Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Scalable privacy-preserving data sharing methodology for genome-wide association studies

Fei Yu [a,*], Stephen E. Fienberg [b], Aleksandra B. Slavković [c], Caroline Uhler [d]

[a] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA
[b] Department of Statistics, Heinz College, Machine Learning Department, and Cylab, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA
[c] Department of Statistics, Department of Public Health Sciences, Penn State University, University Park, PA 16802, USA
[d] Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria

ARTICLE INFO

*Article history:*
Received 25 August 2013
Accepted 23 January 2014
Available online xxxx

*Keywords:*
Differential privacy
Genome-wide association study (GWAS)
Pearson $\chi^2$-test
Allelic test
Contingency table
Single-nucleotide polymorphism (SNP)

ABSTRACT

The protection of privacy of individual-level information in genome-wide association study (GWAS) databases has been a major concern of researchers following the publication of "an attack" on GWAS data by Homer et al. (2008). Traditional statistical methods for confidentiality and privacy protection of statistical databases do not scale well to deal with GWAS data, especially in terms of guarantees regarding protection from linkage to external information. The more recent concept of differential privacy, introduced by the cryptographic community, is an approach that provides a rigorous definition of privacy with meaningful privacy guarantees in the presence of arbitrary external information, although the guarantees may come at a serious price in terms of data utility. Building on such notions, Uhler et al. (2013) proposed new methods to release aggregate GWAS data without compromising an individual's privacy. We extend the methods developed in Uhler et al. (2013) for releasing differentially-private $\chi^2$-statistics by allowing for arbitrary number of cases and controls, and for releasing differentially-private allelic test statistics. We also provide a new interpretation by assuming the controls' data are known, which is a realistic assumption because some GWAS use publicly available data as controls. We assess the performance of the proposed methods through a risk-utility analysis on a real data set consisting of DNA samples collected by the Wellcome Trust Case Control Consortium and compare the methods with the differentially-private release mechanism proposed by Johnson and Shmatikov (2013).



## 1. Introduction

A genome-wide association study (GWAS) tries to identify genetic variations that are associated with a disease. A typical GWAS examines single-nucleotide polymorphisms (SNPs) from thousands of individuals and produces aggregate statistics, such as the $\chi^2$-statistic and the corresponding $p$-value, to evaluate the association of a SNP with a disease.

For many years researchers have assumed that it is safe to publish aggregate statistics of SNPs that they found most relevant to the disease. Because these aggregate statistics were pooled from thousands of individuals, they believed that their release would not compromise the participants' privacy. However, such belief was challenged when Homer et al. [1] demonstrated that, under certain conditions, given an individual's genotype, one only needs the minor allele frequencies (MAFs) in a study and other publicly available MAF information, such as SNP data from the HapMap[1] project, in order to "accurately and robustly" determine whether the individual is in the test population or the reference population. Here, the test population can be the cases in a study, and the reference population can be the data from the HapMap project. Homer et al. [1] defined a distance metric that contrasts the similarity between an individual and the test population and that between the individual and the reference population, and constructed a $t$-test based on this distance metric. They then showed that their method of identifying an individual's membership status has almost zero false positive rate and zero false negative rate.

However, Braun et al. [4] argued that the key assumptions of the Homer et al. [1] attack are too stringent to be applicable in realistic settings. Most problematic are the assumptions that (i) the SNPs are in linkage equilibrium and (ii) that the individual, the reference population, and the test population are samples from the same underlying population. They presented a sensitivity analysis

* Corresponding author.
*E-mail addresses:* feiy@stat.cmu.edu (F. Yu), fienberg@stat.cmu.edu (S.E. Fienberg), sesa@psu.edu (A.B. Slavković), caroline.uhler@ist.ac.at (C. Uhler).

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

http://dx.doi.org/10.1016/j.jbi.2014.01.008
1532-0464/

of the key assumptions and showed that violation of the first assumption results in a substantial increase in variance and violation of the second condition, together with the condition that the reference population and the test population have different sizes, results in the test statistic deviating considerably from the standard normal distribution.

Notwithstanding the apparent limitation of the Homer et al. [1] attack, the National Institute of Health (NIH) was cautious about the potential breach of privacy in genetic studies (see Couzin [5] and Zerhouni and Nabel [6]), and swiftly instituted an elaborate approval process that every researcher has to go through in order to gain access to aggregate genetic data.[2,3] This NIH policy remains in effect today.

The paper by Homer et al. [1] attracted considerable attention within the genetics community and spurred interest in investigating the vulnerability of confidentiality protection of GWAS databases. The research efforts include modifications and extensions of the Homer et al. attack, alternative formulations of the identification problem, and different aspects of attacking and protecting the GWAS databases; e.g., see [7–17]. In partial response to this literature, Uhler et al. [2] proposed new methods for releasing aggregate GWAS data without compromising an individual's privacy by focusing on the release of differentially-private minor allele frequencies, $\chi^2$-statistics and $p$-values.

