Journal of Privacy and Confidentiality (2013) **5**, Number 1, 137–166

Privacy-Preserving Data Sharing for Genome-Wide Association Studies

Caroline Uhler*, Aleksandra Slavković†, and Stephen E. Fienberg‡

1 Introduction

Genome-wide association studies (GWAS) focus on finding genetic variations associated with traits such as major diseases often by measuring associations between single-nucleotide polymorphisms (SNPs), i.e., DNA sequence variations at single nucleotides, and a particular disease. A typical study compares the DNA of individuals with the disease (cases) and similar individuals without (controls). For a specific trait, the output of such studies often consists of the χ^2-statistics or the p-values for the most significant SNPs including their minor allele frequencies (i.e., the lowest allele frequency observed for the cases and the controls).

In an article that shocked the genetics community, Homer et al. [13] claimed that, under certain conditions, they could use statistical methods to "accurately and robustly [resolve]" 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; they then used a t-test to assess if the individual was part of the test population.

Although proposed specifically for use in a forensic context and only secondarily for breaking privacy, the Homer et al. [13] "attack" appeared to be generally applicable. As a reference population one might use the publicly available SNP data from the HapMap project[1] which consists of SNP data from four populations varying in size from 45 to 90 individuals. 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 GWAS, which contain both genotype data and disease status. Before the appearance of the article [13], the averaged MAFs of the cases and the averaged MAFs of the controls in a GWAS were typically publicly available.

In response to Homer et al. [13], Braun et al. [3] showed that their proposed test depends heavily on the assumption that the genotypes of the test population, the reference population, and the specific person under consideration are samples from the same

*Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria, mailto:caroline.uhler@ist.ac.

†Department of Statistics, Department of Public Health Sciences, Penn State University, University Park, PA 16802 USA, mailto:sesa@psu.edu.

‡Department of Statistics, Machine Learning Department, Cylab, and Living Analytics Research Centre, Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA, mailto:fienberg@stat.cmu.edu.

[1] `http://hapmap.ncbi.nlm.nih.gov/`.

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. [13] attack lead to a high false-positive rate, see e.g., Braun et al. [3]. Other critiques of Homer et al. suggested alternative formulations of the identification problem, claimed to strengthen the attack, or suggested different ways to protect the data, e.g., see [6, 14, 15, 16, 18, 19, 21, 23, 27]. Despite the apparent limitations of the Homer et al. [13] attack on the privacy of GWAS participants and the controversial and, we believe, exaggerated nature of their statistical claims, NIH immediately removed from open-access databases all aggregate results such as values of averaged MAFs over cases and controls, chi-square (χ^2)-statistics, and p-values (see Couzin [7] and Zerhouni and Nabel [26]). The NIH policy remains in effect today.[2] Every researcher who wants to gain access to any of these data sets needs to go through an elaborate approval process. This is a particularly difficult obstacle for computer scientists, mathematicians, or statisticians who do not have a credible record in GWAS research.

Here we propose methods which allow for the release of aggregate GWAS data without compromising an individual's privacy, and in many ways totally bystep the debate on the validity of the claims by Homer et al. [13] and others on the vulnerability of GWAS databases. Our GWAS privacy guarantees utilize the concept of *differential privacy*, recently introduced by the cryptographic community (e.g., Dwork et al. [10]). 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 [20] use a form of penalized logistic regression to test for detecting gene-gene interactions on a small number of SNPs. By adapting the work of [1] and [5] to this methodology, we derive a privacy-preserving method for GWAS, where both stages in the two-stage approach satisfy ϵ-differential privacy.

[2] `http://gwas.nih.gov/`.

Section 2 describes the basic problem and relevant definitions. In Section 3, we present methods for releasing ϵ-differentially private MAFs, χ^2-statistics, and p-values, and in Section 4 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 5, we propose a differentially-private method for finding genome-wide associations based on a penalized approach to logistic regression.

2 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 sufficient power to detect SNPs which are associated with 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 assume throughout the paper that all margins of such 3×2 contingency tables 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. We measure association between a disease and a SNP by the χ^2-statistic. For a 3×2 table t with counts t_{ij}, row sums s_i and column sums $N/2$ the χ^2-statistic is

$$\chi^2(t) = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(2t_{ij} - s_i)^2}{2s_i}$$

and the corresponding p-value under the χ^2-distribution with 2 degrees of freedom is

$$\exp(-\frac{\chi^2(t)}{2}). \tag{1}$$

Definition 2.1. 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 measurable subset $S \subset \mathbb{R}$,

$$\frac{\Pr(\mathcal{K}(D) \in S)}{\Pr(\mathcal{K}(D') \in S)} \leq e^{\epsilon}.$$

Definition 2.2. The *sensitivity* of a function $f : \mathcal{D}^N \to \mathbb{R}^d$, 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 [10]). This type of release mechanism is often referred to as the *Laplace mechanism*.

Definition 2.3. The Kullback-Leibler (KL) divergence between two probability distributions f and g is defined by

$$D_{KL}(f||g) = \int_{-\infty}^{\infty} f(x)\, log \frac{f(x)}{g(x)} dx. \tag{2}$$

For the analysis of the simulation results in Section 3 we use the KL divergence to measure the difference between two distributions such as the original χ^2-statistic and its corresponding ϵ-differentially private version.

3 Privacy-Preserving Methodology

In this section we compute the sensitivity of MAFs, χ^2-statistics, and p-values needed to release the private versions of these statistics for each SNP via the Laplace mechanism. We also describe an ϵ-differentially private algorithm for the release of the latter two quantities for the M most relevant SNPs.

3.1 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 1. 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 3.1. *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

Table 1: Table showing the averaged MAFs of the cases and the controls for M SNPs.

MAF	**SNP 1**	**SNP 2**	$\cdots$	**SNP M**
Cases	0.29	0.20	$\cdots$	0.11
Controls	0.27	0.31	$\cdots$	0.10

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, since $a_i, a'_i \in \{0, 1, 2\}$ we get that $|a_j - a'_j| \leq 2$. 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 the M-dimensional vector where all entries are $\frac{2}{N}$. □

Lemma 3.1 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 1 yields ϵ-differential privacy. This result can be seen as a special case of Example 3 in [10] 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.

3.2 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 [1], for release of frequent itemsets, to release the noisy statistics corresponding to the most relevant SNPs.

Theorem 3.2. *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. Consider the following 3×2 contingency table with positive margins and $N/2$ cases and controls each:

		Disease Status	
		0	1
No. Individuals	0	a	m-a
With Genotype	1	b	n-b
	2	N/2-a-b	N/2-m-n+a+b
Total		N/2	N/2

with $a, b \geq 0$, $m, n > 0$, $a \leq m$, $b \leq n$, $a + b \leq N/2$, and $m + n < N$. Let

$$\mathcal{D} \;=\; \{(a, b, m, n) \in \mathbb{N} \mid m > 0,\ n > 0,\ a \leq m,\ b \leq n, a + b \leq N/2,\ m + n < N\}. \quad (3)$$

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 $d = (-1, 1, -1, 1)$, which normalized (to have length 1) becomes $(-1/2, 1/2, -1/2, 1/2)$. First note that

$$\chi^2(a, b, m, n) \quad = \quad \frac{(2a - m)^2}{m} + \frac{(2b - n)^2}{n} + \frac{(2a - m + 2b - n)^2}{N - m - n}. \tag{4}$$

We then compute the directional derivative of $\chi^2(a, b, m, n)$ in direction $d = (-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, \frac{1}{2})$, which 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 for releasing the χ^2-statistic for a single SNP, we need to add Laplace noise with scale $\frac{1}{\epsilon}\frac{4N}{N+2}$ to the true χ^2-statistic. Thus for increasing N, the perturbed (private) χ^2-statistics get more accurate.

