2}{n}$$

2. Choose the equilibrium point a^*:

    let $\chi^2>\chi^2_{dp}>\chi^2_{\alpha,v}$ to calculate $a'\in(a_1,a_2)\cup(a_3,a_4)$;

    if original a is far from (a_1,a_2), let $a^*=a_1+1$;

```
        else if original a is far from $(a_3, a_4)$, let $a^*=a_4-1$;
        return $a^* \in \{a_1+1, a_4-1\}$
3.  Function GenerateNoise( $\varepsilon$ , $\Delta f$ )
    Suppose the size of noises Y is k.
    for i =1 : k ,  generate noise Y(i), and the probability P(i) of each random noise
4.  Function Imnoise ( $M$ , Y, P, $a^*$ , N,m,n )
    Initialize NUMa=0, NUMb=0, j=0,i=0;
    while NUMa+NUMb<N/2
        $s=round\left(M_{1j}+Y_i\right)\bmod 2$
            if $s=0$ && NUMa< $a^*$ , $M_{1j}'=s$ , Numa++,  j++;
            else if  $s=1$ && NUMb<N/2- $a^*$ , $M_{1j}'=s$ , Numb++,  j++;
        i++;
    return $M'=round\left(M+Y\right)\bmod 2$
```

The NEDP algorithm is as follows. The function name followed by parameters.

There are two obvious problems which are solved through our approach: (i) It guarantees the accuracy of GWAS outcome from the disturbed data set, which means that the utility should not loose much and the disturbance should not be excessive, (ii) It balances the privacy-preserving level and the utility of the data set to satisfy both GWAS participants and researchers. For the first problem, we use the Inequality (12) to constraint the utility. For the second problem, we introduce the game theory to realize the equilibrium between utility and privacy.

6. Experimental Analysis

We firstly show NEDP approach and analyze the privacy and utility in the approach. We will use discrete Laplace mechanism (DLM) and Laplace mechanism (LM) in NEDP, and we call NEDP with DLM as NEDPD and NEDP with LM as NEDPL. Then we perform a comparison analysis between independent variable perturbation (IVP) [29] and our NEDPD and NEDPL approach.

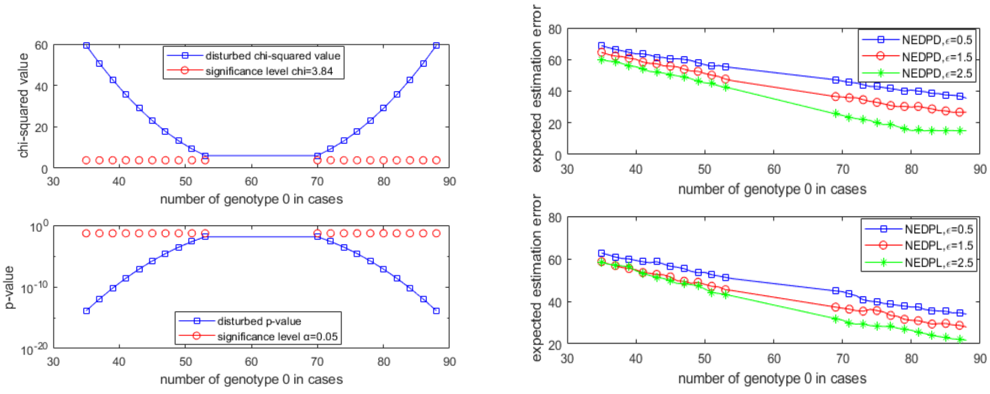
Data set. We use synthetic dataset based on genotypes of chromosome 22 in the 1000 genome project (phase 3). We get $N=200$ participants with $N/2$ cases and $N/2$ controls, the number of genotype 0 in cases is $a=89$, the total number of genotype 0 is $m=123$, the total number of genotype 1 is $n=77$, then we get $b=N/2-a=11$, $c=m-a=34$, $d=N/2-c=66$, so we have the contingency table (**Table 4**). The raw genotype matrix M with $M_{11}=O_{89\times1}$, $M_{12}=O_{34\times1}$, $M_{21}=I_{11\times1}$ and $M_{22}=I_{66\times1}$.

