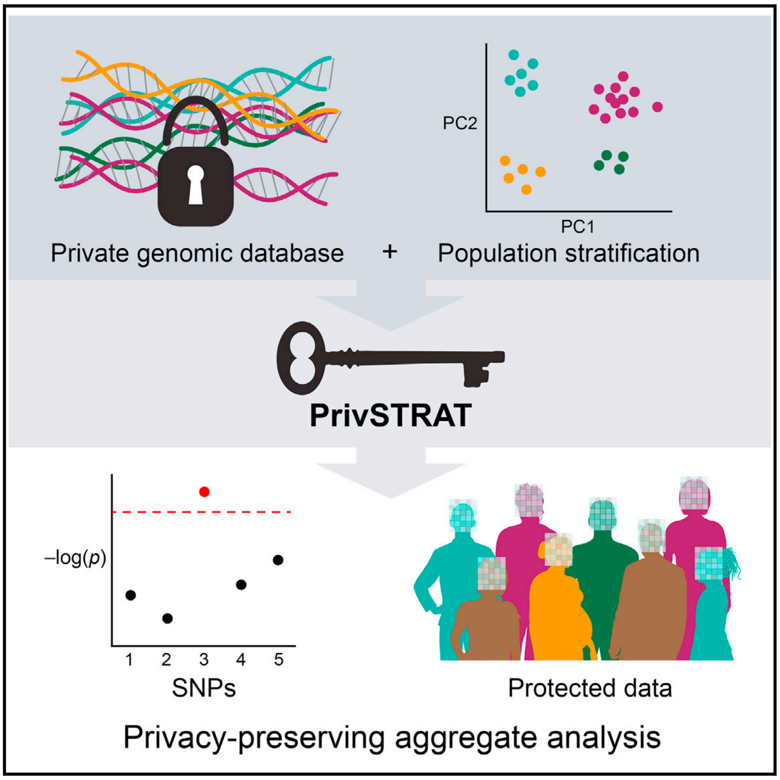
Cell Systems

Report

Enabling Privacy-Preserving GWASs in Heterogeneous Human Populations

Graphical Abstract



Highlights

- We introduce a novel variant of differential privacy tailored to genomic databases
- We enable privacy-preserving GWASs in the presence of population stratification
- We implement and test these algorithms on numerous real and synthetic datasets

Authors

Sean Simmons, Cenk Sahinalp, Bonnie Berger

Correspondence

bab@mit.edu

In Brief

Simmons et al. introduce a scalable framework for allowing privacy-preserving queries on genomic databases using cutting-edge genome-wide association study statistics that account for population stratification.

Simmons et al., 2016, Cell Systems *3*, 54–61
July 27, 2016 © 2016 The Authors. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2016.04.013

OPEN ACCESS
CellPress

Cell Systems
Focus on RECOMB
Report

Enabling Privacy-Preserving GWASs in Heterogeneous Human Populations

Sean Simmons,[1,2,3] Cenk Sahinalp,[3,4] and Bonnie Berger[1,2,*]
[1]Department of Mathematics
[2]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[4]School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA
*Correspondence: bab@mit.edu

http://dx.doi.org/10.1016/j.cels.2016.04.013

SUMMARY

The proliferation of large genomic databases offers the potential to perform increasingly larger-scale genome-wide association studies (GWASs). Due to privacy concerns, however, access to these data is limited, greatly reducing their usefulness for research. Here, we introduce a computational framework for performing GWASs that adapts principles of differential privacy—a cryptographic theory that facilitates secure analysis of sensitive data—to both protect private phenotype information (e.g., disease status) and correct for population stratification. This framework enables us to produce privacy-preserving GWAS results based on EIGENSTRAT and linear mixed model (LMM)-based statistics, both of which correct for population stratification. We test our differentially private statistics, PrivSTRAT and PrivLMM, on simulated and real GWAS datasets and find they are able to protect privacy while returning meaningful results. Our framework can be used to securely query private genomic datasets to discover which specific genomic alterations may be associated with a disease, thus increasing the availability of these valuable datasets.