Before we consider the sensitivity of the p-values, we derive the asymptotic distribution of the perturbed χ^2-statistic which is a convolution of its (asymptotic) sampling distribution and perturbation.

Theorem 3.3. *Let a χ^2 test statistic T have the χ^2 sampling distribution with 2 degrees of freedom and let the perturbation $Y \sim Laplace(0, 4/\epsilon)$. Then, the distribution of the perturbed χ^2 test statistic, $X = T + Y$, has the following probability density function:*

$$f_X(x) = \begin{cases} \frac{\epsilon}{4}\frac{1}{\epsilon+2}\exp\left(\frac{\epsilon x}{4}\right) & \text{if } x < 0 \\ \\ \frac{\epsilon}{4}\left[\left(\frac{1}{\epsilon-2} + \frac{1}{\epsilon+2}\right)\exp\left(-\frac{x}{2}\right) - \frac{1}{\epsilon-2}\exp\left(-\frac{\epsilon x}{4}\right)\right] & \text{if } x \geq 0 \end{cases},$$

and the following cumulative distribution function

$$F_X(x) = \begin{cases} \frac{1}{\epsilon+2} \exp\left(\frac{\epsilon x}{4}\right) & \textit{if } x < 0 \\ \\ 1 - \frac{\epsilon}{2}\left(\frac{1}{\epsilon-2} + \frac{1}{\epsilon+2}\right)\exp\left(-\frac{x}{2}\right) + \frac{1}{\epsilon-2}\exp\left(-\frac{\epsilon x}{4}\right) & \textit{if } x \geq 0 \end{cases}.$$

Proof. Since T and Y are independent random variables, the distribution of X is the convolution of the given χ^2 and *Laplace* distributions:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_Y(x-t) f_T(t)\, dt \\ &= \frac{1}{2}\int_0^{\infty} f_Y(x-t)\ \exp(-\frac{t}{2})\, dt \\ &= \begin{cases} \frac{\epsilon}{16}\int_0^\infty \exp\left(\frac{(x-t)\epsilon}{4} - \frac{t}{2}\right)\ dt & \text{if } x < 0 \\ \frac{\epsilon}{16}\left[\int_0^x \exp\left(\frac{(t-x)\epsilon}{4} - \frac{t}{2}\right)\ dt + \int_x^\infty \exp\left(\frac{(x-t)\epsilon}{4} - \frac{t}{2}\right)\ dt\right] & \text{if } x \geq 0 \end{cases} \\ &= \begin{cases} \frac{\epsilon}{16}\exp\left(\frac{x\epsilon}{4}\right)\int_0^\infty \exp\left(-\frac{\epsilon+2}{4}\,t\right)\ dt & \text{if } x < 0 \\ \\ \frac{\epsilon}{16}\left[\exp\left(-\frac{x\epsilon}{4}\right)\int_0^x \exp\left(\frac{\epsilon-2}{4}\,t\right)\ dt + \exp\left(\frac{x\epsilon}{4}\right)\int_x^\infty \exp\left(-\frac{\epsilon+2}{4}\,t\right)\ dt\right] & \text{if } x \geq 0 \end{cases} \\ &= \begin{cases} \frac{\epsilon}{4}\frac{1}{\epsilon+2}\exp\left(\frac{\epsilon x}{4}\right) & \text{if } x < 0 \\ \\ \frac{\epsilon}{4}\left[\left(\frac{1}{\epsilon-2} + \frac{1}{\epsilon+2}\right)\exp\left(-\frac{x}{2}\right) - \frac{1}{\epsilon-2}\exp\left(-\frac{\epsilon x}{4}\right)\right] & \text{if } x \geq 0. \end{cases} \end{aligned}$$

The cumulative distribution function F_X can easily be computed by integrating f_X. □

We show through simulations in Section 4 that the finite sample distribution is well-approximated by this asymptotic distribution even for tables with low total count, marginal counts, or individual counts. This is in contrast to the poor finite sample behavior of the χ^2 test statistics arising when the noise is added directly to the underlying cell counts (see Section 4); the latter mechanism has been considered by many (e.g., [10, 11]). 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 [24].

We now prove that the asymptotic distribution of the perturbed χ^2-statistic arising from perturbing the cell counts is the same as for the unperturbed χ^2-statistic, namely a χ^2-distribution with two degrees of freedom.

Theorem 3.4. *Let $X^{(n)}$ denote a 6-dimensional random variable corresponding to the entries of a 3×2 contingency table based on n individuals. Let Y denote a 6-dimensional random variable drawn from Laplace$(0, \frac{2}{\epsilon})$. Then the perturbed χ^2-statistic arising from perturbed cell counts $(X^{(n)} + Y)$ asymptotically has a χ^2-distribution with two degrees of freedom.*

Proof. Let $p_0, p_1, p_2, q_0, q_1 \in [0,1]$ such that $p_0 + p_1 + p_2 = 1$ and $q_0 + q_1 = 1$. Under the null hypothesis of independence on a 3×2 contingency table, the data are sampled from a multinomial distribution with probability vector $\hat{p} = (p_0q_0, p_0q_1, p_1q_0, p_1q_1, p_2q_0, p_2q_1)^T$. The central limit theorem implies that

$$\sqrt{n}\left(\frac{X^{(n)}}{n} - \hat{p}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right),$$

where Σ is the covariance matrix of the product multinomial, i.e.,

$$\Sigma = \Gamma - \hat{p}\hat{p}^T$$

and $\Gamma = \operatorname{diag}(\hat{p})$. Note that Σ has rank 2 and therefore also $\Gamma^{-\frac{1}{2}}\Sigma\Gamma^{-\frac{1}{2}}$. Let $Y \sim$ Laplace$(0, \frac{2}{\epsilon})$. Slutsky's theorem implies that

$$\sqrt{n}\left(\frac{X^{(n)} + Y}{n} - \hat{p}\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right),$$

and therefore that

$$\sqrt{n}\,\Gamma^{-\frac{1}{2}}\left(\frac{X^{(n)} + Y}{n} - \hat{p}\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-\frac{1}{2}}\Sigma\Gamma^{-\frac{1}{2}}\right).$$

Finally, by invoking the continuous mapping theorem, we prove the claim, namely

$$\chi^2_{\text{perturbed}} = n\left(\frac{X^{(n)} + Y}{n} - \hat{p}\right)^T \Gamma^{-1}\left(\frac{X^{(n)} + Y}{n} - \hat{p}\right) \xrightarrow{d} \chi^2_2.$$

□

This theorem establishes that for a fixed $\epsilon \in (0,1)$ the asymptotic distribution of the perturbed χ^2-statistic arising from perturbing the cell counts is independent of ϵ. However, the convergence rate does depend on ϵ. More precisely, it depends on the convergence rate of

$$\frac{Y}{\sqrt{n}} \xrightarrow{d} 0.$$

In particular, when considering the situation where ϵ varies as a function of n (we denote this by $\epsilon(n)$), we require that $\epsilon(n)^{-1} = o(n)$ for Theorem 3.4 to hold.

