

Privacy Preserving GWAS Data Sharing

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15217
Email: fienberg@stat.psu.edu

Aleksandra Slavković
Department of Statistics
The Pennsylvania State University
University Park, PA 16802
Email: sesa@stat.psu.edu

Carline Uhler
Institute for Mathematics and its Applications
University of Minnesota
Minneapolis, MN 55455
Email: uhler@ima.umn.edu

Abstract—Traditional statistical methods for the confidentiality protection for statistical databases do not scale well to deal with GWAS (genome-wide association studies) databases and external information on them. The more recent concept of *differential privacy*, introduced by the cryptographic community, is an approach which provides a rigorous definition of privacy with meaningful privacy guarantees in the presence of arbitrary external information. Building on such notions, we propose new methods to release aggregate GWAS data without compromising an individual’s privacy. We present methods for releasing differentially private minor allele frequencies, chi-square statistics and p -values. We compare these approaches on simulated data and on a GWAS study of canine hair length involving 685 dogs. We also propose a privacy-preserving method for finding genome-wide associations based on a differentially private approach to penalized logistic regression.

Keywords—contingency tables; differential privacy; genome-wide association studies (GWAS); chi-square statistics; logistic regression; p -values; single nucleotide polymorphism (SNP).

I. INTRODUCTION

In an article that shocked the genetics community, Homer et al. [1] claimed that, under certain conditions, they could use statistical methods to infer the presence of an individual with known genotype in a mix of DNA samples from which only the minor allele frequencies (MAFs) are known. Their approach compared the MAFs of a specific individual to the distribution of MAFs in a reference population and the distribution of MAFs in a test population and then used a t -test to assess if the individual was part of the test population.

The Homer et al. [1] “attack” appeared to be generally applicable. As a reference population one might use the publicly available single nucleotide polymorphism (SNP) data from the HapMap project¹ which consists of SNP data from 4 populations with about 60 individuals each. Note that the HapMap data set does not contain any information regarding the health status of the individuals. For the test population one might use the cases in genome-wide association studies (GWAS). Before the appearance of the article [1], the averaged MAFs of the cases and the averaged MAFs of the controls were usually publicly available.

In response to Homer et al. [1], Braun et al. [2] showed that their test depends heavily on the assumption that the

genotypes of the test population, the reference population and the specific person under consideration are samples of the same underlying population and that the SNPs used in the study are independent (i.e., that there is no linkage disequilibrium present). These assumptions are usually not met in practice, and as a consequence, the Homer et al. methods lead to a high false-positive rate (e.g., [2]). Despite the apparent limitations of the Homer et al. attack on the privacy of GWAS participants, NIH immediately removed all aggregate results (averaged MAFs over cases and controls, chi-square (χ^2)-statistics and p -values) from open-access data bases. Every researcher, who wants to gain access to any of these data sets, needs to go through an approval process. This is a particularly difficult obstacle for computer scientists, mathematicians or statisticians who do not have a research record in GWAS.

Here we propose methods which allow for the release of aggregate GWAS data without compromising an individual’s privacy. Our GWAS privacy guarantees utilize the concept of *differential privacy*, recently introduced by the cryptographic community (e.g., [3]). Differential privacy provides a rigorous definition of privacy with meaningful privacy guarantees in the presence of arbitrary external information. Our contributions are as follows:

- We propose a method for the release of the averaged MAFs for the cases and for the controls in GWAS without compromising an individual’s privacy.
- We compute ϵ -differentially private χ^2 -statistics and p -values and provide a differentially private algorithm for releasing these statistics for the most relevant SNPs.
- Conditions such as cancer, heart disease, and diabetes are caused by the interaction of various genes and possibly the environment. Detecting such interaction among SNPs related to a specific phenotype (i.e., epistasis) is a main goal of GWAS. Most methods for finding epistasis are based on a two-stage approach: (1) Filtering all SNPs, e.g., using χ^2 -statistics or a simple logistic regression, to reduce the potentially interacting SNPs to a small number; (2) Further examining the loci achieving some threshold for interactions. For example, Park and Hastie [4] use a form of penalized logistic

¹<http://hapmap.ncbi.nlm.nih.gov/>

regression to test for detecting gene-gene interactions on a small number of SNPs. By adapting the work of [5] and [6] to this methodology, we derive a privacy-preserving method for GWAS, where both stages in the two-stage approach satisfy ϵ -differential privacy.

Section II describes the basic problem and relevant definitions. In Section III, we present methods for releasing ϵ -differentially private MAFs, χ^2 -statistics and p -values, and in Section IV we evaluate their statistical utility on data based on a simulation study and on a GWAS study of canine hair length involving 685 dogs. In Section V, we propose a privacy-preserving method for finding genome-wide associations based on a differentially private approach to logistic regression.

II. MAIN DEFINITIONS AND NOTATION

In a typical GWAS setting, we study the interaction between various SNPs and a binary phenotype, as for example the disease status of an individual. The binary phenotype takes values 0 (e.g., non-diseased) and 1 (e.g., diseased). We denote the total number of individuals in a GWAS by N and assume throughout the paper that the number of cases and controls is equal, i.e., there are $N/2$ cases and $N/2$ controls. This corresponds to the usual setting in GWAS and is necessary in order to achieve enough power to detect SNPs which are associated to a disease. We denote the total number of SNPs in a GWAS by M' and the number of SNPs for which we would like to release aggregate data by M . We assume that the SNPs are polymorphic with only two possible nucleotides; the SNPs therefore take values 0, 1 and 2 representing the number of minor alleles. We summarize the data for each SNP in a 3×2 contingency table, where the count in cell (i, j) consists of the number of individuals with genotype i and disease status j . We also assume throughout the paper that all margins of such a 3×2 contingency table are positive. This is motivated by the fact that in GWAS usually all SNPs with a MAF smaller than 0.05 are removed from the study.