Table 4. The contingency table in the GWAS experiment

genotype	cases	controls	total
0	89	34	123
1	11	66	77
total	100	100	200

Calculate the chi-squared value and the Choose the equilibrium point a^* . For the 2×2 contingency table, the critical value is $\chi^2_{\alpha,v}=\chi^2_{0.05,1}=3.84$. The test statistic $\chi^2=63.879$ and the *p*-value is 1.3229×10^{-15}. These illustrate the SNP makes a lot of sense for the disease. We calculate the interval of a' is $(34,55)\cup(68,89)$**. According to section 5.2,** we should choose the interval that is far from the original value $a=89$. The interval (34,55) has greater deviation than the interval (68,89), so we just consider the interval (34,55). In NEDP, we choose $a^*=34+1$. The following figures will contain all interval, just for intuitive and comprehensive understanding. In practical application, we should know that just $a^*=34+1$ can be chosen.

Utility analysis. Here we show the *p*-value and the chi-squared value for each a' in **Fig. 4(a)**. We find all the *p*-values are smaller than the significant level $\alpha=0.05$, and all the chi-squared values are larger than $\chi^2_{\alpha,v}=\chi^2_{0.05,1}=3.84$, which means whichever a' we choose, it satisfied the expected utility--the SNP after disturbance still remains significant. We use scatter diagram so that there is a line segment in (55,68), actually it is meaningless. Because the Equation (8) is a quadratic equation, so the *p*-values are distributed symmetrically between these two intervals $(34,55)\cup(68,89)$.



(a) (b)

Fig. 4. Utility and privacy using NEDP

Privacy analysis. After we guarantee the utility, we should consider the privacy metric *EEE* --the expected estimation error. Each a' corresponds to different matrix M'. *EEE* shows the difference between M and M'. **Fig. 4(b)** shows the relationship between a' and *EEE*. We can see intuitively that whether in NEDPD or in NEDPL, the more deviation between a' and $a=89$, the larger the *EEE*. Essentially the greater the deviation between a' and the original value a, the larger the *EEE*.

Equilibrium between utility and privacy. In Fig. 4(a), when $a'=35$, the *p*-value and chi-squared value are the closest to original *p*-value and chi-squared value, so when $a'=35$ we set the payoff of utility as 1, and the other as 0. In Fig. 4(b), under NEDPD and NEDPL , we choose any user defined ε, are found that the largest *EEE* is when $a'=35$. For patients, the highest *EEE* is their expected privacy. Thus, we set the payoff of privacy as 1 when $a'=35$, and the others are 0. We get a payoff matrix shown in Table 5. Here (u_1, u_2) represents the payoff of utility and privacy.

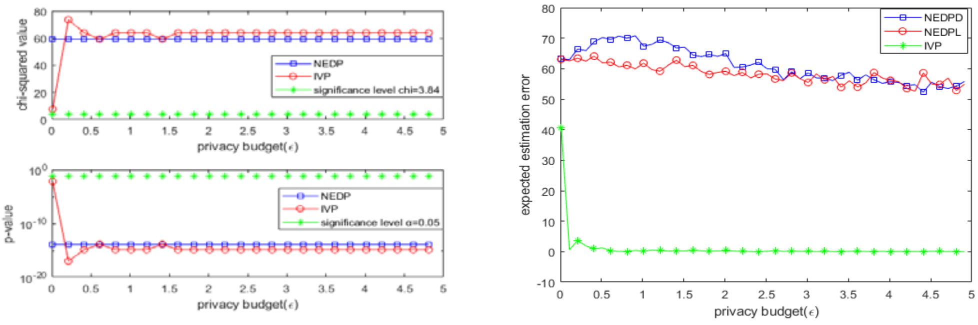
Table 5. The payoff matrix between utility and privacy

(u_1, u_2)	**35**	⋯	**54**
35	(1,1)	(1,0)	(1,0)
⋯	(0,1)	(0,0)	(0,0)
54	(0,1)	(0,0)	(0,0)

Actually the patients and the researchers should choose the same a' when we publish the genotype database, so here we just consider the payoff on the diagonal. When researchers and patients choose $a^*=35$, both of them receive the expected utility and expected privacy. Thus, $a^*=35$ is a Nash equilibrium point in this example. We disturb the Table 4 to Table 6 and choose the genotype matrix M' when $a^*=35$ in NEDPD or NEDPL with expected ε.