Given the distributions derived in Theorem 3.3 and Theorem 3.4, the researcher can now compute the p-values for the test of independence using the perturbed χ^2-statistics (when perturbing the test statistic itself or when adding noise at the level of the cell counts).

We also consider releasing differentially private p-values (without perturbing the counts or the related statistic first). We perform a similar sensitivity analysis on the p-values corresponding to the χ^2-statistics when assuming a χ^2-distribution with 2 degrees of freedom as null distribution, cf. [2].

Theorem 3.5. *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 a 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

$$\exp(-\frac{x}{2}), \qquad x \geq 0.$$

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

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

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 3.6. *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 3.2 and Theorem 3.5. We here give an overview. The contingency table

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

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

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

This new contingency table has χ^2-statistic

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

□

In GWAS settings, however, 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. [1] show in the context of frequent pattern recognition how to release the most significant patterns together with their frequencies while satisfying differential privacy. We adapt their method by incorporating our results from Theorem 3.2 and Theorem 3.5 to GWAS, and state the main result of this section: Algorithm 1 for releasing the private χ^2-statistics (resp. p-values) of the M most relevant SNPs.

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 Algorithm 1, however, the Laplace noise only scales with the number of statistics M actually released.

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

```
Input: The $\chi^2$-statistics (resp. $p$-values) for all $M'$ SNPs and the number of statistics, $M$, we want to release.
Output: The $M$ noisy $\chi^2$-statistics (resp. $p$-values).
  1. Add Laplace noise with mean zero and scale $\frac{4M}{\epsilon}\frac{4N}{N+2}$ to the $\chi^2$-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 $\chi^2$-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 $\chi^2$-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 3.7. *Algorithm 1 is ϵ-differentially private.*

Proof. Using the sensitivities computed in Theorem 3.2 and Theorem 3.5, the proof follows immediately from Theorem 5 in [1]. □

4 Evaluation of Methodology and 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 [4] and further been analyzed in [17]. 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. [4] 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 also use the simulations from [17] performed using HAP-SAMPLE [25]. 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 [17] 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., [10], [24]). Instead, we focus on the release of differentially-private χ^2-statistics, p-values, and the most relevant SNPs.

4.1 Asymptotic distribution of the perturbed χ^2-statistic