Definition II.1. A randomized mechanism \mathcal{K} is ϵ -differentially private if, for all data sets D and D' which differ in at most one individual and for any $t \in \mathbb{R}$,

$$\frac{\Pr(\mathcal{K}(D) = t)}{\Pr(\mathcal{K}(D') = t)} \leq e^\epsilon.$$

Definition II.2. The *sensitivity* of a function $f : \mathcal{D}^N \rightarrow \mathbb{R}^d$, where \mathcal{D}^N denotes the set of all databases with N individuals, is the smallest number $S(f)$ such that

$$\|f(D) - f(D')\|_1 \leq S(f),$$

for all data sets $D, D' \in \mathcal{D}^N$ differing in a single individual.

Releasing $f(D) + b$, where b is random noise drawn from a Laplace distribution with mean 0 and scale $\frac{S(f)}{\epsilon}$ satisfies the definition of ϵ -differential privacy (e.g., see [3]).

Definition II.3. The KL divergence between two probability distributions f and g is defined by

$$D_{KL}(f||g) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx. \quad (1)$$

For the analysis of the simulation results in Section III we use the Kullback-Leibler (KL) divergence to measure the difference between two distributions such as the original χ^2 -statistic and its corresponding ϵ -differentially private version.

III. PRIVACY-PRESERVING METHODOLOGY

A. Privacy-preserving release of aggregate MAFs

We now describe a method for releasing the averaged MAFs for the cases and for the controls in GWAS which satisfies differential privacy. The true data form a table consisting of the MAFs of the cases and the controls for M SNPs; e.g., see Table I. In the following, we compute the amount of Laplace noise we need to add to such a table in order to satisfy ϵ -differential privacy.

Lemma III.1. *The sensitivity of the averaged MAFs of the cases and the controls based on N individuals, with $N/2$ cases and $N/2$ controls, for M SNPs is $\frac{2M}{N}$.*

Proof: Without loss of generality, we can assume that the individual, whose genotype we can change, belongs to the cases. Denote this individual by j . For a given SNP we denote the number of minor alleles of individual i before adding noise by a_i and the perturbed counts by a'_i . Note that $a_i = a'_i$ for all $i \neq j$. In addition, $|a_j - a'_j| \leq 2$. Therefore, for a given SNP we can compute the sensitivity of the averaged MAF as follows:

$$\left| \frac{1}{N/2} \sum_{i=1}^{N/2} \frac{a_i}{2} - \frac{1}{N/2} \sum_{i=1}^{N/2} \frac{a'_i}{2} \right| = \frac{1}{N/2} \left| \frac{a_j}{2} - \frac{a'_j}{2} \right| \leq \frac{2}{N}.$$

This holds for every SNP. As a consequence, for M SNPs the sensitivity is $\frac{2M}{N}$, namely the 1-norm of a M -dimensional vector where all entries are $\frac{2}{N}$. ■

Lemma III.1 shows that a data release mechanism that adds Laplace noise with mean 0 and scale $\frac{2M}{N\epsilon}$ to each cell entry in Table I yields ϵ -differential privacy. This result can be seen as a special case of Example 3 in [3] where every cell entry is a histogram by itself.

Similarly, if instead of releasing the averaged MAFs, we want to release M 3×2 tables containing the counts for each genotype and disease status, the sensitivity would be $2M$. Therefore, we have to add Laplace noise with mean 0 and scale $\frac{2M}{\epsilon}$ to ensure ϵ -differential privacy.

Table I
TABLE SHOWING THE AVERAGED MAFs OF THE CASES AND THE CONTROLS FOR M SNPs.

MAF	SNP 1	SNP 2	...	SNP M
Cases	0.29	0.20	...	0.11
Controls	0.27	0.31	...	0.10

B. Privacy-preserving release of χ^2 -statistics and p -values

In many GWAS settings, researchers report the χ^2 -statistics and the p -values of the most relevant SNPs. We propose a method for releasing these quantities in a differential privacy-preserving way, by first computing the sensitivity and then modifying a method proposed in [5] to release the noisy statistics corresponding to the most relevant SNPs.