Table 6. The contingency table about the significant SNP after our approach

genotype	**cases**	**controls**	**total**
0	35	88	123
1	65	12	77
total	100	100	200



(a) (b)

Fig. 5. Utility and privacy comparison between NEDP and IVP

Comparison between our approach and IVP. In Section 3.6 of Simmons's paper [29], they denoted the independent variables of the statistical test value (corresponding to the χ^2 in this paper) by *x* and *y* (corresponding to the "*a*" in this paper), then they made $x_{dp} = x + Lap(\Delta f / \varepsilon)$ as well as *y*, so (x_{dp}, y_{dp}) is a ε- differentially private estimate of (x, y). For convenience, we refer to that as independent variable perturbation (IVP). IVP achieves good data utility, but IVP did not conduct privacy analysis. Here we compare the utility and the privacy between NEDP and IVP. Firstly, we show the utility comparison – *p*-values and chi-squared values of IVP and NEDP in **Fig. 5(a)**. In IVP, different ε will lead to different $a^{'}$, each $a^{'}$ lead to different chi-squared values and *p*-values. In NEDP, when the original $a = 89$, we already calculate the $a^{*} = 35$, so the chi-squared values and *p*-values are also fixed, depending on $a^{*} = 35$. Here we choose ε from 0.01 to 5, and calculate the chi-squared values and *p*-values when the original $a = 89$. As we can see, the chi-squared values of IVP and NEDP are very close to original chi-squared value $\chi^{2} = 63.879$, and still larger than the significance level. As a result, both IVP and NEDP have met our highest expectations for utility.

Then we show the privacy comparison between our approach and IVP. In **Fig. 5(b)**, the *EEE* of IVP is obviously smaller than that of our approach, which means our approach for privacy protection is better, whether in NEDPD or NEDPL. IVP method is enough to release the statistical value of SNP, but the perturbation of raw data is very small, therefore, that it’s privacy for protecting genotype data is not so good. We also find that when ε is small, the *EEE* of NEDPD is larger than NEDPL, which means that the privacy of NEDPD is better than NEDPL. However, when ε increases, the *EEE* of NEDPD and NEDPL are close.

7. Conclusion

To share disturbed genotype data without compromising participants’ privacy, this paper proposes a differential privacy preserving approach by adding noises to raw data, and also guarantees the equilibrium between privacy and data utility. In GWAS, for the 2×2 contingency table such as **Table 2**, the Nash equilibrium point(s) a^{*} between utility and privacy can be found in NEDP. According to our experimental results on synthetic data, the disturbed genotype data generated by a^{*}=35 satisfied expected p-value and EEE.Moreover, we find that the perturbance of discrete Laplace mechanism is more than Laplace mechanism, and the data privacy of NEDPD is better than NEDPL when ε is small, especially in the range of (0, 2). In addition, we compare our approach with IVP approach and found out that IVP has poor privacy protection and it’s *EEE* is much lower than our approach. Our approach can also be used to perturb data and prevent privacy attacks on other bioinformatics data that can be represented by 2×2 contingency tables, promoting the data sharing and research of bioinformatics and biostatistics.

References

[1] C. M. Lewis, “Genetic association studies: design, analysis and interpretation,” *Briefings in Bioinformatics*, vol. 3, no. 2, pp. 146-153, Jun. 2002. Article(CrossRef Link)

[2] P. M. Visscher, M. A. Brown, M. I. McCarthy and J. Yang, “Five years of GWAS discovery,” *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7-24, Jan. 2012. Article (CrossRef Link)

[3] A. Morgan, D. Vuckovic, N. Krishnamoorthy, E. Rubinato, U. Ambrosetti, P. Castorina, A. Franze, D. Vozzi, M. L. Bianca, S. Cappellani, M. D. Stazio, P. Gasparini and G. Girotto, “Next-generation sequencing identified SPATC1L as a possible candidate gene for both early-onset and age-related hearing loss,” *European Journal of Human Genetics*, vol. 27, no.1, pp.70-79, Sep. 2019. Article (CrossRef Link)

[4] A. Xue, Y. Wu, Z. Zhu, F. Zhang, K. E. Kemper, Z. Zheng, L. Yengo, L. R. Lloyd-Jones, J. Sidorenko, Y. Wu, A. F. McRae, P. M. Visscher, J. Zeng and J. Yang, "Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes," *Nature Communications*, vol. 9, no. 1, Article no. 2941, Jul. 2018. Article (CrossRef Link)