We first present results on the asymptotic distribution of the perturbed χ^2-statistic arising from adding noise directly to the statistic, as derived in Theorem 3.3, and evaluate the accuracy of the asymptotic approximation. The distribution for $\epsilon = 0.2$ is described in Figure 1, and a comparison of three distributions, namely the asymptotic χ^2-distribution, the asymptotic Laplace distribution, and their convolution for different values of the privacy parameter ϵ are shown in Figure 2; we can observe that the asymptotic distribution of the perturbed χ^2-statistic is very similar to the underlying Laplace distribution as expected based on the convolution derived in Theorem 3.3.

Through simulations, we analyzed at which point the asymptotic approximation seems to be accurate for finite samples. It turns out that even for tables with very small cell counts or marginal counts, the finite sample distribution of the private χ^2-statistic is well-approximated by its asymptotic distribution, although it is well known that the exact distribution of the original χ^2-statistic is very poorly approximated by the χ^2-distribution for small samples. As an example we discuss the following 3 x 2 contingency table:

$$\begin{bmatrix} 1 & 3 \\ 8 & 12 \\ 41 & 35 \end{bmatrix}.$$

We ran a Markov chain on the set of contingency tables which have the same margins as the above table using tools from Algebraic Statistics, namely elements of a Markov basis as moves (e.g., see [8]). At each step (table), we computed the corresponding χ^2-statistic and added Laplace noise with scale $4/\epsilon$. The resulting posterior distribution is an approximation to the true distribution of the perturbed χ^2-statistic and corresponds

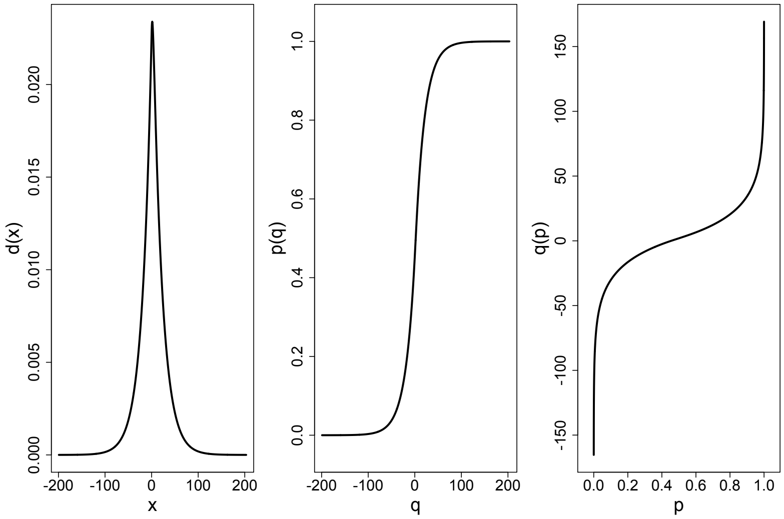


Figure 1: Asymptotic distribution of the perturbed χ^2 test statistic for $\epsilon = 0.2$: density function (left), cumulative distribution function (middle), and quantile function (right).

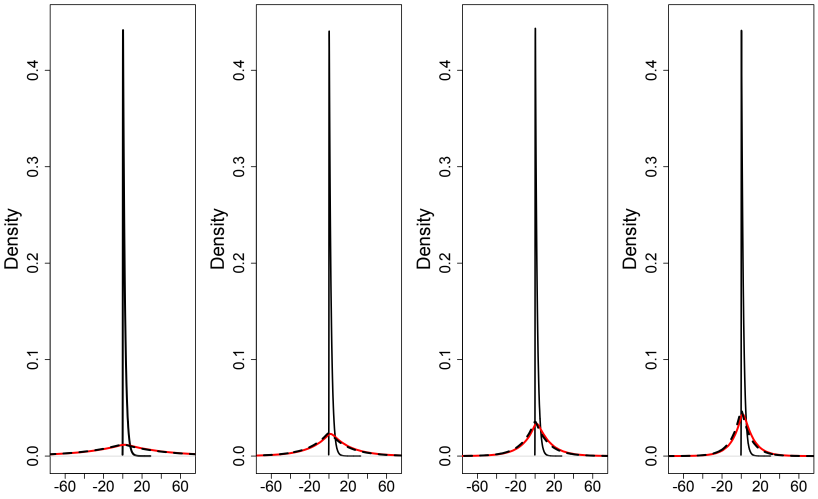


Figure 2: Comparison of the asymptotic sampling distribution (black line), perturbation (black dotted line) and its convolution (red line) for $\epsilon = 0.1$ (left), $\epsilon = 0.2$ (middle left), $\epsilon = 0.3$ (middle right), and $\epsilon = 0.4$ (right).

to the black dotted line in Figure 3. The asymptotic distribution of the perturbed χ^2-statistic derived in Theorem 3.3 is shown in red. These plots and additional simulations show that the asymptotic approximation is accurate even for tables with a low total count, marginal counts, or individual cell counts.

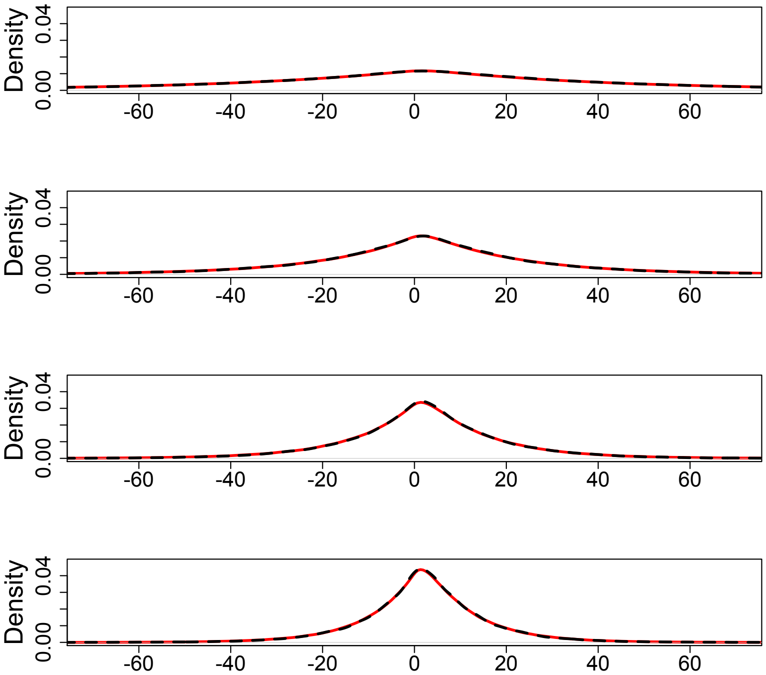


Figure 3: Asymptotic distribution of the perturbed χ^2-statistic (red line) and its true distribution (black dotted line) for $\epsilon = 0.1$ (top), $\epsilon = 0.2$ (second plot), $\epsilon = 0.3$ (third plot), and $\epsilon = 0.4$ (bottom).

Similarly, we now analyze under which conditions the asymptotic distribution of the perturbed χ^2-statistic arising from perturbing the cell counts, as shown in Theorem 3.4, appears to be accurate for finite samples. As we will see, when adding noise to the cell counts instead of the χ^2-statistic, the asymptotic distribution of the computed statistic is only accurate for a very large total cell count and large expected frequencies. We have performed extensive simulations and we here present three representative cases. We analyze the following three 3 x 2 contingency tables, one with a total cell count of 10,000 and two with a total cell count of 100,000 (one with large expected frequencies and one with a small expected frequency):

$$\begin{bmatrix} 1400 & 1600 \\ 1900 & 1300 \\ 1700 & 2100 \end{bmatrix}$$

(a) Table 1

$$\begin{bmatrix} 14000 & 16000 \\ 19000 & 13000 \\ 17000 & 21000 \end{bmatrix}$$

(b) Table 2

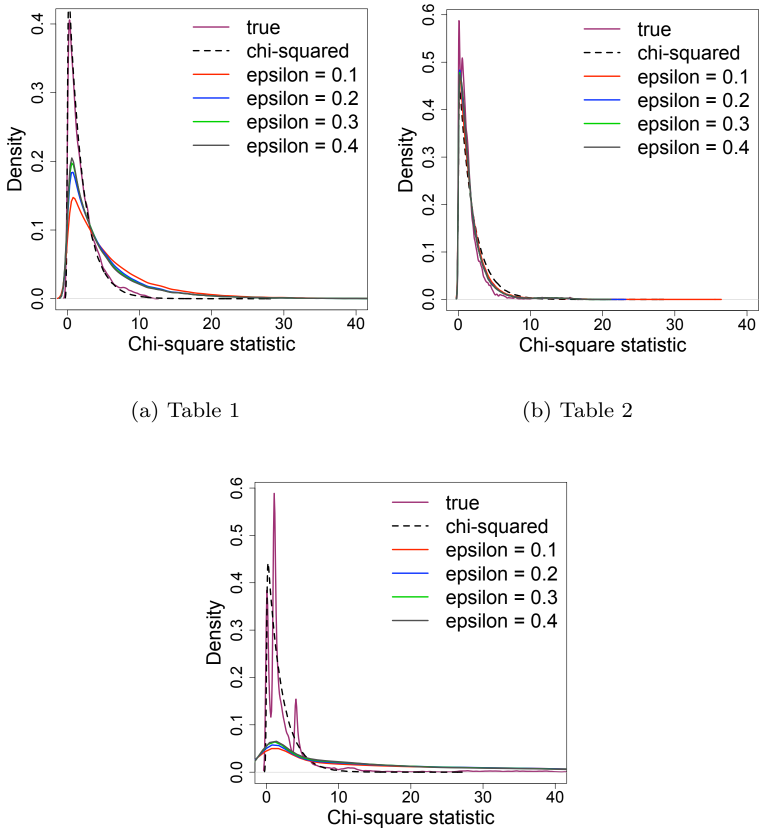
