Privacy-Preserving Data Exploration in Genome-Wide Association Studies

Aaron Johnson*
U.S. Naval Research Laboratory
aaron.m.johnson@nrl.navy.mil

Vitaly Shmatikov
The University of Texas at Austin
shmat@cs.utexas.edu

ABSTRACT

Genome-wide association studies (GWAS) have become a popular method for analyzing sets of DNA sequences in order to discover the genetic basis of disease. Unfortunately, statistics published as the result of GWAS can be used to identify individuals participating in the study. To prevent privacy breaches, even previously published results have been removed from public databases, impeding researchers' access to the data and hindering collaborative research. Existing techniques for privacy-preserving GWAS focus on answering specific questions, such as correlations between a given pair of SNPs (DNA sequence variations). This does not fit the typical GWAS process, where the analyst may not know in advance which SNPs to consider and which statistical tests to use, how many SNPs are significant for a given dataset, etc.

We present a set of practical, privacy-preserving data mining algorithms for GWAS datasets. Our framework supports *exploratory* data analysis, where the analyst does not know a priori how many and which SNPs to consider. We develop privacy-preserving algorithms for computing the number and location of SNPs that are significantly associated with the disease, the significance of any statistical test between a given SNP and the disease, any measure of correlation between SNPs, and the block structure of correlations. We evaluate our algorithms on real-world datasets and demonstrate that they produce significantly more accurate results than prior techniques while guaranteeing differential privacy.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; J.3 [**Life and Medical Sciences**]: *Biology and genetics, health*

General Terms

Algorithms, Security

Keywords

Differential privacy, genome-wide association studies

*This research was primarily undertaken while the author was at The University of Texas at Austin.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

1. INTRODUCTION

Genome-Wide Association Studies (GWAS) have become a popular method to investigate the genetic basis of disease. A typical study examines thousands of singe-nucleotide polymorphism locations (SNPs) in a given population of patients for statistical links to a disease. Recent work has shown, however, that the large volume of data collected from each patient exposes them to privacy breaches—even if only the aggregate statistics are reported! Homer et al. showed that a patient's disease status can be inferred from the p-values of the statistical tests associating each SNP with the disease [12]. Furthermore, Wang et al. showed how correlation statistics among SNPs can be used to reconstruct patients' actual genomes [33]. As a consequence, NIH removed even aggregate GWAS results from public-access databases, hindering collaborative research on the genetic factors of disease [35].

To support such collaborative research, investigators must be able to perform typical data mining tasks involved in GWAS in a manner that preserves privacy of study participants. Differential privacy provides a mathematically rigorous framework for designing privacy-preserving algorithms, but standard differential privacy mechanisms [6, 22, 25] cannot be applied directly to GWAS. In a typical GWAS, the number of outputs (e.g., correlations between SNPs) is orders of magnitude greater than the number of patients in the study. The amount of random perturbation that must be applied to mask the contribution of any single patient scales with the number of outputs and renders the perturbed results unusable.

Fortunately, the purpose of GWAS is to discover "interesting" statistics that are most likely to indicate the genetic basis for a given disease. Privacy-preserving GWAS can be designed to produce a small number of outputs and thus achieve higher accuracy. Existing techniques for doing this [2, 8] assume that the analyst knows beforehand which questions to ask: for example, top k most significant SNPs (the analyst must know k) or a specific correlation measure between a certain pair of SNPs.

In a typical GWAS, however, the number of significant SNPs and the pairs of correlated SNPs are the *output* of the study, not the input! The analyst *explores* the dataset [24], choosing the most appropriate statistical tests, discovering the regions of the genome that look interesting, etc. Existing methods for privacy-preserving analysis of genetic data do not support this kind of exploration.

Our contributions. We develop an algorithmic framework for *ab initio* exploration of case-control GWAS datasets that allows an analyst to obtain privacy-preserving answers to key GWAS queries without prior knowledge of the "right" questions to ask. Specifically, we design and implement algorithms for accurate, differentially private computation of: (1) the number of SNPs that are significantly associated with the disease; (2) the location of the most significant SNPs; (3) p-values for any statistical test between

a given SNP and the disease; (4) the block structure of correlations among SNPs; (5) any measure of correlation between a pair of SNPs. To achieve this, we develop a general *distance-score mechanism* which may be of independent interest.

We analytically and experimentally evaluate our algorithms on real and simulated data, demonstrating their practical utility. In the few cases where a direct comparison with prior techniques is possible, we show that our algorithms produce significantly more accurate results. We are able to query hundreds of thousands of SNPs and get the exact number of significant SNPs the majority of the time; discover SNPs with p-values whose magnitude is within 20% of the maximum possible value in cases where previous work produced nearly 100% error; and also release the p-values themselves with an order of magnitude better accuracy for typical values than previous work. Finally, when patient populations are on the order of several thousand, we can release correlation blocks covering most of the interesting parts of the genome that are within 5% error 95% of the time.

2. PRIVACY-PRESERVING GWAS

A *case-control GWAS* examines a case population of patients with a given disease and a control population without the disease. The genomes of all patients are sequenced at a large number of SNP locations. GWAS aims to find association between the alleles at these locations and the disease status of the patient.

Association between SNPs and the disease. The association between the disease status and individual SNPs is typically measured using the p-value of a statistical test for independence. p-value indicates the probability that the observations are due to chance, assuming the SNP and disease are *not* actually correlated. GWAS aims to find SNPs whose p-values are low enough to be *significant* (i.e., the association of these SNPs with the disease is unlikely to be explained by pure chance) in order to further investigate the biological function of the genes covering these locations and the causal relationship between these genes and the disease. SNPs with low significance are not targeted for further research.

Therefore, one of the goals of privacy-preserving GWAS is to publish—as accurately as possible, while preserving privacy—the significance levels for the SNPs that are most correlated with the disease. The number of such SNPs is much smaller than the total number of SNPs involved in the study.

Correlation among SNPs. Many GWAS also report correlations among SNPs. SNPs that are significantly associated with the disease but not correlated with each another may represent independent risk factors for the disease and indicate the existence of different biological mechanisms associated with the disease.

SNP correlations can be taken from publicly available data rather than the study-specific patient data (e.g., [5]). This is obviously safe from the privacy perspective. Using patient data is preferred, however, when correlations in the study population are expected to be different from those in the public data. This can happen if the study population consists of a different racial group than the public data, if the disease itself affects SNP correlations, or if the set of SNPs examined in the study are not available in a public dataset.

A study may report individual correlations between SNP pairs of interest [5, 27], or a "heat map" of the correlations among all SNPs in a region [5, 13, 27, 28, 34]. Heat maps are often imprecise, but SNP correlations tend to exhibit a structure that allows the genome to be segmented into *correlation blocks* so that correlations are high within a block and low across blocks [9]. Identifying these blocks and the small set of likely allele sequences (*haplotypes*) for each block is the goal of the International HapMap Project [14].

GWAS may also report high-level patient data, such as population demographics and relevant clinical information, and general data about the genetic sequencing, such as success rates.