Theorem III.2. *The sensitivity of the χ^2 -statistic based on a 3×2 contingency table with positive margins and $N/2$ cases and $N/2$ controls is $\frac{4N}{N+2}$.*

Proof: Let

	0	1
0	a	m-a
1	b	n-b
2	$N/2-a-b$	$N/2-m-n+a+b$

with $a, b \geq 0$, $m, n > 0$, $a \leq m$, $b \leq n$, $a + b \leq N/2$, and $m + n < N$ denote a 3×2 contingency table with positive margins and $N/2$ cases and controls each. Let

$$\mathcal{D} = \{(a, b, m, n) \in \mathbb{N} \mid m > 0, n > 0, a \leq m, b \leq n, a + b \leq N/2, m + n < N\}.$$

Then we can view the χ^2 -statistic as a function

$$\chi^2 : \mathcal{D} \longrightarrow \mathbb{R}_{\geq 0},$$

where (a, b, m, n) gets mapped to the χ^2 -statistic of the corresponding contingency table. The sensitivity corresponds to the values of $(a, b, m, n) \in \mathcal{D} \cap \{a \geq 1\}$, which maximize

$$|\chi^2(a, b, m, n) - \chi^2(a - 1, b + 1, m - 1, n + 1)|.$$

Our approach is to compute the sensitivity by maximizing the directional derivative of $\chi^2(a, b, m, n)$ in direction $(-1/2, 1/2, -1/2, 1/2)$. First note that

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

We then compute the directional derivative of $\chi^2(a, b, m, n)$ in direction $(-1/2, 1/2, -1/2, 1/2)$. It is given by

$$\frac{2a^2}{m^2} - \frac{4a}{m} - \frac{2b^2}{n^2} + \frac{4b}{n}.$$

Over $\mathcal{D} \cap \{a \geq 1\}$ this is maximized by the smallest possible value of a , the largest possible value of m , the largest possible value of b and the smallest possible value of n . Consequently, the sensitivity is given by:

$$\left| \chi^2 \left(\begin{bmatrix} 1 & N/2 \\ N/2 - 2 & 0 \\ 1 & 0 \end{bmatrix} \right) - \chi^2 \left(\begin{bmatrix} 0 & N/2 \\ N/2 - 1 & 0 \\ 1 & 0 \end{bmatrix} \right) \right|,$$

which we can easily see to be $\frac{4N}{N+2}$. ■

Note that the sensitivity of the χ^2 -statistic grows as a function of N , but is asymptotically constant. This is interesting since the χ^2 -statistic for a table with fixed frequencies grows proportional to N . In order to achieve ϵ -differential privacy, we need to add Laplace noise with scale $\frac{1}{\epsilon} \frac{4N}{N+2}$ to the true χ^2 -statistics. Thus for increasing N , the perturbed χ^2 -statistics get more accurate. For related simulations that demonstrate the interactive effect of sample size and privacy level ϵ and compare asymptotic efficiency of private and non-private estimators for 2×2 tables and the corresponding χ^2 -statistics, see [7].

We can perform a similar analysis on the p -values corresponding to the χ^2 -statistics assuming a χ^2 -distribution with 2 degrees of freedom as null distribution, cf. [8].

Theorem III.3. *The sensitivity of the p -values of the χ^2 -statistic for a 3×2 contingency table with positive margins and $N/2$ cases and $N/2$ controls is $\exp(-2/3)$, when the null distribution is χ^2 -distribution with 2 degrees of freedom.*

Proof: Under the null χ^2 -distribution with 2 degrees of freedom, the p -value corresponding to a value x of the χ^2 -statistic is

$$\exp\left(-\frac{x}{2}\right), \quad x \geq 0.$$

The first derivative in absolute value is maximized by $x = 0$. Therefore, the sensitivity of the p -value is given by a change of 1 unit in a contingency table with $\chi^2 = 0$, i.e., in a contingency table of the form

$$\begin{bmatrix} a & a \\ b & b \\ N/2 - a - b & N/2 - a - b \end{bmatrix},$$

where $a, b > 0$, and $a + b < N/2$. We therefore need to find a, b which maximize

$$\left| p\text{-value} \left(\begin{bmatrix} a & a \\ b & b \\ N/2 - a - b & N/2 - a - b \end{bmatrix} \right) - p\text{-value} \left(\begin{bmatrix} a - 1 & a \\ b + 1 & b \\ N/2 - a - b & N/2 - a - b \end{bmatrix} \right) \right|,$$

where $a, b > 0$, and $a + b < N/2$. Equivalently, we need to maximize

$$\chi^2 \left(\begin{bmatrix} a - 1 & a \\ b + 1 & b \\ N/2 - a - b & N/2 - a - b \end{bmatrix} \right)$$

over $a, b > 0$, and $a + b < N/2$. The corresponding χ^2 -statistic is given by

$$\frac{1}{2a - 1} + \frac{1}{2b + 1},$$

which is maximized by $a = b = 1$ and results in a χ^2 -statistic of $4/3$. Consequently, the sensitivity of the p -value is $\exp(-2/3)$. ■

The ϵ -differentially private mechanism for a single SNP would then release a private p -value equal to the original value plus Laplace noise with mean zero and scale $\frac{1}{\epsilon} \exp(-2/3)$.

The sensitivity of the χ^2 -statistic corresponds to the most ‘dependent’ contingency table while the sensitivity of the p -value is determined by an ‘independent’ contingency table. By the most ‘dependent’ (resp. ‘independent’) contingency table we mean a table which achieves the maximal (resp. minimal) χ^2 -statistic over all contingency tables with N individuals. The maximal χ^2 -statistic is N , while the minimal χ^2 -statistic is 0.

Since in practice we are not interested in contingency tables with very large p -values, we in effect have overestimated the sensitivity of the p -value, and wish instead to determine the sensitivity of the p -value within the range of “interesting” contingency tables. We therefore analyze what happens if we project all p -values, which are larger than a given value p^* , onto p^* . Since the χ^2 -statistic for a table with fixed marginal frequencies grows in proportion to N , we analyze the situation where p^* decreases with increasing N , i.e., $p^* = \exp(-N/c)$, where c is some constant to be specified by the user. Such a p -value corresponds to a table with χ^2 -statistic $2N/c$ and can be viewed as a contingency table which is at least N/c steps of Hamming distance 1 away from independence.