$$\begin{bmatrix} 1 & 3 \\ 26000 & 21000 \\ 23999 & 28997 \end{bmatrix}$$

(c) Table 3

We again run a Markov chain on the set of contingency tables which have the same margins as the above tables using a Markov basis to move between tables. At each step we perturbed the counts by adding Laplace noise with scale $2/\epsilon$ and computed the corresponding perturbed χ^2-statistic. The resulting posterior distribution is an approximation to the true distribution of the perturbed χ^2-statistic and is shown in Figure 4 for four values of the privacy parameter ϵ. Also, we show the true distribution of the unperturbed χ^2-statistic and the χ^2-distribution for comparison. Note that a total cell count of $10,000$ is not sufficient for a good approximation of the finite sample distribution by the asymptotic distribution. For a total cell count of $100,000$ the approximation appears to be accurate as long as the individual cell counts and margins are not too small, as is the case for Table 3.

4.2 Statistical Utility of Differentially-Private χ^2-Statistics and p-Values

In this section, we evaluate the statistical utility of the three proposed release mechanisms; first, perturbing the χ^2-statistics and releasing them; second, releasing the χ^2-statistics after perturbing the cell counts; third, perturbing the p-values and releasing them. Since the p-values corresponding to a certain χ^2-statistic in a 3×2 table can be computed by a bijection (see (1)), we do not analyze the statistical utility of the differentially-private p-values resulting from perturbing the cell counts or the χ^2-statistics. In these cases the statistical utility corresponds to the statistical utility of the perturbed χ^2-statistics.



(a) Table 1

(b) Table 2

(c) Table 3

Figure 4: Exact and asymptotic distribution of the unperturbed χ^2-statistic and perturbed χ^2-statistic with varied levels of ϵ for Tables 1, 2, and 3.

We base the comparison of the three approaches on 3×2 contingency tables with positive margins and $N/2$ cases and $N/2$ controls generated by assuming a product-multinomial distribution with the following frequencies:

$$\begin{array}{ll} (a)\begin{bmatrix} 0.72 & 0.20 \\ 0.18 & 0.28 \\ 0.10 & 0.52 \end{bmatrix}, & (b)\begin{bmatrix} 0.60 & 0.23 \\ 0.21 & 0.30 \\ 0.19 & 0.47 \end{bmatrix}, \\ (c)\begin{bmatrix} 0.47 & 0.25 \\ 0.45 & 0.51 \\ 0.08 & 0.24 \end{bmatrix}, & (d)\begin{bmatrix} 0.65 & 0.46 \\ 0.29 & 0.43 \\ 0.06 & 0.11 \end{bmatrix}. \end{array} \tag{5}$$

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

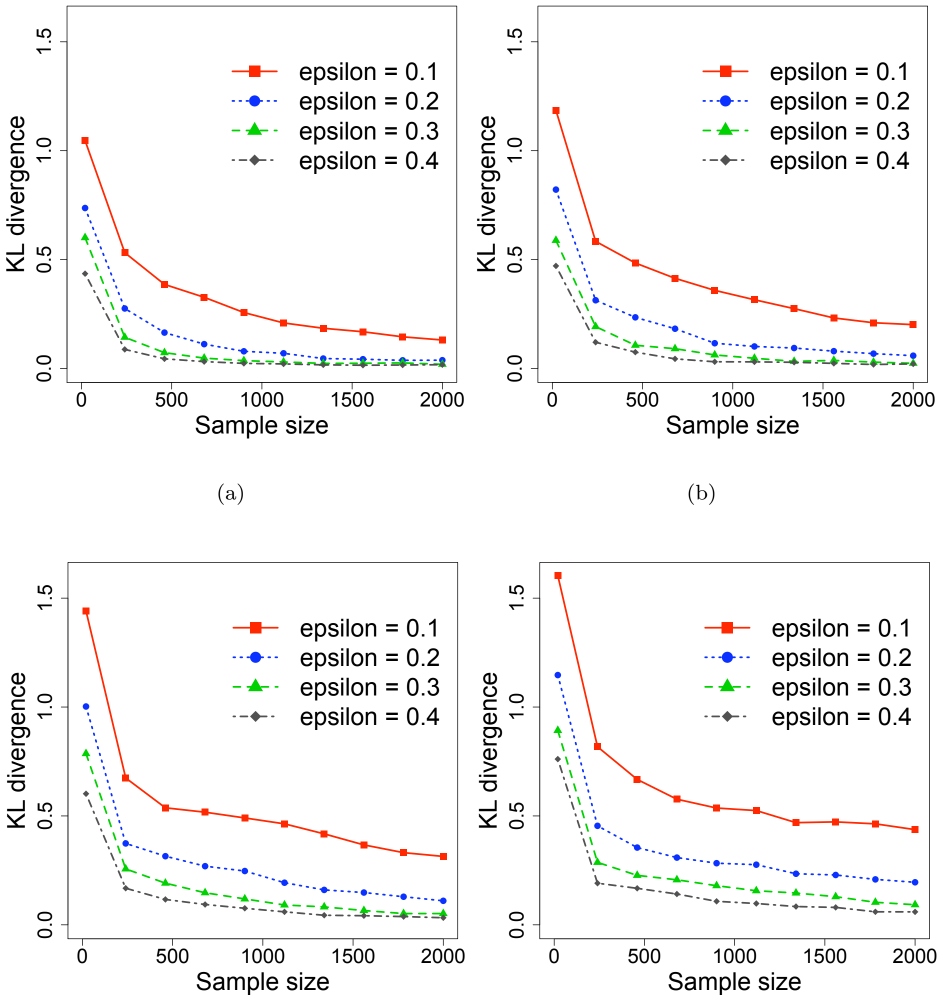
$$\begin{bmatrix} 72 & 20 \\ 18 & 28 \\ 10 & 52 \end{bmatrix},$$

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 (5) is motivated by the GWAS on the hair length of dogs in [4] 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.

4.2.1 Perturbing the χ^2-Statistics

We first compare the ϵ-differentially private χ^2-statistic, resulting from adding Laplace noise directly to the χ^2-statistic, to the original statistic via KL divergence. For a fixed total number of individuals N, we generated 10,000 tables from the frequency tables in (5) and computed the corresponding χ^2-statistics. We also generated 10,000 private χ^2-statistics according to the Laplace mechanism described following Theorem 3.2. In Figure 5 we plotted 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 (5). 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 (i.e., more noise/more privacy protection), 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.



(a)