The challenges of differentially private GWAS. Direct application of standard differential privacy mechanisms to GWAS yields poor accuracy because the number of outputs is very large relative to the number of patients in the study. For example, consider using the basic Laplace mechanism of Dwork et al. [6] to output the degree of independence between each SNP and the disease, that is, the p-value of an independence test. Suppose that there are m SNPs and n patients. The *sensitivity* of the p-value—the maximum amount by which it can change when a single patient is replaced—is $O(1/n)$ because any value in $[0, 1]$ can be reached by replacing all n inputs. The Laplace mechanism then adds random noise with a standard deviation of $O(m/n)$. A typical GWAS analyzes on the order of $m = 10^5$ SNPs over a total patient population on the order of $n = 10^3$, thus $O(m/n)$ noise will render the outputs useless. In general, assuming that the output for every SNP reveals independent information about each patient, one cannot hope that GWAS (1) produces an output for all SNPs, (2) preserves privacy, and (3) provides useful outputs.

One way to limit the amount of noise that must be added in order to guarantee privacy is to limit the number of outputs: for example, publish only the p-values and pairwise correlations of the significant SNPs. However, the analyst must determine beforehand *which* outputs to publish. For example, Fienberg et al. [8] assume that the analyst knows in advance the number of "interesting" SNPs and the statistical tests to use. Similarly, to use the mechanism of Bhaskar et al. [2] to publish correlation blocks, the analyst must know in advance the number of blocks to publish, the correlation measure to use, and a scoring mechanism for choosing the blocks. This is a "Catch-22": in practice, the analyst's choice of these parameters depends on the actual dataset, and direct application of generic differential privacy mechanisms to compute these parameters suffers from the poor accuracy we wanted to avoid in the first place.

3. TECHNICAL PRELIMINARIES

Let each patient record in the database be $\mathcal{I} = \{0, 1\}^{m+1}$, where m is the number of SNPs examined. The first m bits indicate possession of at least one copy of the minor (i.e., less common) allele, the final bit indicates the disease status of the patient.[1] The space of databases is $\mathcal{D} = \mathcal{I}^n$, where n is the total number of patients. We fix m and n to be constant over all input databases, and can thus publish them without privacy loss. Let D_i denote the ith patient vector in database D. Two databases $D, D' \in \mathcal{D}$ are *neighbors* ($D \sim D'$) if they differ only in the data of a single patient.

Let $O_{ij}^{k\ell}(D) = |\{h : D_{hk} = i \wedge D_{h\ell} = j\}|$, where $i, j \in \{0, 1\}$, $1 \le k, \ell \le m + 1$, be the number of patients with a value of i for the kth input bit and j for the ℓth bit. Similarly, $O_i^k(D) = |\{h : D_{hk} = i\}|$. Let $f_{ij}^{k\ell}(D) = O_{ij}^{k\ell}(D)/n$ be the fraction of patients with values i and j for bits k and ℓ, respectively, and $f_i^k(D) = O_i^k(D)/n$. Let $p : \mathcal{N}_0^4 \to [0, 1]$ be a test of independence yielding a p-value, and let $p_i(D)$ be the p-value of the ith SNP as determined by p. Let $c : \mathcal{N}_0^4 \to [0, 1]$ be a correlation metric, and let $c_{ij}(D)$ be the correlation of the ith and jth SNPs as determined by c. In general, we drop the database, superscript, or subscript notation where it is clear from the context.

Disease association. Statistical tests used to obtain p-values for independence between a SNP and the disease include χ^2 [5, 28,

[1]This assumes a *dominant* model of the gene, used, for example, by Sladek et al. [28]. Our framework works similarly for recessive and additive models.

34], Fisher's exact test [5], and logistic regression [13, 27, 34]. The outcome of an independence test is said to be significant, that is, sufficiently unlikely to have occurred by chance, when its p-value falls below a threshold τ (we use $\tau = .05$). When considering m simultaneous tests, an adjusted threshold of τ/m gives the same effective significance for all p-values that fall below the threshold.

Our framework is designed to work with any independence test that yields a p-value. By the nature of differential privacy, the lower the sensitivity of the test, the more accurate the results. We generally use the G test [30] for p-values in our analysis because it has reasonably low sensitivity.

Correlation blocks. Blocks of correlated SNPs can be defined using the "confidence-interval method" of Gabriel et al. [9] or the "four-gamete rule" of Wang et al. [32]. The popular HaploView software package [1] also implements its own "solid spine" method. All of these methods (1) choose some measure of correlation, (2) use it to define when a sequence of SNPs constitutes a *valid block*, and (3) compute a set of disjoint valid blocks.

Our algorithms for privacy-preserving computation of correlation blocks take the correlation measure as a parameter. Common measures include the normalized linkage disequilibrium D' [27, 28], and the correlation coefficient r^2 [5, 13, 34]. We use HaploView's solid-spine definition of a valid block (Definition 1) because it is insensitive and simple. Its parameters are τ_M, a threshold for the minor-allele frequency (MAF) of a SNP, and τ_c, a threshold for the correlation between two SNPs (we use $\tau_M = 0.225$ and $\tau_c = 0.5$). Recall that f_1^k denotes the minor-allele frequency of SNP k.

DEFINITION 1. *The segment from the kth to the ℓth SNP, $k \le \ell$, is a* valid block *if all of the following hold: $f_1^k > \tau_M$; $f_1^\ell > \tau_M$; $c_{k\ell} > \tau_c$; and $\forall_{k<i<\ell}(c_{ki} > \tau_c \wedge c_{\ell i} > \tau_c) \vee f_1^i < \tau_M$.*

This definition uses a MAF threshold because correlation measures are less reliable (and more sensitive) when the number of observations is small. Thus, a block is valid if the ends have a significant correlation and each interior SNP is not significantly *un*correlated with the ends.

To select a set of disjoint blocks, we greedily choose the longest valid block that does not overlap with previously chosen blocks. If the underlying correlation structure is truly a sequence of segments such that SNPs are correlated within segments and uncorrelated between segments, this algorithm produces the correct partition. Furthermore, it helps to maximize the number of SNP pairs that are revealed to be correlated. Haploview uses similar greedy methods.

Differential privacy. Differential privacy [6] is a popular framework for designing privacy-preserving algorithms. If $\mathcal{D}$ is the space of possible input databases and R the range of possible outputs, a mechanism $\mathcal{M} : \mathcal{D} \to R$ is ϵ-differentially private if, for all D, D' that differ only in a single input (i.e., $D \sim D'$), $Pr[\mathcal{M}(D) = r] \le Pr[\mathcal{M}(D') = r]e^\epsilon$, where ϵ is the privacy parameter.

The amount of random noise that needs to be added to the output of a computation to achieve differential privacy depends on the sensitivity of the computation. For $f : \mathcal{D} \to \mathbb{R}^k$, sensitivity is defined as $\Delta_f = \max_{D \sim D' \in \mathcal{D}} \|f(D) - f(D')\|_1$.