Corollary III.4. *Projecting all p -values which are larger than $p^* = \exp(-N/c)$ onto p^* results in a sensitivity of*

$$\exp\left(-\frac{N}{c}\right) - \exp\left(-\frac{N(2Nc - 4N - 4c + c^2)}{2c(Nc - 2N - c)}\right)$$

for any fixed constant $c \geq 3$, which is a factor of $N/2$.

Proof: The proof is similar to the proofs of Theorem III.2 and Theorem III.3. We here give a sketch. The contingency table

$$\begin{bmatrix} 0 & \frac{N}{c} \\ \frac{N}{2c} & 0 \\ \frac{N(c-2)}{2c} & \frac{N(c-2)}{2c} \end{bmatrix}$$

has a χ^2 -statistic $\frac{2N}{c}$ and hence a p -value of $\exp(-N/c)$. This table has the maximal χ^2 -statistic over all tables which are N/c steps of Hamming distance 1 away from independence, i.e., this table is N/c steps away from the following table

$$\begin{bmatrix} \frac{N}{2c} & \frac{N}{2c} \\ \frac{N}{2c} & \frac{N}{2c} \\ \frac{N(c-2)}{2c} & \frac{N(c-2)}{2c} \end{bmatrix}.$$

The largest change in χ^2 -statistic is achieved by moving one individual from cell (3, 2) to cell (1, 2) resulting in the table

$$\begin{bmatrix} 0 & \frac{N+c}{c} \\ \frac{N}{2c} & 0 \\ \frac{N(c-2)}{2c} & \frac{N(c-2)-2c}{2c} \end{bmatrix}.$$

This new contingency table has χ^2 -statistic

$$\frac{N(2Nc - 4N - 4c + c^2)}{c(Nc - 2N - c)}.$$

For large N ,

$$\frac{N(2Nc - 4N - 4c + c^2)}{c(Nc - 2N - c)} \approx \frac{2N}{c},$$

and the corresponding p -value is of the order of p^* . ■

In GWAS settings, researchers typically provide only the χ^2 -statistics or the corresponding p -values of the M most significant SNPs. Since the ranking reveals additional information, it is not sufficient to add the above computed noise to these statistics in order to achieve differential privacy. Bhaskar et al. [5] show in the context of frequent pattern recognition how to release the most significant patterns together with their frequencies while satisfying differential privacy. In the following, we adapt their method to GWAS.

Let M' denote the total number of SNPs in a GWAS and M the number of statistics one would like to release. Naively, one might expect that it is necessary to add Laplace noise with scale $\frac{M'}{\epsilon} \frac{4N}{N+2}$ for the χ^2 -statistics and $\frac{M'}{\epsilon} \exp(-2/3)$ for the p -values. As we see in the following algorithm, however, the Laplace noise only scales with the number of actually released statistics M .

Algorithm 1 The ϵ -Differentially Private Algorithm for Releasing the M Most Relevant SNPs

Input: The χ^2 -statistics (resp. p -values) for all M' SNPs and the number of statistics, M , we want to release.

Output: The M noisy χ^2 -statistics (resp. p -values).

1. Add Laplace noise with mean zero and scale $\frac{4M}{\epsilon} \frac{4N}{N+2}$ to the χ^2 -statistics (resp. Laplace noise with mean zero and scale $\frac{4M}{\epsilon} \exp(-2/3)$ to the p -values).
 2. Pick the top M SNPs with respect to the perturbed χ^2 -statistics (resp. p -values). We denote the corresponding set of SNPs by \mathcal{S} .
 3. Add new Laplace noise with mean zero and scale $\frac{2M}{\epsilon} \frac{4N}{N+2}$ to the true χ^2 -statistics of the SNPs in \mathcal{S} (resp. Laplace noise with mean zero and scale $\frac{2M}{\epsilon} \exp(-2/3)$ to the true p -values) and release these perturbed statistics.
-

Theorem III.5. *Algorithm 1 is ϵ -differentially private.*

Proof: Using the sensitivities computed in Theorem III.2 and Theorem III.3, the proof follows immediately from Theorem 5 in [5]. ■

IV. RESULTS

We now evaluate the performance of the proposed methods based on data from a simulation study and using a

GWAS data set consisting of 685 dogs and their hair length. The GWAS data for the hair length of dogs has first been presented and studied in [9] and further been analyzed in [10]. It consists of 685 dogs, 319 dogs with long hair as cases and 364 with short hair as controls, and contains 40,842 SNPs. Cadieu et al. [9] have shown that the long versus short hair phenotype is associated with a mutation in the *fibroblast growth factor-5* (*FGF5* gene) and the largest χ^2 -statistic is achieved by a SNP located on chromosome 32 at position 7,100,913, i.e., about 300Kb apart from *FGF5*.