INTRODUCTION

We are experiencing unprecedented growth in the amount of personal and clinical genotype data in large repositories (Lowe et al., 2009). However, accessing this growing pool of data poses major privacy concerns for individuals (Gymrek et al., 2013; Malin et al., 2013; Murphy et al., 2011; Nyholt et al., 2009). At the same time, making this data more widely available could lead to novel biomedical insights that could inform medical research (Weber et al., 2009; Lowe et al., 2009). As such, there is hope that the privacy challenges posed in analyzing such data might not simply require tighter regulations over who can use the data—which is often limited to individuals who have gone through a time-consuming and burdensome application process—but may instead benefit from the development of cryptographic tools that allow secure access while ensuring accurate analyses. In particular, there has been increased interest in the usefulness of a cryptographic technique known as differential privacy (Dwork, 2011). This technique allows researchers access to genomic data while preserving every patient's privacy (Jiang et al., 2014; Johnson and Shmatikov, 2013; Uhlerop et al., 2013; Yu et al., 2014a, 2014b; Yu and Ji, 2014; Chen et al., 2014; Zhao et al., 2015; Zhang et al., 2013). Unlike alternative methods for achieving privacy, differential privacy is able to provide formal guarantees of privacy while making minimal assumptions.

Here, we focus on privacy in the context of genome-wide association studies (GWASs) (Weber et al., 2009; Lowe et al., 2009), which are commonly used to identify SNPs associated with a given disease. Numerous works have shown that aggregate genomic data, including GWAS statistics, can leak private information about participants (Homer et al., 2008; Lumley and Rice, 2010; Im et al., 2012; Zhou et al., 2011; Sankararaman et al., 2009). These findings have led the NIH, among others, to place much of its aggregate genomic data into repositories and require researchers to apply for access (Erlich and Narayanan, 2014). Recent work has also shown that a popular method for sharing genomic data, genomic data-sharing beacons, leaks potentially private information about participants (Shringarpure and Bustamante, 2015). These results illustrate the need for new methods that allow privacy-preserving access to genomic data.

Differential privacy (Dwork et al., 2006; Dwork, 2011) (Box 1) has been proposed as one promising solution to the privacy conundrum (Jiang et al., 2014; Yu et al., 2014a, 2014b; Johnson and Shmatikov, 2013; Uhlerop et al., 2013; Yu and Ji, 2014; Chen et al., 2014; Zhao et al., 2015; Zhang et al., 2013; Tramer et al., 2015). The main advantage of differential privacy is that it gives a mathematical guarantee of privacy to all participants in a study. These guarantees make it possible to share genomic data without risking participant privacy. Interest in differential privacy has inspired development of improved methods for performing differentially private GWASs (Jiang et al., 2014), as well as numerous novel applications of differential privacy to other types of research data, including the Privacy Tools for Sharing Research Data project at Harvard University and the Enabling Medical Research with Differential Privacy project at the University of Illinois. Although the initial results of this research have been encouraging, there remain major limitations on the types

Box 1. Phenotypic Differential Privacy

The cryptographic community introduced ϵ-differential privacy as a formal definition of privacy about a decade ago (Dwork et al., 2006). Intuitively, it ensures that the results of an analysis are almost equally likely regardless of whether any one individual participates in the study (more specifically, the probability differs by a factor of $\le \exp(\epsilon)$, where ϵ is a positive real number). This helps ensure that there is negligible private information being leaked.

Here, we introduce phenotypic differential privacy, a formal definition of privacy that attempts to preserve private information about individuals (in this case, disease status). As with all forms of differential privacy, phenotypic differential privacy requires the choice of a privacy parameter (also known as the privacy budget). This parameter, denoted by ϵ, controls the level of privacy guaranteed to all participants in the study: the closer to zero it is, the more privacy is ensured, while the larger it is, the weaker the privacy guarantee is. This means we would like to set ϵ as small as possible; unfortunately, this comes at the cost of less accurate outputs (Dwork, 2011).

It is difficult to reach an intuitive understanding of ϵ. Informally, taking the frequentist's perspective (Wasserman and Zhou, 2010), one can think of exp (ϵ) as bounding the power-to-significance ratio of any statistical test the adversary might use to determine a participant's disease status based on ϵ-phenotypically differentially private data. More formally, assume there is an adversary who would like to determine the ith individual's disease status. This can be thought of as performing a hypothesis test to distinguish between $H_0 : y_i = 1$ and $H_1 : y_i = 0$ based on the output of a ϵ-phenotypically differentially private statistic. The power of such a test (where the power equals the probability of rejecting H_0 given that H_1 is true) is bounded above by exp (ϵ) times that significance level of the test (where the significance level equals the probability of rejecting H_0 given that H_0 is true).

One can also look at exp (ϵ) from a Bayesian perspective (Hsu et al., 2014). If an individual in the study is worried about some negative effect due to participating in the study (such as someone concluding they have a certain disease based on the study results), ϵ-phenotypic differential privacy guarantees the probability of that negative event occurring at most differs by a factor of exp (ϵ), based on whether the individual has the disease.