Differential privacy "composes" [6]: if $\mathcal{M}_1$ and $\mathcal{M}_2$ are ϵ-differentially private, then $\mathcal{M}_2 \circ \mathcal{M}_1$ is 2ϵ-differentially private.

4. GWAS QUERIES

In this section, we give an overview of GWAS queries supported by our framework. In the following, $D \in \mathcal{D}$ is the input database; S is a set of m SNPs; B is a set of blocks $(b_1, b_2), 1 \le b_1 \le b_2 \le m$; i, j, k are integers in $[1, m]$; and $\epsilon > 0$ is the privacy parameter. For any differentially private query $Q(\epsilon, \cdot)$, let Q_0 be the "true" value, which the query would return if privacy had not been a concern.

NumSig(ϵ, D, p, k) returns the number of significant SNPs in D as determined by p, with a factor-2 approximation above k. Let $s = |\{i : p_i(D) \le \tau|$. Then $\mathsf{NumSig}_0 = s$ if $s \le k$ and $\mathsf{NumSig}_0 = \max(2^{\lfloor \lg(s) \rfloor}, k)$ otherwise. Thus for any $\sigma > k$, if NumSig returns σ, then there are at least σ and fewer than 2σ significant SNPs. The approximation above k allows the mechanism to lose accuracy on high values, where it matters less, in exchange for higher accuracy on low values.

LocSig(ϵ, D, p, B, k) returns the location of top k SNPs from B, ordered by increasing p-values as determined by p. B allows the querier to focus on the segments of most interest, for example by excluding regions correlated with SNPs already known to have significant p-values.

LocBlock(ϵ, D, c, B, S) returns the location of the longest correlation block, following the definition from Section 3 with c as the correlation measure, and with the additional constraint that the segment be wholly contained within some block in B and include some SNP in S. B can be used to, for example, exclude blocks that are already known. S allows the querier to find the correlation block containing a known SNP.

SNPpval(ϵ, D, i, p) returns the p-value of a given SNP, that is, $\mathsf{SNPpval}_0 = p(O_{00}^{i(m+1)}, O_{01}^{i(m+1)}, O_{10}^{i(m+1)}, O_{11}^{i(m+1)})$.

SNPcorr(ϵ, D, i, j, c) returns the correlation of two SNPs, that is, $\mathsf{SNPcorr}_0 = c(O_{00}^{ij}, O_{01}^{ij}, O_{10}^{ij}, O_{11}^{ij})$.

Using queries for exploratory GWAS analysis. The above queries allow the analyst to progress from a rough idea about the number of significant SNPs to detailed information about their identities, p-values, and mutual independence. They also enable the analyst to select the statistical tests and correlation measures that are most appropriate to the study.

For example, the analyst may begin by using NumSig with privacy parameter $\epsilon/4$ to obtain an upper bound k on the number of significant SNPs. Then, he can obtain their locations S by executing LocSig with privacy parameter $\epsilon/4$, assuming k is small enough for all of them to be accurately released. The analyst may then try to discover the correlation blocks containing the resulting SNPs by iteratively executing LocBlock with privacy parameter $\epsilon/(4k)$, SNP set S, and a block set B that excludes all previously released blocks. Finally, he can obtain the p-values for the specific significant SNPs by using SNPpval on the SNPs in S, with privacy parameter $\epsilon/(4k)$ each time. The resulting outputs will be ϵ-differentially private.

This process illustrates the general strategy of **using the data itself to focus on the most useful results while maintaining privacy and accuracy**. Other combinations of queries may be useful, too. For example, the analyst may want to obtain *independent* significant SNPs located in different blocks. To achieve this, he can alternate calls to LocSig and LocBlock, asking for only one SNP at a time and excluding those contained in previously released blocks.

Also, because all queries take the statistical measures as parameters, the analyst can choose those most appropriate for the domain and change them as data exploration progresses. He may, for example, use one of the preceding procedures with the G test as the independence test p and the correlation coefficient r^2 as the correlation measure c. These have relatively low sensitivity, and thus they would provide accurate results for the set of significant SNPs and the correlation blocks that contain them. Then he could apply SNPpval with the χ^2 test as p and SNPcorr with linkage disequi-

librium D' as the correlation measure c to obtain metrics that are more familiar to researchers in his field.

5. DISTANCE-SCORE MECHANISM

Many of the queries from Section 4 depend on every SNP in the input database and thus have high sensitivity, making it challenging to produce accurate, yet differentially private answers. Furthermore, some of these queries have complicated output spaces.

A standard tool for designing differentially private algorithms for complex spaces is the *exponential mechanism* [22]. To apply this mechanism, one must define a score function $q : R \times \mathcal{D} \to \mathbb{R}$ that assigns a value to each possible (output, input database) pair. The exponential mechanism $\mathcal{E}^{\epsilon}_{q,\Delta}$ has output distribution $Pr[\mathcal{E}^{\epsilon}_{q,\Delta}(D) = r] \propto e^{\frac{q(r,D)\epsilon}{2\Delta}}$.

In this paper, we define a general score function d called the *distance score* that works for arbitrary output spaces. In addition, Theorem 9 shows that it is highly accurate because it approximates the "best" possible scores under the requirement that the correct output must have the highest score, where we consider a score function to be better on a specific input if it yields a higher probability for the correct output under the exponential mechanism.

Let $f : \mathcal{D} \to R$ be the query. The score is computed as follows:

$$d(r, D) = \begin{cases} -1 \\ \quad \text{if } f(D) \neq r \wedge \exists_{D' \sim D} f(D') = r \\ -1 + max_{D' \sim D} d(D', r) \\ \quad \text{if } f(D) \neq r \wedge \nexists_{D' \sim D} f(D') = r \\ 0 \\ \quad \text{if } f(D) = r \wedge \exists_{D' \sim D} f(D') \neq r \\ 1 + min_{D' \sim D} d(D', r) \\ \quad \text{if } f(D) = r \wedge \nexists_{D' \sim D} f(D') \neq r \end{cases} \quad (1)$$

Intuitively, scores given by d are based on the distance from the input database to the "edge" of the set of databases for which a given output is the true output. Its sensitivity $\Delta_d = 1$ because moving to a neighboring database can only change the distance to the edge by 1. We use the term *distance-score mechanism* for the exponential mechanism equipped with this score function and sensitivity: $\mathcal{E}^{\epsilon}_{d,\Delta_d}$.

When the score d is not efficiently computable, we use a lower bound on the distance that also has sensitivity of at most 1. To the extent that these lower bounds approximate the true distance, they also approximate the best scores possible.

6. QUERY MECHANISMS

In this section, we give privacy-preserving mechanisms for evaluating the queries of Section 4. Proofs of privacy theorems appear in the full version of this paper. We note that these mechanisms can be implemented in polynomial time, although we omit the details.

6.1 Number of significant SNPs

The worst-case sensitivity of NumSig_0 is m because there exists a database in which every SNP count is one patient away from crossing the significance threshold. Applying the Laplace mechanism thus requires $Lap(m/\epsilon)$ noise, completely destroying the utility of NumSig_0 the true value of which ranges from 0 to m.