[5] L. Zhou, and F. Zhao, "Prioritization and functional assessment of noncoding variants associated with complex diseases," *Genome Medicine*, vol. 10, no. 1, pp. 53, Jul. 2018. Article (CrossRef Link)

[6] T. W. Winkler, C. Brandl, F. Grassmann, M. Gorski, K. Stark, J. Loss, B. H. F. Weber and I. M. Heid, "Investigating the modulation of genetic effects on late AMD by age and sex: Lessons learned and two additional loci," *PloS One*, vol. 13, no. 3, Mar. 2018. Article (CrossRef Link)

[7] A. Kong, G. Thorleifsson, M. L. Frigge, B. J. Vilhjálmsson, A. I. Young, T. E. Thorgeirsson, S. Benonisdottir, A. Oddsson, B. V. Halldorsson, G. Másson, D. F. Gudbjartsson, A. Helgason, G. Bjornsdottir, U. Thorsteinsdottir and K. Stefánsson, "The nature of nurture: Effects of parental genotypes," *Science*, vol. 359, no. 6374, pp. 424-428, Jan. 2018. Article (CrossRef Link)

[8] E. Ayday and J. P. Hubaux, "Privacy and Security in the Genomic Era," in *Proc. of ACM Conf. CCS*, Vienna, Austria, pp. 1863-1865, Oct. 2016. Article (CrossRef Link)

[9] T. Wang, Z. Cao, S. Wang, J. Wang, L. Qi, A. Liu, M. Xie, X. Li, "Privacy-Enhanced Data Collection Based on Deep Learning for Internet of Vehicles," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6663-6672, Oct. 2020. Article (CrossRef Link)

[10] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson and D. W. Craig, "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays," *PLoS Genetics*, vol. 4, no. 8, Aug. 2008. Article (CrossRef Link)

[11] E. E. Schadt, S. Woo and K. Hao, "Bayesian method to predict individual SNP genotypes from gene expression data," *Nature Genetic.*, vol. 44, no. 5, pp. 603-608, Apr. 2012. Article (CrossRef Link)

[12] R. Cai, Z. Hao, M. Winslett, X. Xiao, Y. Yang, Z. Zhang and S. Zhou, "Deterministic identification of specific individuals from GWAS results," *Bioinformatics*, vol. 31, no. 11, pp. 1701-1707, Jan. 2015. Article (CrossRef Link)

[13] M. Alibakhshikenari, B. S. Virdee, P. Shukla, C. H. See, R. Abd-Alhameed, M. Khalily, F. Falcone, and E. Limiti, "Antenna Mutual Coupling Suppression Over Wideband Using Embedded Periphery Slot for Antenna Arrays," *Electronics*, vol. 7, no. 9, Sep. 2018. Article (CrossRef Link)

[14] M. Alibakhshikenari, B. S. Virdee, E. Limiti, "Compact Single-Layer Traveling-Wave Antenna Design Using Metamaterial Transmission Lines," *Radio Science*, vol. 52, no. 12, pp.1510-1521, Dec. 2017. Article (CrossRef Link)

[15] M. Alibakhshikenari, B. S. Virdee, E. Limiti, "Wideband planar array antenna based on SCRLH-TL for airborne synthetic aperture radar application," *Journal of Electromagnetic Waves and Applications*, vol. 32, no. 12, pp.1586-1599, May. 2018. Article (CrossRef Link)

[16] S. Wang, Y. Zhang, W. Dai, K. E. Lauter, M. Kim, Y. Tang, H. Xiong and X. Jiang, "HEALER: homomorphic computation of exact logistic regression for secure rare disease variants analysis in GWAS," *Bioinformatics*, vol. 32, no. 2, pp. 211-218, Oct. 2015. Article (CrossRef Link)

[17] R. Mott, C. Fischer, P. Prins, R. W. Davies, "Private Genomes and Public SNPs: Homomorphic Encryption of Genotypes and Phenotypes for Shared Quantitative Genetics," *Genetics*, vol. 215, no. 2, pp. 359–372, June 2020. Article (CrossRef Link)

[18] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. S. Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. Ford, J. P. Hubaux, "MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1328-1341, Jul. 2019. Article (CrossRef Link)

[19] C. Dwork, "Differential Privacy," in *Proc. of ICALP*, Venice, Italy, pp. 1-12, 2006. Article (CrossRef Link)