For example, $\epsilon \le 2$ implies that, for any participant with the disease under investigation, releasing F(D, y) does not increase the adversary certainty in the participant's disease status by more than a (multiplicative) factor of exp(2) < 7.5 compared to the case in which the participant does not have the disease. In agreement with previous work (Vinterbo et al., 2012), we consider $\epsilon \le 2$ to be a realistic level of privacy, though exact thresholds differ from application to application.

The privacy guarantee decreases as the number of queries increases. If the user makes k queries, where the ith query is ϵ_i-phenotypically differentially private, the result is $(\epsilon_1 + \cdots + \epsilon_k)$-phenotypically differentially private.

In practice, anyone who wants privacy-preserving access to a database begins by signing up for an account. Unlike in the controlled access model, no time-consuming application is necessary. All that is needed is to ensure a given individual has not previously signed up for an account (for example, a system could be based on academic email addresses or PubMed author identifiers). Upon registering, each user is given a privacy budget ϵ. The individual can use that privacy budget to query the database in a differentially private way until his or her privacy budget is used up (for example, the user may make k queries, where each query is (ϵ/k)-differentially private). After the privacy budget is used up, the user must apply for an increase in privacy budget. More details are given in Supplemental Experimental Procedures.

of genomic analyses that can be performed accurately and efficiently (Fredrikson et al., 2014).

Privacy concerns are not the only hurdles facing modern GWASs. Such analyses are complicated by systematic differences among different human populations (Yang et al., 2014). Biologically meaningful mutations are often inherited jointly with unrelated mutations, leading to false GWAS associations. For example, the lactase gene is responsible for the ability to digest lactose (such as in milk) and is more common in people of northern European ancestry than in those of East Asian ancestry. People from northern Europe are also, on average, taller than those from East Asia. This observation would lead a naive statistical method to erroneously suggest that the lactase gene is related to height. Such confounding effects are a major problem that can render the results of a GWAS (particularly one with a large sample size) nearly nonsensical (Marchini et al., 2004). To avoid this common problem, known as population stratification, various methods have been employed, including EIGENSTRAT (Price et al., 2006), linear mixed models (LMMs) (Yang et al., 2014), and genomic control (Devlin and Roeder, 1999). In recent years, there has been a growing interest in using LMMs for this task because of improved algorithms (Kang et al., 2010; Lippert et al., 2011; Tucker et al., 2014; Loh et al., 2015; Furlotte and Eskin, 2015). Still, EIGENSTRAT remains a common approach for correcting for population stratification. However, previous works on differentially private GWASs have not addressed population stratification, greatly limiting their real-world applicability.

Here, we jointly address the population stratification and privacy issues that arise when using private data to answer GWAS queries. We focus on two types of queries: (1) GWAS statistics at SNPs of interest and (2) lists of SNPs highly associated with diseases of interest. We develop a differential privacy framework that can transform GWAS statistics commonly used to answer these queries into tools for privacy-preserving GWASs. We demonstrate this approach on two state-of-the-art statistics, EIGENSTRAT (Price et al., 2006) and LMM-based statistics (Yang et al., 2014), through our PrivSTRAT and PrivLMM methods, respectively. We test these methods on real and synthetic data and show that they perform well in terms of both accuracy and runtime; their accuracy improves as sample sizes increase.