Instead, NumSig uses the distance-score mechanism of Section 5 with an approximate distance function. NumSig focuses on accuracy where most important—where there are a small number of significant SNPs—by reducing the output space to a set of intervals such that, for a given interval, the largest contained value is within a factor of two of the smallest contained value. This provides constant relative accuracy. In addition, values of k or less are output directly, without being given as a larger containing interval.

Instead of the actual distances, we use lower bounds that can be computed efficiently. We first compute, for each individual SNP i, the distance d^p_i to a database at which the significance of this SNP changes, where significance is determined by p. The sign of d^p_i indicates if the SNP gains significance (positive) or loses it (negative) We then consider each output of NumSig larger than NumSig_0. Let j be the number of additional SNPs that must be significant for this value to be the true query answer. To lower-bound the distance to a database such that NumSig_0 takes this larger value, we order the non-significant SNPs by their distances d^p_i and use the jth smallest one. Lower bounds on distances for outputs smaller than the true answer are calculated similarly.

Formally, let $d^p_i(D)$ be the minimum number of rows in D that must be modified to change whether or not the ith SNP is below the threshold; the sign of d^p_i indicates the significance of the ith SNP:

$$d^p_i(D) = \begin{cases} \max\{r \in \mathbb{N} : |\{j : D_j \neq D'_j\}| < r \\ \quad \Rightarrow p_i(D') \geq \tau \} & \text{if } p_i(D) \geq \tau \\ -\max\{r \in \mathbb{N} : |\{j : D_j \neq D'_j\}| < r \\ \quad \Rightarrow p_i(D') < \tau \} & \text{otherwise} \end{cases} \quad (2)$$

Let π be the permutation that sends SNPs sorted by increasing distances d^p_i to their original position (that is, $\pi^{-1}(i) < \pi^{-1}(j) \Rightarrow d^p_i(D) < d^p_j(D)$). Let R be the range of NumSig. Let σ_i be the ith smallest output in R, that is, $\sigma_i = i$ for $0 \leq i \leq k$ and $\sigma_i = 2^{\lfloor i \rfloor}$ for $i \geq k$. The following score function provides the desired lower bound on the distance score (Equation 1):

For $0 < i < |R| - 1$,

$$q_1(\sigma_i, D) = \begin{cases} \min\left(-d^p_{\pi(\sigma_i)}(D), d^p_{\pi(\sigma_{i+1})}(D)\right) - 1 \\ \quad \text{if } (p_{\pi(\sigma_i)}(D) < \tau) \wedge (p_{\pi(\sigma_{i+1})}(D) \geq \tau) \\ d^p_{\pi(i+1)}(D) \\ \quad \text{if } (p_{\pi(\sigma_i)}(D) < \tau) \wedge (p_{\pi(\sigma_{i+1})}(D) < \tau) \\ -d^p_{\pi(\sigma_i)}(D) \text{ otherwise} \end{cases},$$

$$q_1(\sigma_0, D) = \begin{cases} d^p_{\pi(\sigma_1)}(D) - 1 & \text{if } p_\pi(\sigma_1) \geq \tau \\ d^p_{\pi(\sigma_1)}(D) & \text{otherwise} \end{cases},$$

and

$$q_1(\sigma_{|R|-1}, D) = \begin{cases} -d^p_{\pi(\sigma_{|R|-1})}(D) - 1 & \text{if } p_\pi(\sigma_{|R|-1}) < \tau \\ -d^p_{\pi(\sigma_{|R|-1})}(D) & \text{otherwise} \end{cases}.$$

NumSig then uses the exponential mechanism and outputs $\mathcal{E}^{\epsilon}_{q_1,1}(D)$.

The sensitivity of the score function q_1 is, as with a true distance function, at most 1. Differential privacy then follows directly from the exponential mechanism.

THEOREM 1. NumSig *satisfies ϵ-differential privacy.*

6.2 Location of significant SNPs

To allow accurate release of the identities of the significant SNPs, LocSig limits the number of revealed SNPs to a user-specified parameter k (e.g. as obtained via NumSig). The smaller k, the more accurate the output of LocSig. The LocSig mechanism is given in Algorithm 1. It uses the distance to significance as the score for a SNP and then iteratively applies the exponential mechanism on the

Algorithm 1 LocSig mechanism

function LocSig(ϵ, D, p, B, k)
  $q_2(i, D) \leftarrow \begin{cases} -\infty & \text{if } \nexists_{b \in B} i \in b \\ d_i^p(D) & \text{otherwise} \end{cases}$
  for $j \leftarrow 1, k$ **do**
    repeat
      $sig \leftarrow \mathcal{E}^{\epsilon/k}_{q_2,1}(D)$
    until $sig \notin sigs$
    $sigs[j] \leftarrow sig$
  return $sigs$

SNPs in B with the privacy budget of ϵ/k for each iteration, where the user specifies B to indicate which regions to search.

Privacy of LocSig follows from that of the exponential mechanism and the compositionality of differential privacy.

THEOREM 2. LocSig *satisfies ϵ-differential privacy.*

6.3 Location of correlation blocks

There are 2^{m-1} ways to fully partition a region with m SNPs into blocks. The huge size of this output space does not allow both privacy *and* accuracy when the size of the input $n << m$.

LocBlock just outputs the longest block such that the output is within some block $b \in B$ and contains some SNP $s \in S$. Thus the output range can be limited only to the areas of most interest, improving accuracy. LocBlock uses the distance-score mechanism but with an approximate distance function that uses efficiently-computable distance lower bounds. We build up these bounds through a series of intermediate bounds for functions related to the longest block.

Let $d^1_k(D)$ be the minimum number of patients whose input to D must be changed to reduce the MAF of the kth SNP to below τ_M. Let $d^2_{k\ell}(D, c)$ be the minimum number of patients whose input must be changed to bring the MAF of SNPs k and ℓ above τ_M and the correlation between the SNPs above τ_c. Let $d^3_{k\ell}(D)$ be the minimum number of patients whose input must be changed to bring the MAF of SNPs k and ℓ above τ_M and the correlation between the SNPs below τ_c.