We use the simulations from [10] performed using HAP-SAMPLE [11]. HAP-SAMPLE generates the cases and controls by resampling from HapMap. The simulated data show linkage disequilibrium and allele frequencies similar to real data. The simulated association studies consist of 400 cases and 400 controls with about 10,000 SNPs per individual (SNPs typed with the Affy CHIP on chromosome 9 and chromosome 13 of the Phase I/II HapMap data). Two SNPs were chosen to be causative and the simulations were performed for three different MAFs (0.1, 0.25 and 0.4) and two different models of interaction (additive effect and multiplicative effect of the two SNPs). See [10] for more details.

For this paper, we omit the simulation results on the statistical utility of ϵ -differentially private release of aggregate MAFs. Our results are similar to those reported in the current literature on Laplace mechanism for noise addition to histograms or smaller contingency tables with proportions (e.g., [3], [7]). Instead we focus on the release of differentially-private χ^2 -statistics, p -values and the most relevant SNPs.

A. Differentially private χ^2 -statistics

We evaluate the statistical utility of the proposed release mechanism following Theorem III.2. We compare the ϵ -differentially private χ^2 -statistic to the original statistic via KL divergence. We generated 3×2 contingency tables with positive margins and $N/2$ cases and $N/2$ controls assuming a product-multinomial distribution with the following frequencies:

$$\begin{aligned}
 (a) \begin{bmatrix} 0.72 & 0.20 \\ 0.18 & 0.28 \\ 0.10 & 0.52 \end{bmatrix}, & \quad (b) \begin{bmatrix} 0.60 & 0.23 \\ 0.21 & 0.30 \\ 0.19 & 0.47 \end{bmatrix}, \\
 (c) \begin{bmatrix} 0.47 & 0.25 \\ 0.45 & 0.51 \\ 0.08 & 0.24 \end{bmatrix}, & \quad (d) \begin{bmatrix} 0.65 & 0.46 \\ 0.29 & 0.43 \\ 0.06 & 0.11 \end{bmatrix}.
 \end{aligned} \tag{2}$$

For the χ^2 -distribution with 2 degrees of freedom, an observed value of 6 corresponds to a p -value of $\exp(-3) \approx 0.05$. The preceding frequency tables correspond to contingency tables for which we expect a p -value of 0.05 for

$$(a) N = 20, \quad (b) N = 40, \quad (c) N = 80, \quad (d) N = 160.$$

For example, for $N = 200$ individuals and underlying frequency table (a) we expect a table of the form

$$\begin{bmatrix} 72 & 20 \\ 18 & 28 \\ 10 & 52 \end{bmatrix},$$

which has a χ^2 -statistic of 60. Therefore, for $N = 20$ we expect a χ^2 -statistic of 6. If we fix the number of individuals N , then the χ^2 -statistic corresponding to frequency table (a) is the largest, namely 8 times the χ^2 -statistic corresponding to frequency table (d).

The choice of the frequency tables in (2) is motivated by the GWAS on the hair length of dogs in [9] and our simulations using HAP-SAMPLE. The χ^2 -statistic resulting from the frequency table (a) is comparable to the χ^2 -statistic of the SNP most associated to the hair length in dogs (on chromosome 32 at position 7,100,913 in the CanMap data set). The χ^2 -statistic resulting from the frequency table (c) is comparable to the χ^2 -statistic of a causative SNP in a simulated association study under the additive model (i.e., main effects only model) for $MAF = 0.4$, and (d) is comparable to a causative SNP under the additive model for $MAF = 0.25$. The frequency table (b) corresponds to an intermediate model for a causative SNP with high MAF and was added for consistency.

For a fixed total number of individuals N , we generated 10,000 tables from the frequency tables in (2) and computed the corresponding χ^2 -statistics. We also generated 10,000 private χ^2 -statistics according to the release mechanism described following Theorem III.2. In Figure 1 we plotted

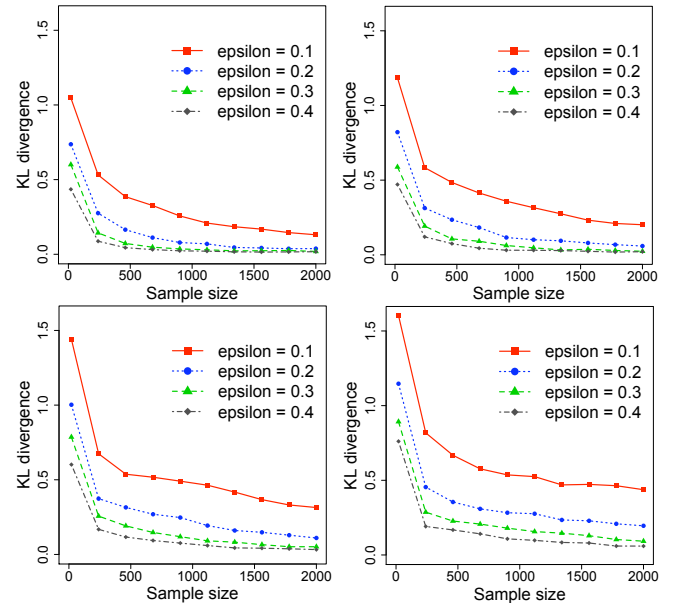


Figure 1. KL divergence between the original χ^2 -statistic and the private χ^2 -statistic based on the frequency table (a) top left, (b) top right, (c) bottom left, and (d) bottom right.

the KL divergence between the original and the private χ^2 -statistics for increasing N and for four different levels of privacy. The four plots correspond to the four frequency tables in (2). We see that the KL divergence depends on the χ^2 -statistic of the underlying frequency table, the total number of individuals N , and the privacy level ϵ . Since the added noise is asymptotically $Laplace(0, 4)$ distributed, the larger the original χ^2 -statistic, the smaller the KL divergence is. Similarly, a larger number of individuals N leads to a larger χ^2 -statistic and hence to a smaller KL divergence. The scale of the Laplace noise is inverse proportional to the privacy parameter ϵ . Therefore, the smaller ϵ , the larger the KL divergence is. These simulations demonstrate that it is possible to release ϵ -differentially private χ^2 -statistics and maintain good statistical utility in a realistic GWAS setting.