(b)

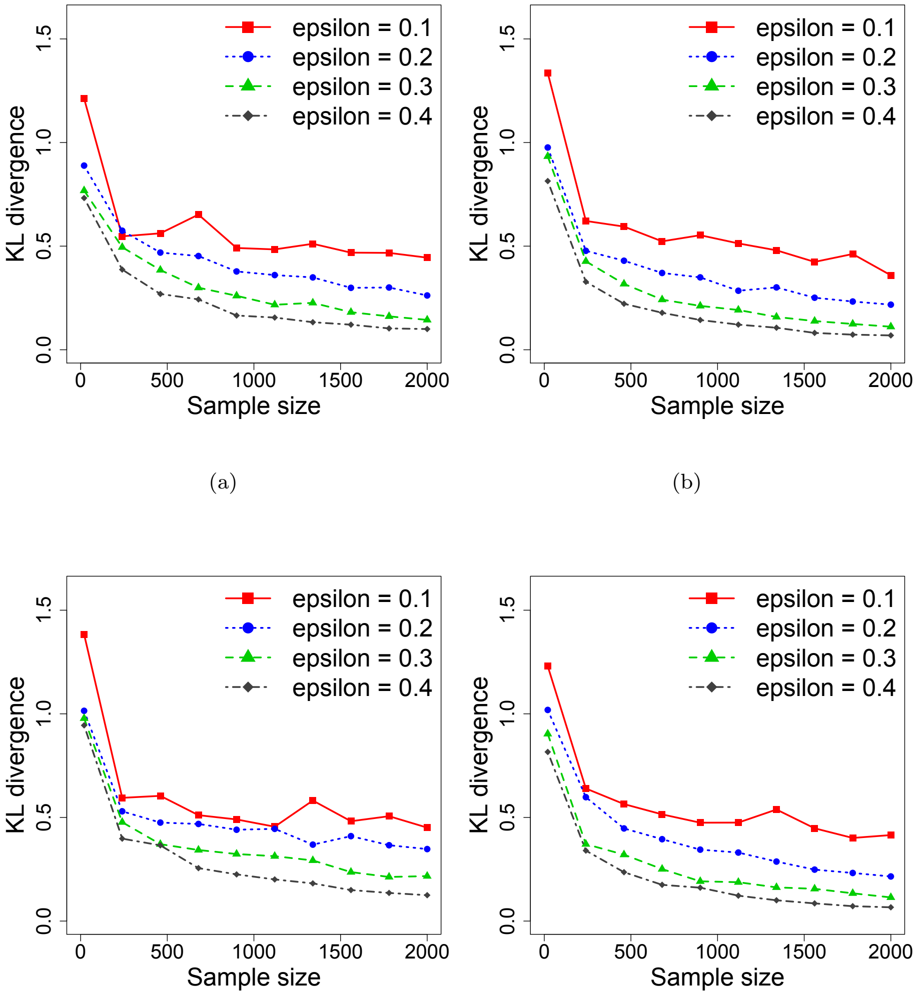
(c)

(d)

Figure 5: KL divergence between the original χ^2-statistic and the private χ^2-statistic resulting from perturbing the χ^2-statistics based on the frequency table (a) top left, (b) top right, (c) bottom left, and (d) bottom right.

4.2.2 Perturbing the Cell Counts

We did a similar analysis on the ϵ-differentially private χ^2-statistics resulting from adding Laplace noise to the cell counts. When generating ϵ-differentially private χ^2-statistics using this approach, we need to be careful to protect the number of cases and controls and the positivity of the margins. One possibility is to apply the correction proposed by Dinur and Nissim [9]. Another simpler solution is to follow the idea in the proof of Theorem 3.2. We add i.i.d. Laplace noise with scale $2/\epsilon$ to a, b, m, and n, apply corrections if necessary (according to the definition of $\mathcal{D}$ in (3)), and compute



(a)

(b)

(c)

(d)

Figure 6: KL divergence between the original χ^2-statistic and the private χ^2-statistic resulting from perturbing the cell counts based on the frequency table (a) top left, (b) top right, (c) bottom left, and (d) bottom right.

the perturbed χ^2-statistics by the formula in (4). Based on this simpler approach and the frequency tables in (5), we computed the KL divergence between the original and private χ^2-statistics for increasing N and for four different privacy levels. The resulting plots are shown in Figure 6. The plots are similar to the ones in the previous section resulting from perturbing the χ^2-statistics directly. However, the KL-divergence when perturbing the cell counts are slightly larger in most scenarios and do not depend on the value of χ^2-statistic of the underlying frequency table.

4.2.3 Perturbing the p-Values

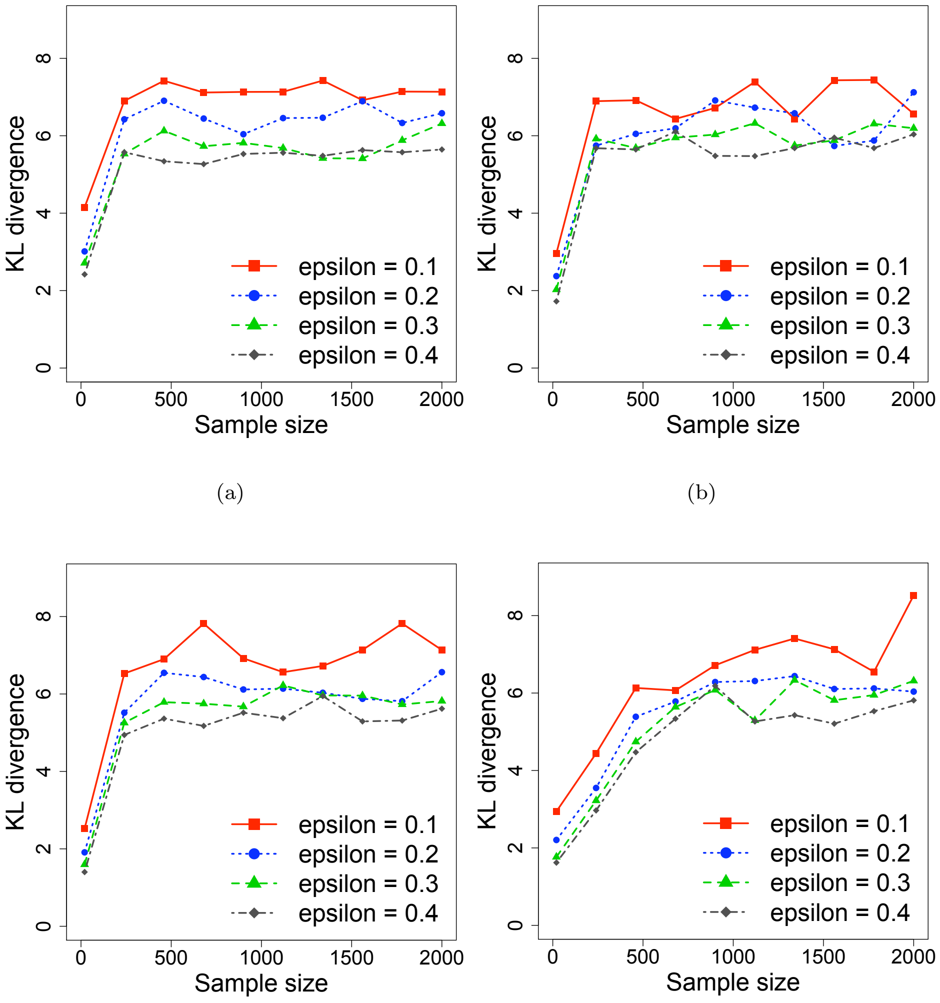
We did a similar analysis on the p-values following the proposed release mechanism of adding Laplace noise according to Theorem 3.5. Based on the frequency tables in (5), we computed the KL divergence between the original and private p-values for increasing N and for four different privacy levels. The resulting plots are shown in Figure 7. Similarly to the χ^2-statistics, the smaller the 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-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 3.6 results in plots similar to those in Figure 7; the plots depend on how much smaller the p-value under consideration is compared to 1 in the case of Theorem 3.5 and p^* in the case of Corollary 3.6.

Our analysis and the plots in Figure 7 strongly suggest that perturbing the p-values to achieve ϵ-differential privacy leads to too much noise. Making inference based on such perturbed p-values seems impossible. However, it is a valid question to ask whether there might exist a cut-off which could control the Type I & Type II errors.

We analyze this question by sampling 500 true positives (p-values in $[0, 0.05]$) and 500 true negatives (p-values in $[0.05, 1]$) uniformly and adding Laplace noise with scale $\exp(-\frac{2}{3})/\epsilon$. We represent the simulated data in an ROC plot, where we report for all possible cut-off values the resulting Type I and Type II errors. These plots for four levels of privacy, namely $\epsilon = 0.1, 0.2, 0.3, 0.4$ are shown in Figure 8. We especially indicate the point corresponding to the usual cut-off of 0.05.