$d^4_{k\ell}(D, c)$ is a lower bound on the distance from D to a database for which the (k, ℓ) segment is a valid block. To obtain d^4, we observe that *(i)* the distance to a database satisfying a conjunction of conditions is at least the maximum of the distances to each of the conditions individually and *(ii)* the distance to a database satisfying a disjunction of conditions is at least the minimum of the distances. A lower bound on the distance to satisfying the first three conditions of Definition 1 is already given by $d^2_{k\ell}(D, c)$. A lower bound on satisfying all four conditions can thus be given by

$$d^4_{k\ell}(D, c) = \max\Bigg(d^2_{k\ell}(D, c), \max_{k<i<\ell}\left(\min(d^1_i(D), d^2_{ki}(D, c)), \min(d^1_i(D), d^2_{i\ell}(D, c))\right)\Bigg).$$

We can similarly obtain a lower bound $d^5_{k\ell}(D, c)$ on the distance to a database for which (k, ℓ) violates some condition of Definition 1 and thus is not a valid block: $d^5_{k\ell}(D, c) =$

$$\min\left(d^1_k(D), d^1_\ell(D), d^3_{k\ell}(D, c), \min_{k<i<\ell}\left(d^3_{ki}(D, c), d^3_{i\ell}(D, c)\right)\right).$$

$d^6_{k\ell}(D, c, B, S)$ gives a lower bound on the distance to a database for which the (k, ℓ) segment is the longest valid block among those within some $b \in B$ and containing some $s \in S$. For (k, ℓ) to be such a block, *(i)* it must be a valid block within some $b \in B$ and containing some $s \in S$, and *(ii)* every longer block within some $b \in B$ and containing some $s \in S$ must not be valid. Let the set of blocks longer than (k, ℓ), within some $b \in B$, and containing some $s \in S$ be $L_{k\ell}$. A lower bound on the distance to becoming the longest valid block is

$$d^6_{k\ell}(D, c, B) = \max\left(d^4_{k\ell}(D, c), \max_{(g,h)\in L_{k\ell}}\left(d^5_{gh}(D, c)\right)\right).$$

$d^7_{k\ell}(D, c, B, S)$ gives a lower bound on the distance to a database for which the (k, ℓ) segment is not the longest valid block among those contained within some $b \in B$ and containing some $s \in S$. For this to be the case, either *(i)* (k, ℓ) must not be a valid block within some $b \in B$ and containing some $s \in S$, or *(ii)* there must exist a longer valid block within B and intersecting S. Thus, the following provides the desired lower bound:

$$d^7_{k\ell}(D, c, B, S) = \min\left(d^5_{k\ell}(D, c), \min_{(g,h)\in L_{k\ell}}\left(d^4_{gh}(D, c)\right)\right).$$

The mechanism LocBlock(ϵ, D, c, B, S) has range $R = \{(i, j) : 1 \le i \le j \le m\}$. The following score function on R uses the distance lower bounds as the distance score (Equation 1) would use the true distances:

$$q_3(B, S; (i, j), D) = \begin{cases} -\infty & \text{if } (\nexists_{(b_1,b_2)\in B} i \ge b_1 \wedge j \ge b_2) \vee \\ & \quad \nexists_{s\in S} i \le s \le j \\ d^7_{ij}(D, c, B, S) - 1 & \text{if } d^7_{ij}(D, c, B, S) > 0 \\ -d^6_{ij}(D, c, B, S) & \text{otherwise} \end{cases}.$$

The sensitivity of q_3 is, as with a true distance function, at most 1. LocBlock simply applies the exponential mechanism $\mathcal{E}^{\epsilon}_{q_3,1}$.

THEOREM 3. LocBlock *satisfies ϵ-differential privacy.*

6.4 P-value of a given SNP

As explained in Section 2, the sensitivity of the p-value is at least $1/n$, while significant values themselves are at most the threshold value, which is less than $1/m$ (where $n << m$). A direct application of the Laplace mechanism would add noise with a standard deviation of at least $1/n$, completely obscuring the values.

Instead, we use the Laplace mechanism to output the counts for each possible allele-population pair for the ith SNP. These counts have low sensitivity, and independence tests are generally robust to small changes in the counts, allowing the p-value to be computed with good accuracy. Let the noise random variables N_{jk}, $j, k \in \{0, 1\}$, be independent and have distribution $Lap(2/\epsilon)$.[2] We compute the privacy-preserving versions of these counts as $\hat{O}_{jk}(D) = \max(O^{i(m+1)}_{jk}(D) + N_{jk}, 0)$. We then compute SNPpval as $p(\hat{O}_{00}(D), \hat{O}_{01}(D), \hat{O}_{10}(D), \hat{O}_{11}(D))$.

Privacy of SNPpval follows directly from that of the Laplace mechanism [6].

THEOREM 4. SNPpval *satisfies ϵ-differential privacy.*

6.5 Correlation of SNPs

For similar reasons as in Section 6.4, we add noise to the counts of pairwise values for i and j rather than directly to the correlation measure. The privacy-preserving correlation measure is then computed from the noisy counts. Let the random variables $N_{k\ell}$, $k, \ell \in \{0, 1\}$, be independent and have distribution $Lap(2/\epsilon)$. The noisy pairwise counts for SNPs i and j are $\hat{O}_{k\ell}(D) = \max(O^{ij}_{k\ell}(D) + N_{k\ell}, 0)$. SNPcorr is then computed as $c(\hat{O}_{00}(D), \hat{O}_{01}(D), \hat{O}_{10}(D), \hat{O}_{11}(D))$.

SNPcorr uses essentially the same mechanism as SNPpval.

THEOREM 5. SNPcorr *satisfies ϵ-differential privacy.*

[2]The distribution $Lap(b)$ has density function $f(x) = e^{-|x|/b}/(2b)$

7. UTILITY

We evaluate the utility of our mechanisms using both theoretical and experimental analysis. For two query types (location and p-values of significant SNPs), we show that our mechanisms give significantly more accurate answers than prior state of the art [8]; for the other queries, this paper gives the first known construction. Finally, we prove that the distance-score mechanism provides good accuracy in general relative to other score functions. Proofs of the theorems and detailed experimental descriptions appear in the full version of the paper [15].

7.1 Number of significant SNPs

We evaluate NumSig on data from the GWAS on Irritable Bowel Syndrome (IBS) by Duerr et al. [5]. This study involves 1138 patients, 567 of whom have the disease and 571 do not, and sequences a total of 308,332 SNPs. Among these, it finds several SNP associations within a region of the genome that codes for a known gene and publishes the counts for all 42 studied SNPs in this region. We include the counts for these 42 SNPs into our input dataset. We extend these to m total counts by randomly selecting minor-allele frequencies and then counts based on those frequencies. These added counts simulate the SNPs not published in the IBS study because they are likely independent of the disease. The accuracy of privacy-preserving outputs is very dependent on the size of the database; thus we consider different population sizes n. For the significant SNPs taken from the IBS study, we increase n above 1138 by taking the observed allele frequencies in the case and control populations as probabilities and sampling n times.

We consider the accuracy of NumSig on $m = 10^5$ total SNPs, which is a typical magnitude for GWAS. We set the NumSig accuracy threshold to $k = 1$, and we use the significance threshold of $\tau = .05/10^5$. The G test is used to calculate p-values because it is simple yet has low sensitivity. The correct output in our experiment is $\mathsf{NumSig}_0 = 2$.