B. Differentially private p -values

We did a similar analysis on the p -values following the proposed release mechanism of adding Laplace noise according to Theorem III.3. Based on the frequency tables in (2), we computed the KL divergence between the original and private p -values for increasing N and for four different privacy levels. The resulting plots are shown in Figure 2. Similarly to the χ^2 -statistics, the smaller ϵ , the larger the KL divergence is. However, the relation between the KL divergence and the number of individuals, resp. the original χ^2 -statistic, is reversed since, for the χ^2 -distribution with 2 degrees of freedom, the χ^2 -statistic is proportional to the logarithm of the p -value. The larger the χ^2 -statistic, the smaller the p -value and hence the smaller the signal

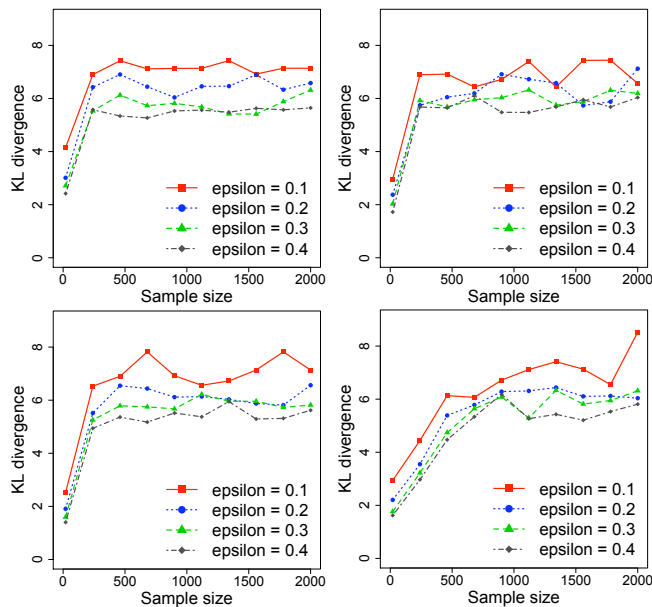


Figure 2. KL divergence between the original p -values and the private p -values based on the frequency table (a) top left, (b) top right, (c) bottom left, and (d) bottom right.

to noise ratio. The jumps in the figures arise because we project the perturbed p -values which fall outside the interval $[0, 1]$ to 0 or 1, respectively. Although there is a one-to-one correspondence between the χ^2 -statistics and the p -values, the χ^2 -statistics have a much smaller KL divergence and are therefore better suited for privacy purposes.

Projecting the p -values onto a region of interest as described in Corollary III.4 results in plots similar to those in Figure 2; the plots depend on how much smaller the p -value under consideration is compared to 1 in the case of Theorem III.3 and p^* in the case of Corollary III.4.

C. Releasing the M most relevant SNPs with respect to a specific phenotype

Practitioners are often interested in finding and releasing the most relevant (i.e., most statistically and practically significant) SNPs. Here we analyze what sample size N is needed in order to recover the two causative SNPs in the HAP-SAMPLE simulations from the private χ^2 -statistics. We chose $M = 3$ and plotted the frequencies (based on 1,000 private χ^2 -statistics) for which one or both of the two causative SNPs were among the three highest ranked private χ^2 -statistics computed according to Algorithm 1. We performed this analysis for increasing sample size N and for four different privacy levels. We used the simulated HAP-SAMPLE data consisting of around 10,000 SNPs total with two causative SNPs under the additive model with MAF=0.25 and MAF=0.4. The resulting bar charts are shown in Figure 3.

As we expect, a larger value of ϵ (i.e., less noise/less privacy) results in a higher chance of releasing the truly causative SNPs. We also observe that the smaller the MAF, the more data we need to detect the causative SNPs at a fixed level of ϵ . For example, for $\epsilon = 0.4$, Figure 3 shows that for MAF=0.4 we need about 7,500 individuals to detect the causative SNPs whereas for MAF=0.25 we need about 10,000 individuals. A smaller MAF corresponds to a sparser table, and we are in a similar situation to that described in [12], who show that for sparse tables differential privacy requires adding a lot of noise, often with the result of impairing statistical inference. Our results support the traditional trade-off: in order to detect important effects, we need to either relax the privacy constraint or increase the total number of individuals massively.

An alternative to adding noise to the data we want to release, is to add noise to the analysis itself. We explain this approach for GWAS in the following section.

V. EXTENDED WORK: DIFFERENTIALLY PRIVATE ALGORITHM FOR DETECTING EPISTASIS

As we just saw, the sparseness of GWAS data requires an unrealistically large number of individuals in each study or a relaxation of the privacy level. In order to deal with sparseness, methods have been proposed, where the Laplace

noise is added to the analysis directly instead of to the output. Another advantage of such an approach is that it allows the analysis of models that integrate information across SNPs. Here we present an ϵ -differentially logistic regression approach that is directly applicable to GWAS.