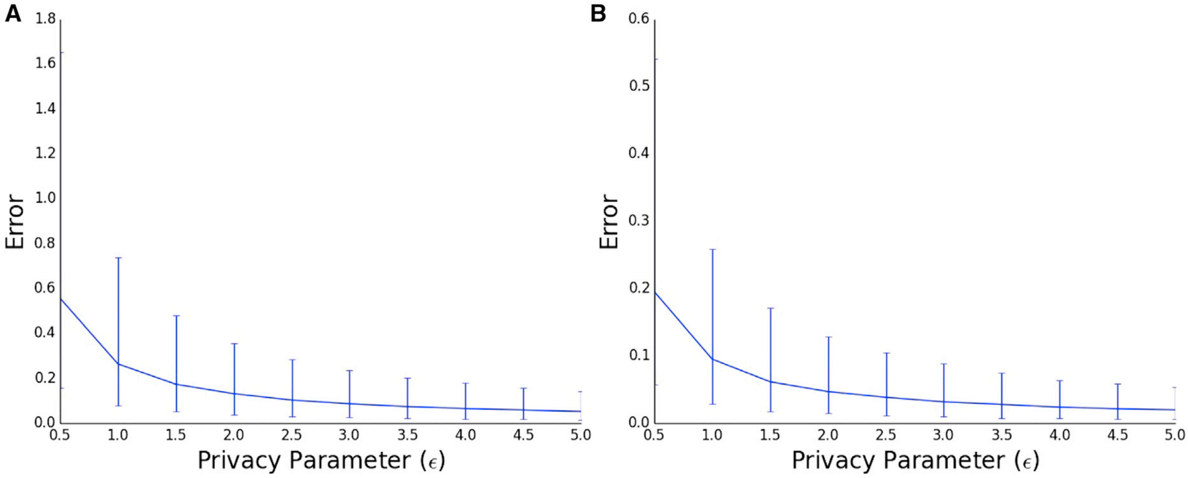


Figure 1. Accuracy of Our Mechanisms for Approximating EIGENSTRAT Statistics

(A and B) PrivSTRAT statistic accuracy of (A) real and (B) simulated GWAS data for various privacy parameters ϵ. Median error is over all SNPs, with error bars representing the 25% and 75% quantiles. As expected, accuracy increases as ϵ increases (i.e., as privacy decreases).

RESULTS

Motivating Scenario

We begin with a massive database consisting of phenotype and genotype data from a large number of individuals—for example, from the Database of Genotypes and Phenotypes or electronic health records (EHRs). The curators of the database would like to make the data available to as many individuals as possible, with the hope of supporting new research. At the same time, the database curator is responsible for protecting the privacy of the individuals in the database.

Our aim is to allow researchers access to this data while preserving privacy using a technique known as differential privacy. Because we are focusing on protecting phenotype data, we introduce a slightly modified definition known as phenotypic differential privacy (see Box 1 for an overview and Experimental Procedures for technical details). Intuitively, phenotypic differential privacy guarantees that an analysis performed on any dataset is statistically indistinguishable from the same analysis performed on any dataset that differs in any individual's disease status. This helps prevent the use of genotype information to learn about private phenotype information, and vice versa.

The exact definition of indistinguishability depends on a user-defined privacy parameter ϵ (see Box 1 for details about this parameter). The closer to zero this ϵ parameter, the greater the privacy. This indistinguishability ensures that our database offers negligible information about the participants' private phenotype information. The current work focuses on developing methods to return answers to common genomic queries that are phenotypically differentially private.

Privately Estimating EIGENSTRAT Statistics

Because privacy methods do not produce correct results on typical GWASs without population stratification (Supplemental Experimental Procedures), we start by looking at the most basic queries in GWASs, namely, the calculation of the GWAS statistic for a given SNP. In particular, we are interested in calculating a χ^2 distributed statistic, known as the EIGENSTRAT statistic, for a given SNP in our database while preserving privacy. As detailed in Experimental Procedures, this is achieved using a modified version of the Laplacian mechanism (Dwork, 2011).

We studied the trade-off between privacy (the ϵ parameter) and accuracy on real GWAS data for a rheumatoid arthritis dataset (Figure 1A) (Plenge et al., 2007), as well as on simulated data with two subpopulations (Figure 1B) (see Experimental Procedures). As expected, our method quickly increases in accuracy as privacy decreases (i.e., as ϵ increases). For reasonable values of ϵ (ϵ around 1 or 2) (Vinterbo et al., 2012), the median error introduced by our method is around 0.1 or 0.2. This is fairly small, corresponding to about a 5%–10% error in the EIGENSTRAT statistic, an amount that is unlikely to affect the final conclusions of an analysis (see Case Study). We can also calculate a phenotypically differentially private p value for a given SNP using this returned statistic (Supplemental Experimental Procedures).

Privately Selecting Highly Associated SNPs

Besides calculating χ^2 statistics, users may be interested in determining which SNPs are most highly correlated with a given disease. A simple way to make this identification would be to use the preceding method to estimate the EIGENSTRAT statistics for all SNPs and return the highest-scoring SNPs. However, the large number of queries required to do this necessitates a small ϵ parameter for each query, resulting in poor accuracy. More accurate methods have been proposed that identify high-scoring SNPs while preserving privacy (Uhlerop et al., 2013; Johnson and Shmatikov, 2013; Yu et al., 2014a). We focus on one of these methods, known as the distance-based method (Johnson and Shmatikov, 2013; Simmons and Berger, 2016). We compare the distance-based method with other methods in Supplemental Experimental Procedures, showing that it performs the best in practice, a result that is consistent with previous work (Simmons and Berger, 2016).