Figure 1 displays as a function of the patient population the probability of the correct output and the upper limit of the output values in the 95th and 99th percentiles. It also shows the effect of changing the privacy parameter by using the values $\epsilon \in \{0.2, 0.6, 1\}$. The figure shows that, for $\epsilon = 1$ and at the IBS-study population

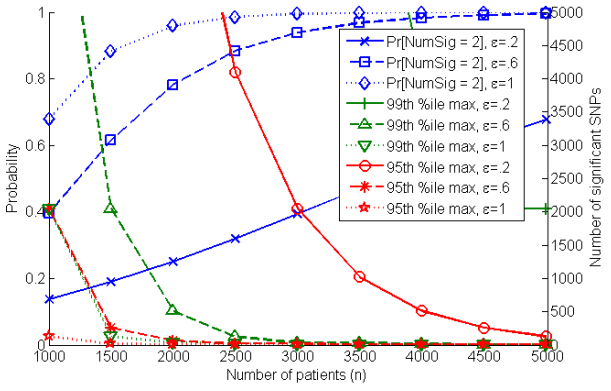

Figure 1: Output of NumSig, $m = 10^5$, $\mathsf{NumSig}_0 = 2$

size of $n = 1138$, the probability of $\mathsf{NumSig} = 2$ is greater than .5, and the probability that the output range extends beyond 128 is less than 5%. The accuracy of NumSig quickly improves with larger patient populations: when $n = 3000$ and $\epsilon = 1$, the probability of an incorrect output is less than 1%. The figure also shows that increasing ϵ by some factor provides about the same improvement in accuracy as if instead the population were increased by the same factor.

To understand the dependence of NumSig on n, consider increasing n by scaling each O_{ij} by some constant ρ. Then the G test statistic also increases by a factor ρ. This is essentially what happens to the significant SNPs as we increase n in our experiments (although with sampling error). The G test statistics of the other SNPs remain small, assuming they are truly independent. Thus the distances to changing the significance of a SNP increase linearly with ρ, thus the probability of producing the correct output increases exponentially.

7.2 Location of significant SNPs

LocSig may publish k SNPs that are very different from the true top k, but the published SNPs are likely to have p-values that are nearly as small as those of the true top k SNPs. The usefulness of the published SNPs increases as their p-values decrease, thus **the output of LocSig is likely to be as useful as the true top k.**

It is straightforward from the definition of the exponential mechanism that a given SNP is exponentially less likely to be output by LocSig the further its counts get from significance. To be able to make a statement about the p-values themselves, suppose that significance is determined via the G test. Let $p_1 \le \ldots \le p_k$ be the k smallest p-values for the SNPs in the *input*. Let $P_{\min}$ and $P_{\max}$ be the smallest and largest p-values in the *output* of LocSig, respectively. Theorem 6 says that the exponents of the p-values of the released SNPs are exponentially likely to be near those of the true top k SNPs. Observe that the rate at which the probability increases is relative to k and has only a weak dependence on m.

THEOREM 6. *For $p_{max} \ge p_k$ and $p_{min} \ge p_1$,*

$$Pr[\ln(1/P_{min}) > \ln(1/p_{min}) \wedge \ln(1/P_{max}) > \ln(1/p_{max})] \ge \left(1 - me^{\epsilon(\ln(1/p_{min}) - (1-o(1))\ln(1/p_1))/(2\Delta_q)}\right) \cdot \left(1 - mke^{\epsilon(\ln(1/p_{max}) - (1-o(1))\ln(1/p_k))/(2k\Delta_q)}\right)$$

We evaluate LocSig, too, on the data from the IBS GWAS by Duerr et al. [5] extended with extra SNPs. This time, to obtain an even more realistic distribution of p-values, we extend the number of SNP counts to $m = 10^5$ by taking the extra SNPs from the HapMap [14] data, specifically from a contiguous region that covers the one published by the IBS study. We change the population size by independently sampling the SNP values for both case and control populations using their minor-allele frequencies as probabilities.

The results of evaluating LocSig on this dataset are in Figure 2. We use the G test as our test of independence p. For comparison, we also show the results of the top-k SNP publication mechanism by Fienberg et al. [8]. It applies the Laplace mechanism to the χ^2 statistic of each SNP, then releases k SNPs with the highest perturbed values. It relies on the assumption that case population is always restricted to $n/2$, which makes the privacy guarantee much weaker; we show the performance with and without this assumption (denoted "restricted" and "unrestricted", respectively). The experiment uses $\epsilon = 1$ and $k = 2$.

To evaluate the output, we sum the p-value exponents $-\ln(p_i)$ of the output SNPs, and we consider the difference between this sum and the same sum for the correct output. The figure shows the relative value of this difference, i.e., what fraction of the exponent sum is lost by the mechanism. Figure 2 shows that by $n = 3000$ LocSig has no error with probability greater than 0.5, and that by

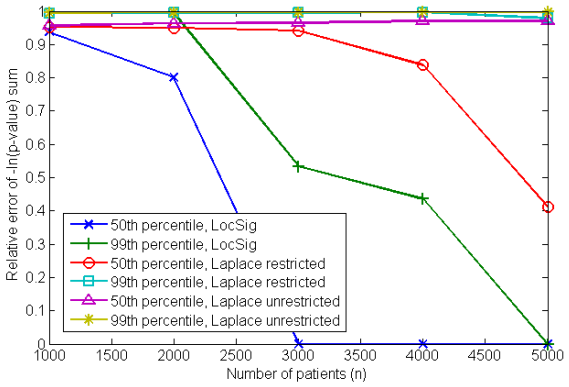

Figure 2: Relative error in top k significant SNPs: $\epsilon = 1$, $m = 10^5$, $k = 2$

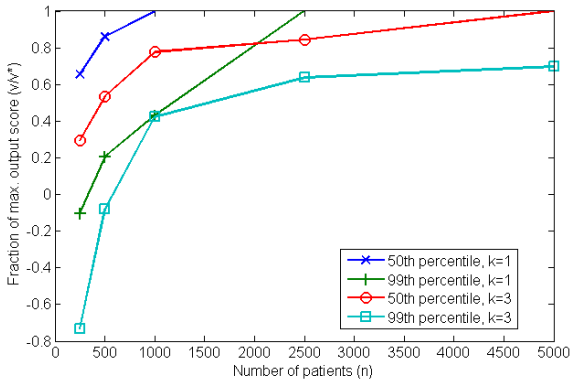

Figure 3: Value percentiles of LocBlock: $m = 250$, $\epsilon = 1$

$n = 5000$ it has no error with probability greater than 0.99. On the other hand, the restricted Laplace-based mechanism has greater than 40% error with probability at least 0.5 even for populations as large as $n = 5000$. The unrestricted Laplace mechanism is completely unusable, as for all population sizes it produces over 95% error with probability at least 0.99.

7.3 Location of correlation blocks

We evaluate LocBlock over the SNP region published in the Duerr et al. GWAS [5]. That study published a correlation heat map of a region with $m = 264$, in which three blocks are sufficient to cover over half the region. We use individual SNP sequences from HapMap starting from the beginning of that region and including $m = 250$ following SNPs. The HapMap data included $n = 1011$ patients, which we increase by resampling and decrease by subsampling. We use the correlation coefficient (i.e.$c = r^2$), and set the privacy parameter $\epsilon = 1$.