Figure 8 confirms that using the perturbed p-values as a test for independence is not much better than a random test, independent of the chosen cut-off. Choosing a cut-off of 0.05 seems reasonable, but it is anyways impossible to control the Type I & Type II errors. An interesting feature in the plots are the long straight lines going from both corners along the diagonal. These lines arise since we project the perturbed p-values which fall outside the interval $[0, 1]$ to either 0 or 1. These plots show again that the perturbed p-values are dominated by these projected 0's and 1's rendering a test based on the perturbed p-values uninformative.



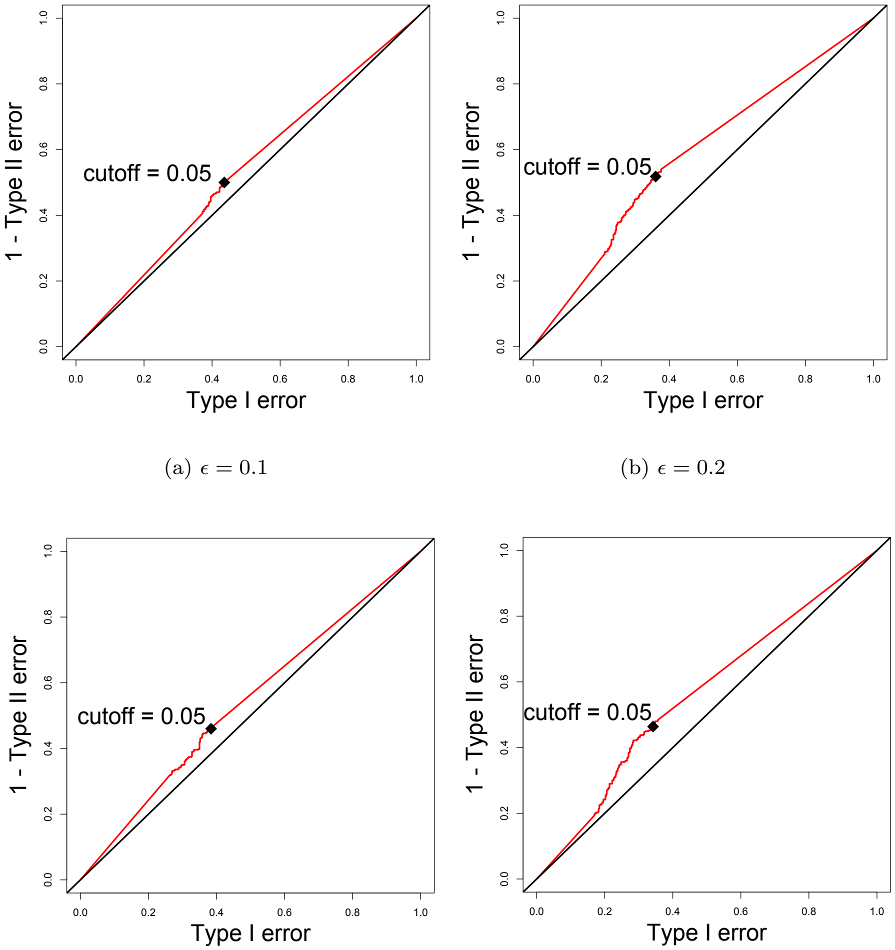
(a)

(b)

(c)

(d)

Figure 7: 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.



(a) $\epsilon = 0.1$

(b) $\epsilon = 0.2$

(c) $\epsilon = 0.3$

(d) $\epsilon = 0.4$

Figure 8: ROC curves for the perturbed p-values for different values of ϵ.

4.3 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 9. Note that when no noise is added (i.e., non-private version) the true causative SNPs are among the three highest rated SNPs in all settings.

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 9 shows that for MAF=0.4 we need about 20,000 individuals to detect both causative SNPs whereas for MAF=0.25 we need about 50,000 individuals. A smaller MAF corresponds to a sparser table, and we are in a similar situation to that described in [11], where it is shown 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.

5 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 [20]. By adapting the work of Bhaskar et al. [1] and Chaudhuri et al. [5], 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 [20] suggest an L_2-regularized logistic regression in order to detect epistasis within a small subset of SNPs. Chaudhuri et al. [5] demonstrates 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

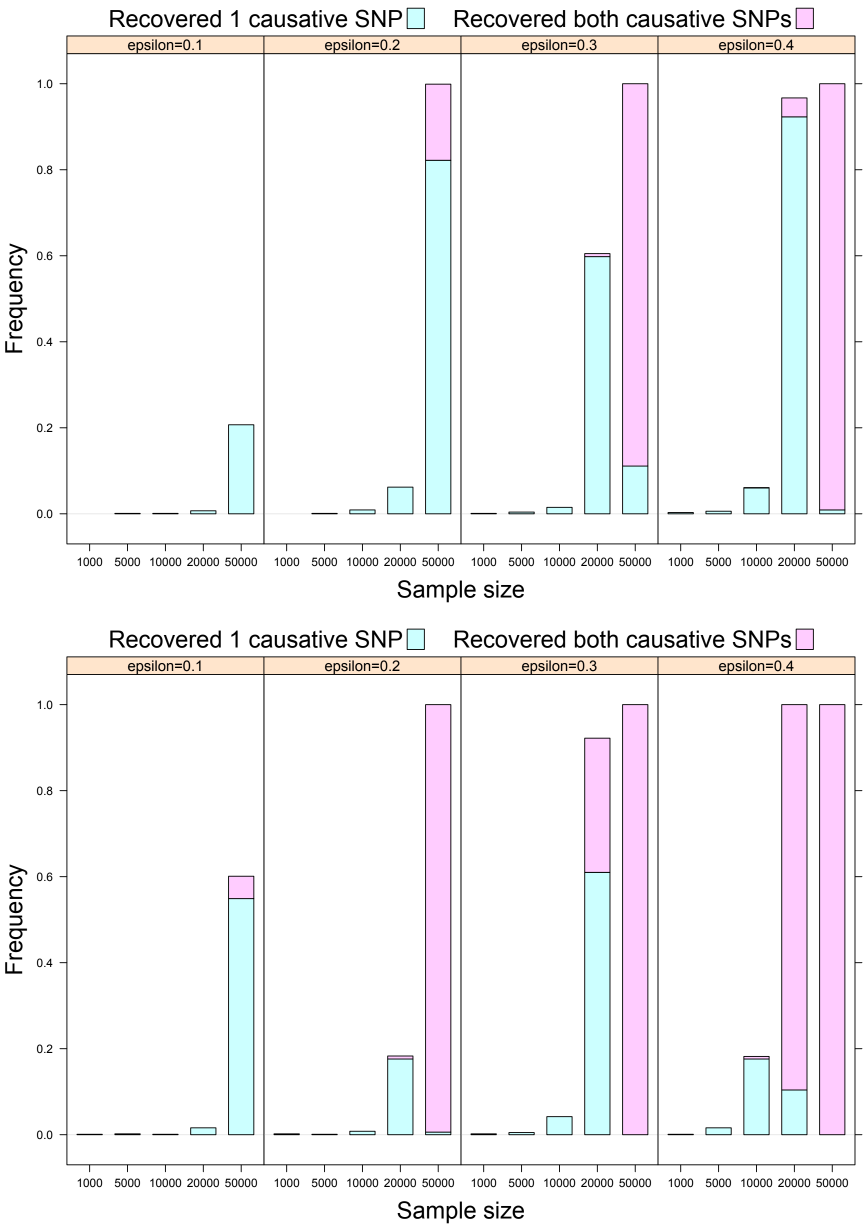


Figure 9: Bar charts representing the frequencies for which one or both of the two causative SNPs were among the three highest ranked private χ^2-statistics under the additive model with MAF=0.25 (top) and MAF=0.4 (bottom).

objective perturbation in order to find a differentially private algorithm for detecting epistasis.

Let $y = (y_1, \ldots, y_N)$ 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×SNP2 takes the states 00, 01, 02, 10, 11, 12, 20, 21, 22 and is encoded by $100000000, 010000000, \ldots, 000000001$. So an individual with genotype 12 who is not diseased would have

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

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 Park and Hastie [20] is

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

where Λ is of the form $(0, \lambda, \ldots, \lambda)$, 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 Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) score to select model size and important factors.

We can apply the approach of Chaudhuri et al. [5] to perturb the objective function such that the algorithm satisfies ϵ-differential privacy. We are interested in the following perturbed objective function:

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

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.

Following the proposal by Park and Hastie [20] we make use of 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 Algorithm 2.

Theorem 5.1. *Algorithm 2 is ϵ-differentially private.*

Proof. The proof follows by taking into account that in our application $\|x_i\|_2 \leq K$ and following step-by-step the proof of Theorem 9 in [5]. □

Algorithm 2 ϵ-Differentially Private Algorithm for Detecting Epistasis