Most methods for detecting epistasis are based on a two-stage approach. First, all SNPs are filtered e.g. using χ^2 -statistics or p -values, to reduce the potential interacting SNPs to a small number. The loci achieving some threshold are then further examined for interactions. A widely used test for detecting gene-gene interactions on a small number of SNPs is a penalized logistic regression, e.g. the L_2 -regularized logistic regression proposed by Park and Hastie [4]. By adapting the work of Bhaskar et al. [5] and Chaudhuri et al. [6], we derive a privacy-preserving method for detecting epistasis, where both stages in the two-stage

approach satisfy differential privacy.

We use the first two steps in Algorithm 1 to chose a subset of interesting SNPs of size M in a differentially private way. Park and Hastie [4] suggest an L_2 -regularized logistic regression in order to detect epistasis within a small subset of SNPs. Chaudhuri et al. [6] demonstrated how to perturb the objective function for privacy-preserving machine-learning algorithm designs if the loss function and the regularizer satisfy certain convexity and differentiability criteria. In the following, we outline how to apply their objective perturbation in order to find a differentially private algorithm for detecting epistasis.

Let $y = (y_1, \dots, y_N)$ denote the disease status of the N individuals. Note that in this section we encode the diseased status by 1 and the non-diseased status by -1. Let $x_i \in \mathbb{R}^{p+1}$ denote the feature vector for the i^{th} individual. The first entry corresponds to the intercept. The encoding of the features is explained via an example. We will look at a model with two SNPs including their interaction. SNP1 takes the three states 0, 1, and 2, which are encoded by 100, 010, and 001. Similarly for SNP2. The interaction term SNP1 \times SNP2 takes the states 00, 01, 02, 10, 11, 12, 20, 21, 22 and is encoded by 100000000, 010000000, \dots , 000000001. So an individual with genotype 12, who is not diseased would have

$$x = (1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0), \quad y = -1.$$

Let $K - 1$ be the total number of effects in the model (including main and higher-order effects). It is important to note that $\|x_i\|_2 \leq K$.

The objective function described in [4] is

$$L(\beta) = \sum_{i=1}^N \log(1 + \exp(-y_i \beta^T x_i)) + \frac{1}{2} \beta^T \Lambda \beta,$$

where Λ is of the form $(0, \lambda, \dots, \lambda)$, i.e. β_0 is not penalized. They use the Newton-Raphson method for the optimization and forward selection and backward deletion steps based on an AIC or BIC score to select model size and important factors.

We can apply the approach of Chaudhuri et al. [6] to perturb the objective function such that the algorithm satisfies ϵ -differential privacy. We are interested in the following perturbed objective function:

$$L_{\text{priv}}(\beta) = \sum_{i=1}^N \log(1 + \exp(-y_i \beta^T x_i)) + \frac{1}{2} \beta^T \Lambda \beta + \frac{1}{N} b^T \beta,$$

where b is noise drawn from a distribution with density

$$f(b) = \frac{1}{\alpha} \exp(-k \|b\|_2)$$

and k is a constant and α the normalizing constant.

As proposed by Park and Hastie [4] we use forward selection and backward deletion steps based on an AIC or BIC score to select model size. However, we replace the optimization step in their method by the following algorithm:

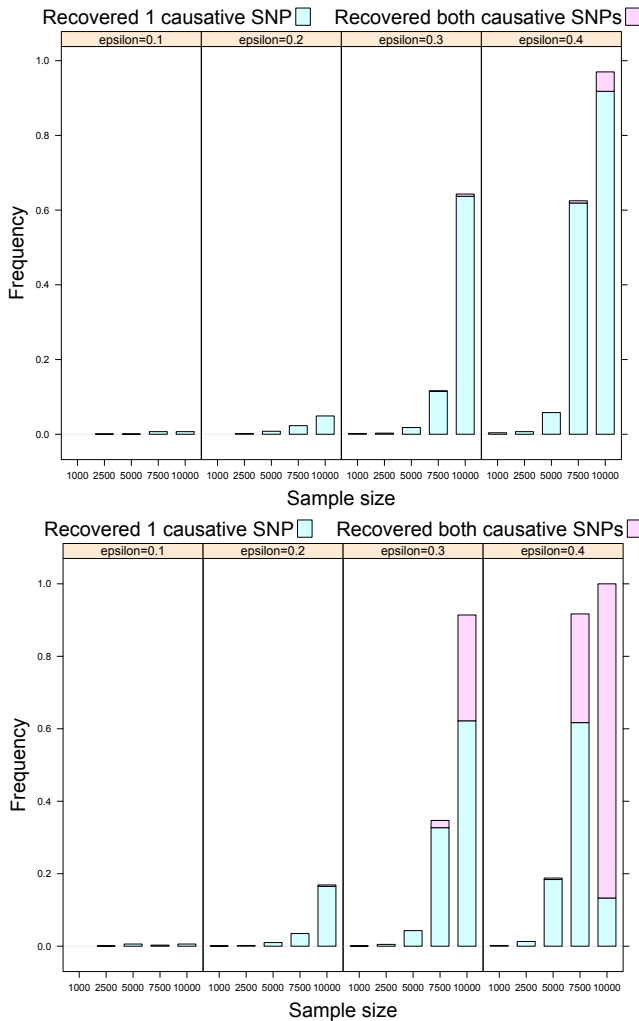


Figure 3. Bar charts representing the frequencies for which one or both of the two causative SNPs were among the three highest ranked private χ^2 -statistics under the additive model with MAF=0.25 (top) and MAF=0.4 (bottom).