In this paper, we develop a differentially-private allelic test statistic and extend the results on differentially-private $\chi^2$-statistics in [2] to allow for an arbitrary number of cases and controls. We start with some main definitions and notation in Section 2. The new sensitivity results are presented in Section 3. Uhler et al. [2] proposed an algorithm based on the *Laplace mechanism* for releasing the $M$ most relevant SNPs in a differentially-private way. In the same paper they also developed an alternative approach to differential privacy in the GWAS setting using what is known as the *exponential mechanism* linked to an objective function perturbation method by Chaudhuri et al. [18]. This was proposed as a way to achieve a differentially-private algorithm for detecting epistasis. But the *exponential mechanism* could in principle have also been used as a direct alternative to the *Laplace mechanism* of Uhler et al. [2]. This is in fact what Johnson and Shmatikov [3] proposed. Their method selects the top-ranked $M$ SNPs using the exponential mechanism. In Section 4 we review the algorithm based on the Laplace mechanism from [2] and propose a new algorithm based on the exponential mechanism by adapting the method by Johnson and Shmatikov [3]. Finally, in Section 5 we compare our two algorithms to the algorithm proposed in [3] by analyzing a data set consisting of DNA samples collected by the Wellcome Trust Consortium (WTCCC)[4] and made available to us for reanalysis.

## 2. Main definitions and notation

The concept of differential privacy, recently introduced by the cryptographic community (e.g., Dwork et al. [19]), provides a notion of privacy guarantees that protect GWAS databases against arbitrary external information.

**Definition 1.** Let $\mathcal{D}$ denote the set of all data sets. Write $D \sim D'$ if $D$ and $D'$ differ in one individual. A randomized mechanism $\mathcal{K}$ is $\epsilon$-differentially private if, for all $D \sim D'$ and for any measurable set $S \subset \mathbb{R}$,

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

[2] http://gwas.nih.gov/pdf/Data%20Sharing%20Policy%20Modifications.pdf.
[3] http://epi.grants.cancer.gov/dac/da_request.html.
[4] http://www.wtccc.org.uk/.

**Table 1**
Genotype distribution.

| | # Of minor alleles | | | Total |
|---|---|---|---|---|
| | 0 | 1 | 2 | |
| Case | $r_0$ | $r_1$ | $r_2$ | R |
| Control | $s_0$ | $s_1$ | $s_2$ | S |
| Total | $n_0$ | $n_1$ | $n_2$ | N |

**Table 2**
Allelic distribution.

| | Allele type | | Total |
|---|---|---|---|
| | Minor | Major | |
| Case | $r_1 + 2r_2$ | $2r_0 + r_1$ | 2R |
| Control | $s_1 + 2s_2$ | $2s_0 + s_1$ | 2S |
| Total | $n_1 + 2n_2$ | $2n_0 + n_1$ | 2N |

**Definition 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

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

for all data sets $D, D' \in \mathcal{D}^N$ such that $D \sim D'$.

Releasing $f(D) + b$, where $b \sim \text{Laplace}\left(0, \frac{S(f)}{\epsilon}\right)$, satisfies the definition of $\epsilon$-differential privacy (e.g., see [19]). This type of release mechanism is often referred to as the *Laplace mechanism*. Here $\epsilon$ is the privacy budget; a smaller value of $\epsilon$ implies stronger privacy guarantees.

### 2.1. SNP summaries using contingency tables

Following the notation in [20], we can summarize the data for a single SNP in a case-control study with $R$ cases and $S$ controls using a $2 \times 3$ genotype contingency table shown in Table 1, or a $2 \times 2$ allelic contingency table shown in Table 2. We require that margins of the contingency table be positive.

**Definition 3.** The (Pearson) $\chi^2$-statistic based on a genotype contingency table (Table 1) is

$$Y = \frac{(r_0 N - n_0 R)^2}{n_0 RS} + \frac{(r_1 N - n_1 R)^2}{n_1 RS} + \frac{(r_2 N - n_2 R)^2}{n_2 RS}.$$

**Definition 4.** The allelic test is also known as the Cochran–Armitage trend test for the additive model. The allelic test statistic based on a genotype contingency table (Table 1) is equivalent to the $\chi^2$-statistic based on the corresponding allelic contingency table (Table 2). The allelic test statistic can be written as

$$Y_A = \frac{2N^3}{RS} \frac{\left\{(s_1 + 2s_2) - \frac{S}{N}(n_1 + 2n_2)\right\}^2}{2N(n_1 + 2n_2) - (n_1 + 2n_2)^2}.$$

The Pearson $\chi^2$-test for genotype data and the allelic test for allele data are among the most commonly used statistical tests for association in GWAS. Zheng et al. [21] suggest using the allelic test when the genetic model of the phenotype is additive, and the Pearson $\chi^2$-test when the genetic model is unknown.

## 3. Sensitivity results

Under the assumption that there are an equal number of cases and controls, Uhler et al. [2] found the sensitivities of the