The distance-based method has previously been used with simple statistics such as Pearson or allelic test statistics. Here,

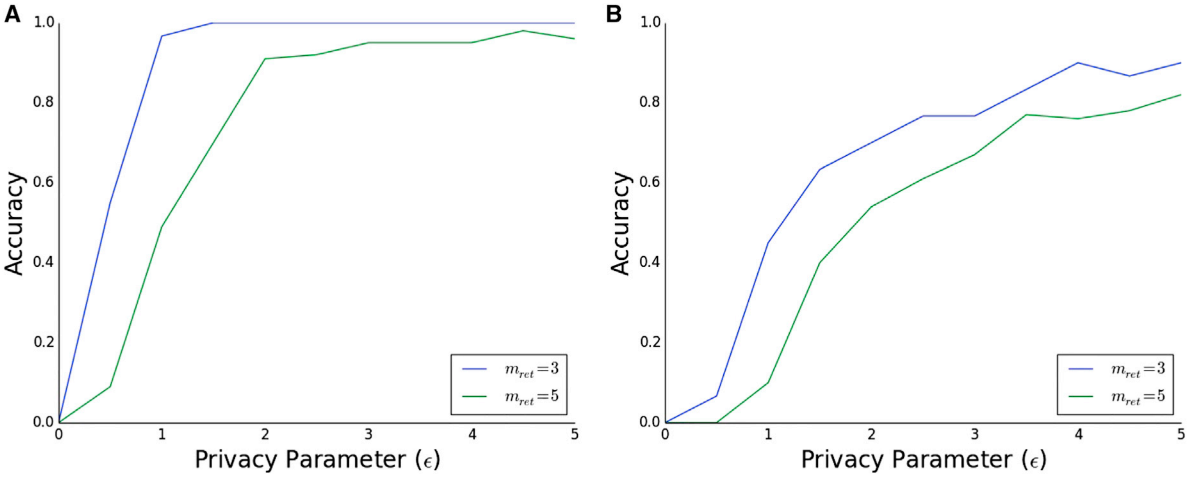


Figure 2. Accuracy of the PrivSTRAT Algorithms for Selection of Top SNPs

(A and B) Percentage of the top SNPs correctly returned for the PrivSTRAT algorithms, with m_{ret} (the number of SNPs being returned) equal to 3 and 5 for (A) the rheumatoid arthritis GWAS dataset and (B) our simulated dataset and varying values of the privacy parameter ϵ. In all four cases, accuracy increases as privacy decreases (as ϵ increases). More importantly, in three of the four cases, high accuracy is achieved for a reasonable level of privacy (ϵ around 2), which should increase with sample size. These results are averaged over 20 iterations.

we show that, with a few algorithmic insights, we are able modify this approach to work for the more complicated EIGENSTRAT statistics. This result is notable, given that it was only recently shown that this approach could be made computationally tractable even for relatively simple allelic test statistics (Yu and Ji, 2014; Simmons and Berger, 2016).

Our algorithm for selecting highly associative SNPs takes a privacy parameter ϵ (see Box 1 for details) and the number of SNPs to be returned $\boldsymbol{m_{ret}}$. The algorithm returns $\boldsymbol{m_{ret}}$ SNPs in a way that ensures that the returned SNPs are almost equally likely to have been produced (more specifically, the likelihood differs by at most a multiplicative factor of $\exp(\epsilon)$) if we changed an individual's disease status in our dataset (which we denote as ϵ-phenotypic differential privacy) (see Box 1 and Experimental Procedures) while maximizing the number of returned high-scoring SNPs.

We tested the accuracy of PrivSTRAT for selecting high-scoring SNPs and found we can obtain high accuracy for reasonable levels of privacy (ϵ around 1 or 2) (Figure 2). More specifically, we used the algorithm to return the top $\boldsymbol{m_{ret}}$ SNPs, where $\boldsymbol{m_{ret}} \in \{3, 5\}$ for various values of the privacy parameter ϵ. (This choice is based on previous work by Yu et al., 2014a. Other values are explored in Supplemental Experimental Procedures.) The accuracy of the returned results (averaged over 20 trials) is measured by the percentage overlap between the returned results and the true results (Yu et al., 2014a). For both the rheumatoid arthritis dataset (Figure 2A) and the simulated dataset (Figure 2B), accuracy increases as ϵ increases (privacy decreases), as expected. Moreover, near perfect accuracy is obtained for realistic values of ϵ (values around 1 or 2) (Vinterbo et al., 2012) on real GWAS data, which will only increase as datasets grow (Supplemental Experimental Procedures). We also performed experiments on datasets with higher levels of population stratification (Figure S3; Supplemental Experimental Procedures). In particular, we took data from the HapMap project (consisting of 880 individuals of different ancestries) and simulated corresponding phenotype data. We see that our method for picking high-scoring SNPs has lower accuracy, but this is consistent with the decrease in sample size (Supplemental Experimental Procedures).