[20] S. E. Fienberg, A. B. Slavkovic and C. Uhler, "Privacy Preserving GWAS Data Sharing," in *Proc. of ICDMW*, Vancouver, BC, Canada, pp. 628-635, 2011. Article (CrossRef Link)

[21] F. Tramèr, Z. Huang, J. P. Hubaux and E. Ayday, "Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies," in *Proc. of CCS*, New York, NY, USA, pp. 1286-1297, 2015. Article (CrossRef Link)

[22] M. Wang, Z. Ji, S. Wang, J. Kim, H. Yang, X. Jiang and L. Ohno-Machado, "Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies," *Bioinformatics*, vol. 33, no. 23, pp. 3716-3725, July, 2017. Article (CrossRef Link)

[23] E. M. Ramos, C. Din-Lovinescu, E. B. Bookman, L. J. McNeil, C. C. Baker, G. Godynskiy, E. L. Harris, T. Lehner, C. McKeon, J. M. Moss, V. L. Starks, S. T. Sherry, T. A. Manolio and L. L. Rodriguez, "A mechanism for controlled access to GWAS data: experience of the GAIN Data Access Committee," *American Journal of Human Genetics*, vol. 92, no. 4, pp. 479-488, April, 2013. Article (CrossRef Link)

[24] L. Kamm, D. Bogdanov, S. Laur and J. Vilo, "A new way to protect privacy in large-scale genome-wide association studies," *Bioinformatics*, vol. 29, no. 7, pp. 886-893, April, 2013. Article (CrossRef Link)

[25] S. S. Shringarpure and C. D. Bustamante, "Privacy Risks from Genomic Data-Sharing Beacons," *American Journal of Human Genetics*, vol.97, no. 5, pp. 631-646, November, 2015. Article (CrossRef Link)

[26] A. Johnson and V. Shmatikov, "Privacy-preserving data exploration in genome-wide association studies," in *Proc. of KDD*, Chicago, USA, pp.1079-1087, 2013. Article (CrossRef Link)

[27] C. Uhler, A. Slavkovic and S. E. Fienberg, "Privacy-Preserving Data Sharing for Genome-Wide Association Studies," *The Journal of Privacy and Confidentiality*, vol. 5, no. 1, pp. 137-166, Oct. 2013. Article (CrossRef Link)

[28] F. Yu, S. E. Fienberg, A. B. Slavkovic and C. Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies," *Journal of Biomedical Informatics*, vol. 50, pp. 133-141, Aug., 2014. Article (CrossRef Link)

[29] S. Simmons and B. Berger, "Realizing privacy preserving genome-wide association studies," *Bioinformatics*, vol. 32, no. 9, pp. 1293-1300, May, 2016. Article (CrossRef Link)

[30] C. Dwork and G. N. Rothblum, "Concentrated Differential Privacy," *ArXiv*, pp. 1-28, Mar. 2016. Article (CrossRef Link)

[31] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211-407, Aug. 2014. Article (CrossRef Link)

[32] S. Inusah and T. J. Kozubowski, "A discrete analogue of the Laplace distribution," *Journal of Statistical Planning and Inference*, vol. 136, no. 3, pp. 1090-1102, March, 2006. Article (CrossRef Link)

[33] I. Wagner, "Evaluating the Strength of Genomic Privacy Metrics," *ACM Transactions on Privacy and Security*, vol. 20, no. 1, pp. 1-34, February, 2017. Article (CrossRef Link)

Jun Yan received the M.S. degree in College of Earth Exploration Science and Technology from Jilin University. He is currently pursuing the Ph.D. degree in School of Computer Sciensce, Shaanxi Normal University. His research interests include network security and privacy preserving.

Ziwei Han received the B.E. degree in computer science and technology from NorthWest University and the M.E. degree in computer science and technology from Shaanxi Normal University, China, respectively. Her research interests include Security and Privacy Preserving

Yihui Zhou received her B.E. degree, M.S. degree and Ph.D. degree in College of Mathematics and Information Science from Shaanxi Normal University, Shaanxi, China, in 2003, in 2006 and in 2009, respectively. Now she is a lecturer in School of Computer Science, Shaanxi Normal University. Her research interests include information security and privacy preserving.

Laifeng Lu received M.S. degree and Ph.D.degree in Computer system architecture from Xi'dian University, Shaanxi, China. Now she is an associate professor in Shaanxi Normal University. Her research interests include security and privacy protection.