Algorithm 2 The ϵ -Differentially Private Algorithm for Detecting Epistasis

Input: The data vectors x_i, y_i , where $i = 1, \dots, N$ and parameters ϵ , λ , and c .

Output: The output consists of the noisy effects.

1. Let $\epsilon' = \epsilon - \log(1 + \frac{2cK}{N\lambda} + \frac{c^2K^2}{N^2\lambda^2})$, If $\epsilon' > 0$, then $\delta = 0$, else $\delta = \frac{cK}{N(e^{\epsilon'/4} - 1)} - \lambda$ and $\epsilon' = \epsilon/2K$.
 2. Draw b from a distribution with density $f(b) = \frac{1}{\alpha} \exp(-\frac{\epsilon\|b\|_2}{2})$.
 3. Compute $\beta_{\text{priv}} = \text{argmin}(L_{\text{priv}}(\beta) + \frac{1}{2}\delta\|\beta\|_2)$.
-

Theorem V.1. Algorithm 2 is ϵ -differentially private.

Proof: The proof follows from Theorem 9 in [6], and taking into account that $\|x_i\|_2 \leq K$ for our application. ■

Thus we can move away from a SNP-by-SNP analysis to an integrated approach without relaxing privacy. Applying this method to actual GWAS data is part of ongoing work.

VI. CONCLUSION

In this paper, we have demonstrated that it is possible, using the formal privacy guarantees of differential privacy, for NIH or other data owners to release at least some genetic data required by practitioners. More specifically, we described a privacy-preserving release of aggregate minor allele frequencies and the release of differentially-private χ^2 -statistics and p -values. We also provided a differentially private algorithm for releasing these statistics for the most relevant SNPs.

Our simulations, however, indicate that for bigger and sparse data the release of simple summary statistics is problematic and not sufficient from both privacy and utility perspectives. The release of summary statistics is in part sufficient for traditional piece-wise analysis, but more complex methodology is needed to deal with more sparse data and models that integrate across SNPs to detect epistasis. To address this problem, we outlined an ϵ -differentially private algorithm for a specific form of penalized logistic regression. This is but one of the newer methods being introduced into the statistical literature for GWAS, but we expect that the general strategy suggested here might be adaptable for other statistical methods, e.g., for sparse partitioning [13].

Since the introduction of differential privacy by [3], and in particular ϵ -differential privacy, many additional variations along with their considerations with respect to statistical analysis have been proposed (e.g., more recently [14]). To further improve the privacy-utility tradeoffs for GWAS, the future research would consider such alternate mechanisms.

ACKNOWLEDGMENT

This research was supported in part by NSF Grants BCS-0941553 and BCS-0941518 to Pennsylvania State University and Carnegie Mellon University, respectively.

REFERENCES

- [1] N. Homer, S. Szelling, 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, 2008.
- [2] R. Braun, W. Rowe, C. Schaefer, J. Zhan, and K. Buetow, "Needles in the haystack: Identifying individuals present in pooled genomic data," *PLoS Genetics*, vol. 5, 2009.
- [3] C. Dwork, F. McSherry, M. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *Theory of Cryptography Conference*, pp. 265–284, 2006.
- [4] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, pp. 30–50, 2008.
- [5] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," *16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.
- [6] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, pp. 1069–1109, 2011.
- [7] D. Vu and A. Slavkovic, "Differential privacy for clinical trial data: Preliminary evaluations," in *IEEE International Conference Data Mining Workshop*, 2009, pp. 138–143.
- [8] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press, 1975.
- [9] E. Cadieu, M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M. Vonholdt, A. Rhue, A. Boyko, A. Byers, A. Wong, D. S. Mosher, A. G. Elkahoulou, T. C. Spady, C. Andre, K. G. Lark, M. Cargill, C. D. Bustamante, R. K. Wayne, and E. A. Ostrander, "Coat variation in the domestic dog is governed by variants in three genes." *Science*, vol. 326, no. 5949, pp. 150–153, 2009.
- [10] A. Malaspinas and C. Uhler, "Detecting epistasis via Markov bases," *Journal of Algebraic Statistics*, vol. 2, pp. 36–53, 2011.
- [11] F. A. Wright, H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, F. Pardo-Manuel de Villena, P. F. Sullivan, K. C. Wilhelmsen, and F. Zou, "Simulating association studies: a data-based resampling method for candidate regions or whole genome scans," *Bioinformatics*, vol. 23, pp. 2581–2588, 2007.
- [12] S. Fienberg, A. Rinaldo, and X. Yang, "Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables," in *Proceedings of the 2010 conference on Privacy in Statistical Databases*. Springer-Verlag, 2010, pp. 187–199.
- [13] D. Speed and S. Tavaré, "Sparse partitioning: Nonlinear regression with binary or tertiary predictors, with application to association studies," *The Annals of Applied Statistics*, vol. 5, no. 2A, pp. 873–893, 2011.
- [14] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," 12 2010. [Online]. Available: <http://arxiv.org/abs/1012.4763v1>