Our method requires the user to specify the number of returned SNPs ahead of time (the $\boldsymbol{m_{ret}}$ parameter). This differs from the traditional GWAS approach, in which the researcher sets a p value threshold and returns all SNPs with p values below that threshold. Our framework can be modified to work in this way. In particular, it is possible to estimate the number of SNPs with scores below a certain p value threshold in a privacy-preserving way using neighbor distance (Supplemental Experimental Procedures). However, we do not envision this as the main use of our method—it is meant to return not all possible hits but only the most promising handful (see Case Study).

By running our algorithm on subsets of the rheumatoid arthritis dataset (with n = 200, 400, 600, 800, and 1,000 participants), we see that accuracy increases as sample size increases (Table S2). This suggests that the utility of our method will increase as genomic databases become larger.

Runtime

To assess the effects of the privacy-preserving nature of PrivSTRAT on runtime, we ran PrivSTRAT on the rheumatoid arthritis (RA) dataset described in Experimental Procedures, with m_{ret} = 3, and measured the amount of time taken by each step of the algorithm: performing the singular value decomposition (SVD) using either the smartpca algorithm included in EIGENSTRAT (134.16 s) or an approximate method (14.37 s) (Supplemental Experimental Procedures), calculating the μ vectors (8.6 s), calculating the neighbor distance (26.23 s), and picking the SNPs (.25 s). The results are an average over 10 trials. The calculation of the exact SVD is the slowest of these steps by a factor of >5, while even the approximate SVD calculation is only a factor of

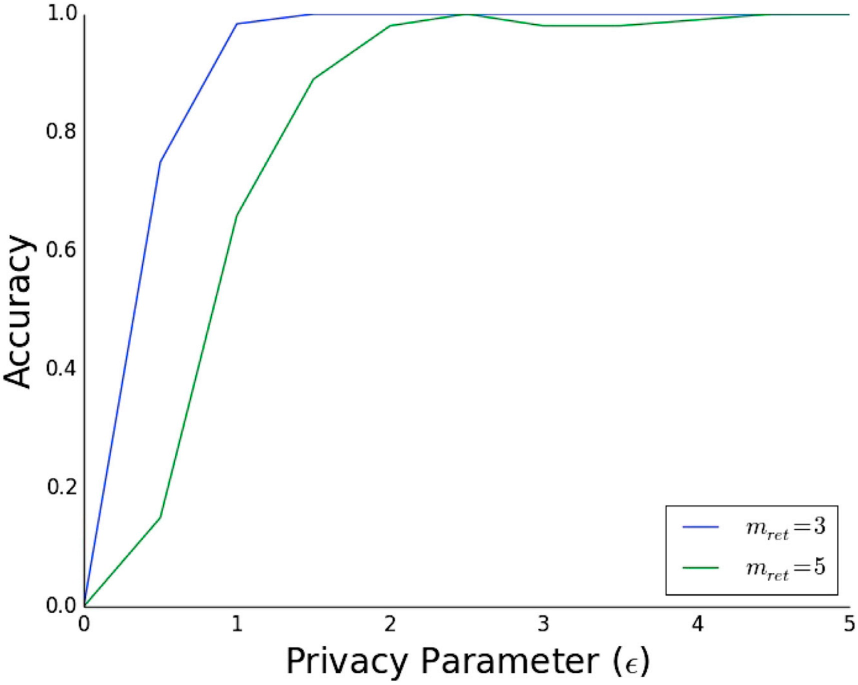


Figure 3. Accuracy of the PrivLMM Method for Selection of Top SNPs

Percentage of top SNPs correctly returned for the PrivLMM method, with m_{ret} (the number of SNPs being returned) equal to 3 and 5 for the rheumatoid arthritis GWAS dataset and varying values of the privacy parameter ϵ. In both cases, high accuracy is achieved for a reasonable level of privacy (ϵ around 2). These results are averaged over 20 iterations.

2 faster than the slowest step in the privacy-preserving algorithm. Because calculation of the SVD is required by standard EIGENSTRAT statistics, the overhead required to preserve privacy is minimal.

Extending to LMMs: PrivLMM

We have thus far focused on performing privacy-preserving GWASs based on EIGENSTRAT statistics. In recent years, however, there has been growing interest in using statistics based on LMMs to perform GWASs. The framework introduced earlier can be used to perform privacy-preserving LMM-based analysis, a method we denote PrivLMM (Supplemental Experimental Procedures).

To demonstrate this application, we tested PrivLMM for returning high-scoring SNPs on the rheumatoid arthritis GWAS dataset. We used the same setup that we used for PrivSTRAT. As expected, privacy decreases as accuracy increases (Figure 3). High accuracy can be obtained for reasonable levels of privacy (ϵ around 1 or 2) (Figure 3). We used values for the variance parameters ($\boldsymbol{\sigma_e}$ and $\boldsymbol{\sigma_g}$) calculated using FaST-LMM software (Lippert et al., 2011). In theory, it is preferable to use a differentially private approach to calculate these parameters. A method to do this, based on previous work (Abowd et al., 2013), is described in Supplemental Experimental Procedures.