```
Input: The data vectors $x_i, y_i$, where $i = 1, \ldots, N$ and parameters $\epsilon$ and $\lambda$.
Output: The output consists of the noisy effects.
    1. Let $\epsilon' = \epsilon - \log(1 + \frac{K}{2N\lambda} + \frac{K^2}{16N^2\lambda^2})$. If $\epsilon' > 0$, then $\delta = 0$, else $\delta = \frac{K}{4N(e^{\epsilon/4}-1)} - \lambda$ 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 $\beta_{\text{priv}} = \text{argmin}(L_{\text{priv}}(\beta) + \frac{1}{2}\delta\|\beta\|_2^2)$.
```

This result allows us to move away from a SNP-by-SNP analysis to an integrated approach without necessarily requiring an unrealistically large number of individuals in a study or relaxing the privacy constraints. Applying this method to actual GWAS data is part of ongoing work.

6 Conclusion

In this paper, we have demonstrated that it is possible, using the formal privacy guarantees of differential privacy, for NIH and other GWAS data repositories as well as "GWAS 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 may be at least in part sufficient for traditional piecewise SNP-by-SNP analysis. More specifically, our results on finite sample properties of differentially-private χ^2-statistics show that by using the Laplace mechanism and adding noise directly to the χ^2-statistic achieves the best trade-off between privacy and utility, in comparison to adding noise to the p-values or cell entries themselves, in particular for tables with small to moderate counts and overall sample size. However, we require more complex methodology 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 [22]. Applying the ϵ-differentially private algorithm outlined in Section 5 to actual GWAS data and analyzing its statistical utility is part of ongoing work.

In the work presented here, we have assumed that the number of SNP statistics one would like to release (i.e., M) is fixed before seeing the data. An interesting extension of this work is to determine M in a private way. This is an important issue, since one would like to release only the "relevant" SNPs, but the number of "relevant" SNPs is not known beforehand. In addition, it would be interesting to better understand in which scenarios perturbing the χ^2-statistic leads to a smaller KL-divergence than perturbing the cell counts. This would require a careful analysis of contingency tables with a fixed χ^2-statistic but varying cell counts. Finally, we have analyzed differentially-private χ^2-statistics and p-values based on the Laplace mechanism. Since the introduction of differential privacy by [10], 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 [12]). To further improve the privacy-utility tradeoffs for GWAS, the future research would consider such alternate mechanisms. For example, it would be interesting to see if the statistical utility of the perturbed χ^2-statistics or p-values could be improved for example by applying smooth sensitivity or the exponential mechanism.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This research was supported in part by NSF Grants BCS-0941553 and BCS-0941518 to Pennsylvania State University and Carnegie Mellon University, respectively. This work was also supported in part by the National Center for Research Resources, Grant UL1 RR033184, and is now at the National Center for Advancing Translational Sciences, Grant UL1 TR000127 to Pennsylvania State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF and NIH.

References

[1] Bhaskar, R., Laxman, S., Smith, A., and Thakurta, A. (2010). Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[2] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

[3] Braun, R., Rowe, W., Schaefer, C., Zhan, J., and Buetow, K. (2009). Needles in the haystack: Identifying individuals present in pooled genomic data. *PLoS Genetics*, 5(9): e1000668.

[4] Cadieu, E., Neff, M. W., Quignon, P., Walsh, K., Chase, K., Parker, H. G., Vonholdt, B. M., Rhue, A., Boyko, A., Byers, A., Wong, A., Mosher, D. S., Elkahloun, A. G., Spady, T. C., Andre, C., Lark, K. G., Cargill, M., Bustamante, C. D., Wayne, R. K., and Ostrander, E. A. (2009). Coat variation in the domestic dog is governed by variants in three genes. *Science*, 326(5949): 150–153.

[5] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12: 1069–1109.

[6] Clayton, D. (2010). On inferring presence of an individual in a mixture: A Bayesian approach. *Biostatistics*, 11(4): 661–673.

[7] Couzin, J. (2008). Genetic privacy: Whole genome data not anonymous, challenging assumptions. *Science*, 321(5894): 1268–1374.

[8] Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26: 363–397.

[9] Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. *Principles of Database Systems 2003*, 202–210.

[10] Dwork, C., McSherry, F., Nissim, M., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284.

[11] Fienberg, S., Rinaldo, A., and Yang, X. (2010). Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proceedings of the 2010 Conference on Privacy in Statistical Databases*, 187–199. Springer-Verlag.

[12] Hardt, M., Ligett, K., and McSherry, F. (2010). A simple and practical algorithm for differentially private data release. ArXiv:1012.4763v1. `http://arxiv.org/abs/1012.4763v1`.

[13] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8): e1000167.

[14] Jacobs, K. B., Yeager, M., Wacholder, S., Craig, D., Kraft, P., Hunter, D. J., Paschal, J., Manolio, T. A., Tucker, M., Hoover, R. N., Thomas, G. D., Chanock, S. J., and Chatterjee, N. (2009). A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, 41: 1253–1257.

[15] Loukides, G., Gkoulalas-Divanis, A., and Malin, B. (2010). Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy Sciences U S A.*, 107(17): 7898–7903.

[16] Lumley, T. and Rice, K. (2010). Potential for revealing individual-level information in genome-wide association studies. *Journal of the American Medical Association*, 303(7): 659–660.

[17] Malaspinas, A. and Uhler, C. (2011). Detecting epistasis via Markov bases. *Journal of Algebraic Statistics*, 2: 36–53.

[18] Masca, N., Burton, P. R., and Sheehan, N. A. (2011). Participant identification in genetic association studies: Improved methods and practical implications. *International Journal of Epidemiology*, 40(6): 1629–1642.

[19] P^{3}G Consortium, Church, G., Heeney, C., Hawkins, N., de Vries, J., Boddington, P., Kaye, J., Bobrow, M., and Weir, B. Public access to genome-wide data: Five views on balancing research with privacy and protection. *PLoS Genetics*, 10(e1000665).

[20] Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1): 30–50.

[21] Sankararaman, S., Obozinski, G., Jordan, M. I., and Halperin, E. (20010). Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41: 965–967.

[22] Speed, D. and Tavaré, S. (2011). Sparse partitioning: Nonlinear regression with binary or tertiary predictors, with application to association studies. *The Annals of Applied Statistics*, 5(2A): 873–893.

[23] Visscher, P. M. and Hill, W. G. (2009). The limits of individual identification from sample allele frequencies: Theory and statistical analysis. *PLoS Genetics*, 5(9): e1000628.

[24] Vu, D. and Slavkovic, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *Proceedings of the IEEE International Conference Data Mining Workshop*, 138–143.

[25] Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., Pardo-Manuel de Villena, F., Sullivan, P. F., Wilhelmsen, K. C., and Zou, F. (2007). Simulating association studies: A data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23: 2581–2588.

[26] Zerhouni, E. and Nabel, E. G. (2008). Protecting aggregate genomic data. *Science*, 321(5894): 1278.

[27] Zhou, X., Peng, B., Li, Y., Chen, Y., Tang, H., and Wang, X. (2011). To release or not to release: Evaluating information leaks in aggregate human-genome data. In V. Atluri and C. Diaz (eds.), *Computer Security – ESORICS 2011*, volume 6879 of *LNCS*. Springer. 607–627.