To evaluate the "quality" of an output, we consider both the correlations within the released blocks and the number of SNPs covered by these blocks. We compute the largest number of SNPs the output shares with *some* valid block and subtract the number of SNPs that are not shared, then take the sum of these values over all blocks. More precisely, let $B \subseteq \{(k,\ell) : 1 \le k \le \ell \le m\}$ be the set of valid, possibly overlapping, blocks (Definition 1). Given an output block (k,ℓ), we define its *value* v as

$$v(k,\ell) = \max_{(i,j)\in B} |\{g : (i \le g \le j) \wedge (k \le g \le \ell)\}| - |\{g : (g > j \vee g < i) \wedge (k \le g \le \ell)\}|.$$

An output block that is different from the actual top block can still have a high value if it represents another valid block of similar total length. In addition, an output block with boundaries that differ by only a small amount from a valid block has only slightly lower value. This captures the utility of a set of blocks for finding SNP correlations.

Figure 3 illustrates the distribution of values obtained by sampling from the output distribution 1000 times, as a function of the patient population size. LocBlock is used three times with the privacy parameter $1/3$, for the overall guarantee of $\epsilon = 1$. The values are shown relative to the maximum possible value v^* for the given number of blocks k.

In general, outputs are accurate when the number of output blocks is small or the patient population is large. For example, the output is optimal over half the time for a typical population size of $n = 1000$ when we limit the number of blocks to one. With three blocks, the output reaches the optimal value over half the time when the number of patients n is increased to 5000. In addition, at only $n = 2500$, with probability 0.99 the released blocks capture over half of the optimal value, indicating that they give a reasonable picture of the correlation structure in that region.

The effect of increasing the population (ignoring sampling error) is to increase all distances linearly and therefore to improve the probability of the blocks with the highest values exponentially. Consider increasing all pairwise counts $O_{ij}^{k\ell}$ by some factor ρ. This increases the distance required to change the MAF of a SNP to any given value ρ times, and, because the value of r^2 is constant in ρ, it increases the distance to any given correlation value ρ times. Thus d^1–d^3 increase by a factor ρ. d^4–d^7 then also scale with ρ because they are simply maxima and minima of d^1–d^3 and constants. Thus the score function q_3 becomes ρq_3, and the probability of the highest-value outputs increases exponentially in ρ.

7.4 P-value of a given SNP

The accuracy of the disease-association p-values calculated from the noisy counts $\hat{O}_{jk}$ depends on the sensitivity of the statistical test with respect to these counts. We consider the p-values calculated according to the G test. Let $\hat{p}$ be the noisy p-value.

The Laplace distribution of the noise has exponentially decreasing tail probabilities, and we combine this fact with bounds on the sensitivity of the G test statistic to give tail bounds for the magnitude of $\hat{p}$. To understand the effect of population size, we consider counts that grow with the population size while keeping constant the count fractions, that is, $O_{ij} = nf_{ij}$.

THEOREM 7.

$$Pr\left[\left|\ln(\hat{p}) - \ln(p)\right| > (1+o(1))\left(16\delta\left|\ln\left(\frac{f_{00}}{f_0 f_1}\right)\right|\right)\right] < 4e^{-\delta\epsilon/2}.$$

Theorem 7 shows that the noisy answers are likely to give good estimates of the smallest (and thus most interesting) p-values.

Figure 4 shows an experimental evaluation of SNPpval, directly comparing our mechanism with Fienberg et al. [8]. The latter releases both private p-values and χ^2 statistics by directly adding Laplace noise. Because it uses Pearson's χ^2 statistic for both, we use it for all mechanisms in the experiments. As before, we give results with and without the assumption that the numbers of case and control patients are both restricted to $n/2$ in all input datasets. The

experiments were done over a range of inputs, each with $n = 1000$ and minor-allele frequency of 0.2, typical for GWAS. The privacy parameter was $\epsilon = 0.2$, and for each input we sampled from the output distribution 1000 times. We measure accuracy as the root mean squared error of the logarithm of the output p-values.

The figure shows our mechanism almost always gives more accurate results. In particular, for p-values of at least 10^{-15}, which is the typical range for GWAS results, **our mechanism is an order of magnitude more accurate**. Furthermore, adding noise directly to the p-value or the χ^2 statistic without the restricted-row-count assumption produces unusable answers. With the prior mechanisms, added noise is often larger than the possible range of the true value.

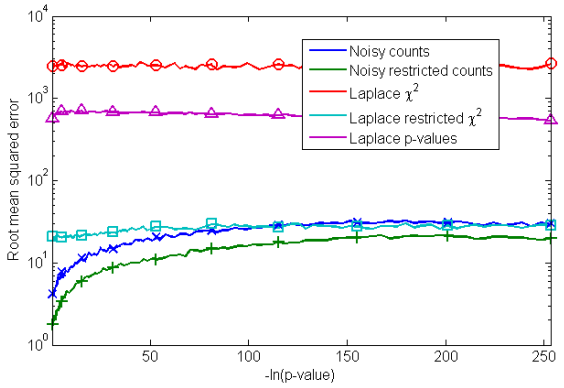

Figure 4: Root mean squared error of noisy p-values: $n = 1000$, $\epsilon = 0.2$, MAF=0.2

7.5 Correlation of SNPs

To evaluate SNPcorr, we again use the correlation coefficient $c = r^2$ and let $O_{ij} = nf_{ij}$. Also, let $\hat{r}^2$ be the noisy correlation coefficient. The exponentially decreasing tail probabilities of the noise's Laplace distribution combined with bounds on the sensitivity of r^2 imply the following bound on the change in r^2:

THEOREM 8. $Pr[|\hat{r}^2 - r^2| > 32\delta/n] < 4e^{-\delta\epsilon/2}$

Theorem 8 shows that the output of SNPcorr is exponentially close to the correct value and the width of the tail distribution is inversely proportional to n.

7.6 Distance-score mechanism

The distance-score mechanism is quite general, and we are able to show that it provides a kind of accuracy guarantee and therefore could be useful more broadly. The intuition is that, under the requirement that the correct output r^* always has the highest output probability, distances measured by d (Equation 1) constrain how much larger the score of r^* can be than the score of any other output r. Maximizing the difference between the scores of r^* and r maximizes their relative probabilities under the exponential mechanism. Theorem 9 shows that d approximately provides the maximum difference between the scores of r^* and r. Let $Q = \{q : R \times \mathcal{D} \to \mathbb{R} | f(D) = r \Rightarrow q(r, D) \geq q(r', D)\}$. Note that $d \in Q$.

THEOREM 9. *For all $q \in Q$, $D \in \mathcal{D}$, and $r \in R$,*

$$d(f(D), D) - d(r, D) \geq (q(f(D), D) - q(r, D))/(2\Delta_q)$$

Without the requirement that the correct output always has the highest score, a single mechanism cannot yield competitive scores on all inputs. For example, a mechanism that keeps the scores constant can assign an arbitrarily high score to any one output. Similarly, normalizing by Δ_q is necessary because scaling a score function does not affect the output distribution of the exponential mechanism.