Case Study

Next, we provide a case study to illustrate one possible way PrivLMM and PrivSTRAT might be used in practice. Consider a lab with limited resources that wants to run a GWAS on a group of individuals. The study will result in a list of high-scoring SNPs, many of which are significant. However, many SNPs may be close to the threshold of being significant. Some of these close calls are due to chance, while some may be SNPs weakly associated with the disease, a particular concern in small studies. The researchers might be interested in assessing several of these SNPs close to the significance boundary for significance on a new, larger dataset as validation. Unfortunately, they may not have direct access to such datasets due to privacy concerns. Our method could allow these researchers access to the databases necessary to validate their results. Such validation is of particular interest in light of the large number of false-positive results that appear in the biomedical literature (Kohane et al., 2012; Ioannidis, 2005).

To demonstrate this utility on a test dataset, we divided the rheumatoid arthritis dataset into two subsets of 450 cases and 450 controls (dataset 1) and the remaining 435 cases and 780 controls (dataset 2). We ran a GWAS on dataset 1 to obtain several significant SNPs below the p value cutoff of 10^{-6} (corresponding to a Bonferroni corrected p value of 0.05). Two additional SNPs, rs498422 ($p = 2.10\times10^{-6}$) and rs9419011 ($p = 1.19\times10^{-6}$), do not quite reach the p value cutoff. Because they are close to reaching statistical significance, we would like to test them on dataset 2 to see whether follow-up studies might be worthwhile. Due to privacy concerns, we are not given direct access to this database. However, we can apply PrivSTRAT with a total privacy budget of $\epsilon = 2$. This gives us an estimate of the EIGENSTRAT statistics for both SNPs, with scores of 25.79 and 1.66 (estimated 95% confidence intervals of 4.21–95.60 and 0–13.75, corresponding to p values of $\sim5\times10^{-7}$ and $\sim$0.19, respectively).

Because we are only testing two SNPs, even after correcting for multiple-hypothesis testing, this is enough to suggest that rs498422 might be worth further investigation. This result is consistent with previous findings on rheumatoid arthritis: rs498422 is close to the human leukocyte antigen locus in the human genome, a region known to be highly associated with rheumatoid arthritis risk. However, this result does not support further study of rs9419011, saving us time and effort. In both cases, the PrivSTRAT statistics are close to the actual EIGENSTRAT statistics (which equal 26.18 and 1.69, respectively). We obtain similar accuracy when repeating this experiment (Supplemental Experimental Procedures).

DISCUSSION

Here, we introduce three advances that make differential privacy useful for real-world GWAS statistics. First, we offer a modified yet practical form of differential privacy, termed phenotypic differential privacy (Box 1), with the aim of efficiently protecting private disease status information from being leaked while accurately answering various common queries on genomic data. This definition does not guarantee that information about whether someone participated in our study is hidden (though it also does not guarantee that such information will be leaked). Instead, it prevents the release of private information that can be used to compromise a patient's disease status using genotype information or can be used to compromise a patient's genotype data using disease information. With EHRs or large genomic databases (such as 23andMe), knowing that someone participated is equivalent to knowing they have their genotype on record, a fact that is unlikely to be private. As such, it makes sense to use phenotypic differential privacy in such settings.

Second, we introduce decompositions of EIGENSTRAT and LMM-based statistics that allow us to use a tool from differential privacy, the Laplacian mechanism (Dwork, 2011), to obtain accurate and fast estimations of the statistical significance of specific SNPs while preserving patient privacy.

Third, we develop a greedy algorithm that allows us to return lists of SNPs highly associated with a disease while ensuring high levels of both accuracy and privacy. This result is particularly noteworthy, because analogous methods for simpler statistics have been devised only recently (Johnson and Shmatikov, 2013; Simmons and Berger, 2016). Combined, our tools demonstrate that it is possible to correct for population stratification while preserving privacy in GWAS results, thus offering the possibility of applying a differentially private framework to large, genetically diverse groups of individuals and patients, such as those present in large genomic databases.

The major computational bottleneck in our methods comes not from the privacy-preserving component but from the original statistics (calculating the eigenvectors in EIGENSTRAT or calculating the variance parameters in LMM-based statistics). As such, our methods are well positioned to exploit computational advances in GWAS analysis. In particular, we are interested in modifying our method to take advantage of computational advances recently introduced for LMM-based association (Loh et al., 2015). (See Supplemental Experimental Procedures for other potential directions.)