8. RELATED WORK

Homer et al. [12], followed by Wang et al. [33], demonstrated the privacy risks of publishing GWAS results. The problems they identified received much attention from medical researchers [4], and NIH immediately removed public access to the data [35].

Sankararaman et al. [26] show how to limit the statistical power of Homer's attack by restricting the amount of data published. Loukides et al. [18] suggest using well-known generalization and suppression techniques [29], but this approach is known to have inherent flaws [17, 19].

Differential privacy [6] is a promising approach to privacy-preserving data publishing. It has been successfully applied to search logs [16], movie-viewing records [21], network traces [20], and social networks [11]. General-purpose mechanisms for differentially private data release (e.g. [6, 22, 25]) cannot be directly applied to GWAS data, however, because both the set of possible inputs for each patient and the number of outputs are very large, while most mechanisms rely on one of these being small in order to provide accurate outputs. We do observe that, similar to our distance-score mechanism, smooth sensitivity [23] does provide instance-dependent accuracy. However, computing the required smooth bound appears difficult in general, and our attempt to use this mechanism did not yield satisfactory accuracy [15].

Dwork and Lei [7] consider general methods for turning robust statistical estimators into differentially private estimators. Our use of the G test and r^2 correlation coefficient for their low sensitivity echo this work.

Our distance-score mechanism defines scores that are approximately optimal; Ghosh et al. [10] and Brenner and Nissim [3] investigate the problem of finding truly optimal mechanisms.

Applying differential privacy to GWAS while preserving utility requires domain-specific solutions. Fienberg et al. [8] consider how to publish minor-allele frequencies, χ^2 statistics, and p-values; they also adapt the method of Bhaskar et al. [2] to release k most significant SNPs. Their approach, however, makes several restrictive assumptions, including a fixed population size, considers only the χ^2 test, and, critically, assumes that the analyst knows in advance k, the number of SNPs to publish. Furthermore, as we show in Section 7, their mechanism often produces highly inaccurate outputs. Vu and Slavković [31] also consider applying differential privacy to medical research data, but focus on determining the population size needed to provide a given power to statistical tests when noise is added using standard mechanisms.

9. CONCLUSION

As a consequence of recent research on privacy risks in genome-wide association studies [12, 33], medical researchers now face obstacles to sharing their data and results. Differential privacy offers an approach to the problem that provides a rigorous privacy guarantee without unrealistic assumptions about the adversary's knowledge (in contrast to other approaches [18]).

Unfortunately, the massive amount of per-patient data involved in GWAS causes standard differential privacy mechanisms to produce highly inaccurate results. Previously proposed adaptations of differential privacy to GWAS are limited in scope and flexibility. In particular, they require the data analyst to know or guess in advance

features of the data, such as the number of genetic locations that are associated with the disease.

In this paper, we develop a toolkit of privacy-preserving queries for GWAS that enable *data exploration* without any background knowledge or assumptions. Our queries allow the analyst to learn the number, location, and p-values of SNPs that are significantly correlated with the disease, as well as the correlation-block structure of the genome in the areas of most interest. The statistical tests and measures can be determined during data exploration itself rather than fixed beforehand. This helps analysts to quickly identify and focus on the aspects of the genome that are most likely to yield clinically interesting results.

Analytical and empirical analysis shows that our mechanisms achieve much better accuracy than prior state of the art. We also develop a general *distance-score* technique for designing accurate, differentially private mechanisms on complex output spaces.

The accuracy of our outputs could be increased through *computational* advances by improving the distance approximations used in several of our queries. In addition, incorporating biological constraints into the possible inputs from a single patient is likely to result in more accurate mechanisms.

A question outside the scope of this paper is how accurate the outputs need to be in order to be useful to medical researchers. Like other applications of differential privacy, our results include non-trivial amounts of random noise, although inaccuracy can be reduced by increasing the number of patients in the study. Finding the right balance between accuracy, privacy, and study costs must be informed by the needs of medical researchers.

Acknowledgments. This work was partially supported by the NSF grant CNS-0746888, the MURI program under AFOSR grant FA9550-08-1-0352, and grant R01 LM011028-01 from the National Library of Medicine, NIH. Support also provided by ONR.

References

[1] J. Barrett et al. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2), 2005.

[2] R. Bhaskar et al. Discovering frequent patterns in sensitive data. In *KDD*, 2010.

[3] H. Brenner and K. Nissim. Impossibility of differentially private universally optimal mechanisms. In *FOCS*, 2010.

[4] G. Church et al. Public access to genome-wide data: Five views on balancing research with privacy and protection. *PLoS Genet*, 5(10), 2009.

[5] R. Duerr et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, 314(5804), 2006.

[6] C. Dwork et al. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[7] C. Dwork and J. Lei. Differential privacy and robust statistics. In *STOC*, 2009.

[8] S. Fienberg, A. Slavković, and C. Uhler. Privacy preserving GWAS data sharing. In *PADM*, 2011.

[9] S. Gabriel et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576), 2002.

[10] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. In *STOC*, 2009.

[11] M. Hay et al. Accurate estimation of the degree distribution of private networks. In *ICDM*, 2009.

[12] N. Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4, 2008.

[13] D. Hunter et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7), 2007.

[14] International HapMap Constortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449, 2007.

[15] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. Technical Report TR-13-13, The University of Texas at Austin, Department of Computer Science, 2013.

[16] A. Korolova et al. Releasing search queries and clicks privately. In *WWW*, 2009.

[17] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

[18] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Anonymization of electronic medical records for validating genome-wide association studies. *Proc. of NAS*, 107(17), 2010.

[19] A. Machanavajjhala et al. l-diversity: Privacy beyond k-anonymity. *TKDD*, 1(1), 2007.

[20] F. McSherry and R. Mahajan. Differentially-private network trace analysis. *CCR*, 41(4), 2010.

[21] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the Netflix Prize contenders. In *KDD*, 2009.

[22] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

[23] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

[24] D. Reshef et al. Detecting novel associations in large data sets. *Science*, 334(6062), 2011.

[25] A. Roth and T. Roughgarden. Interactive privacy via the median mechanism. In *STOC*, 2010.

[26] S. Sankararaman et al. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9), 2009.

[27] L. Scott et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 2007.

[28] R. Sladek et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130), 2007.

[29] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. 10(5), 2002.

[30] A. Tamhane and D. Dunlop. *Statistics and Data Analysis: from Elementary to Intermediate*. Prentice-Hall, Inc., 2000.

[31] D. Vu and A. Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *ICDMW*, 2009.

[32] N. Wang et al. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *The American Journal of Human Genetics*, 71(5), 2002.

[33] R. Wang, Y. Li, et al. Learning your identity and disease from research papers: Information leaks in genome wide association study. In *CCS*, 2009.

[34] M. Yeager et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39(5), 2007.

[35] E. Zerhouni and E. Nabel. Protecting aggregate genomic data. *Science*, 322(5898), 2008.