We are advocating privacy-preserving methods not for all situations in which one might want to conduct a GWAS but rather for only situations in which privacy concerns would make alternative approaches cumbersome or impossible. It is our hope that our Priv suite of tools will be used to improve access to private genomics data. This access will offer researchers new tools that can be used to produce novel hypotheses or validate previous findings in ways that are not currently possible due to privacy concerns.

As with any set of tools, it is important to understand the limitations of the Priv suite. Although our tools are useful for answering questions about large databases while preserving privacy, they are less accurate on small databases (discussed in Experimental Procedures). Even on large databases, while our approach performs well in many circumstances, greater accuracy is desirable. Understanding exactly where our method is most useful will require tests on a large variety of datasets in numerous application domains. Moreover, our use of phenotypic differential privacy cannot guarantee privacy in databases with large levels of case ascertainment (that is, when the percentage of individuals with the disease in the study is larger than the percentage in the background population) but is instead focused on databases that are representative of the background population (such as in 23andMe or similar databases). It is our hope that future work will build upon our results to overcome these limitations.

In the long term, it is possible that differential privacy techniques will no longer be needed as we come to understand exactly how much privacy is lost after releasing aggregate genomic data (Simmons and Berger, 2015). Currently, we are far from this understanding. Thus, differential privacy provides us with the possibility of granting wider access to genomic data now, with immediate benefits for the research community.

Availability

An implementation of our results and simulated data is available on our website, http://groups.csail.mit.edu/cb/PrivGWAS, and on the *Cell Systems* website (Data S1).

EXPERIMENTAL PROCEDURES

Notation

In the following sections, we use $|v|$ to denote the length of the vector v. Moreover, for vectors u and v, we let $u \cdot v$ denote the dot product of u with v.

GWASs Revisited

The aim of GWASs is to link SNPs in a study cohort to a phenotype (e.g., disease) of interest. In a GWAS, the researcher begins with a group of n individuals genotyped at m SNPs. Let D be an n by m genotype matrix, where the ith entry in the jth row of D is equal to the number of times the minor allele occurs in the jth individual at the ith SNP (for autosomal SNPs, this number is equal to 0, 1, or 2). Details on how to handle missing genotypes are provided in Supplemental Experimental Procedures. Let X be the n by m matrix obtained by mean centering and variance normalizing each column of the genotype matrix D. Let x_i be the column of X corresponding to SNP i. Similarly, let $y = (y_1, \ldots, y_n) \in \{0, 1\}^n$ be a vector of phenotypes, where $y_j = 1$ if the jth individual has the disease and $y_j = 0$ otherwise.

Given X and y, we would like to determine which SNPs are associated with the disease phenotype. We focus on two statistics that allow us to test for these associations: EIGENSTRAT (Price et al., 2006) and LMM-based association statistics (Kang et al., 2010).

Phenotypic Differential Privacy

Here, we give a formal definition of phenotypic differential privacy. See Box 1 for a more informal discussion.

Definition 1. Let F be a random function that takes an n by m genotype matrix, D, and an n dimensional phenotype vector, y, and outputs F(D, y), where the output is in some set Ω. We say that F is ϵ-phenotypic differential privacy for some privacy parameter $\epsilon > 0$ if, for all genotype matrices D, all phenotype vectors $y, y' \in \{0, 1\}$, such that y and y′ differ in exactly one coordinate, and for all sets $S \subset \Omega$, we have

$$P(F(D, y) \in S) \leq \exp(E) P(F(D, y') \in S).$$

This definition of privacy can be viewed as a specific instantiation of both induced differential privacy (Kifer and Machanavajjhala, 2011) and the Blowfish framework (Kifer and Machanavajjhala, 2014; Hi et al., 2014).

PrivSTRAT: Privacy-Preserving EIGENSTRAT

The differentially private GWAS literature has largely focused on three tasks: identifying highly associated SNPs, estimating association statistics, and estimating the number of significantly associated SNPs in a study. We consider all three tasks, focusing on the first two (the third is addressed in Supplemental Experimental Procedures).

Estimating χ^2

We would like to estimate the $\boldsymbol{\chi}^2$ statistic from EIGENSTRAT (Supplemental Experimental Procedures). In particular, assume we want an estimate of the EIGENSTRAT statistic $\boldsymbol{\chi}_{\boldsymbol{i}}^2$ for a given SNP i. To do this, if we let $\mu_i = x_i^* / \left| x_i^* \right|$, then

$$\boldsymbol{\chi}_{\boldsymbol{i}}^2 = \frac{(\boldsymbol{n}-\boldsymbol{k}-1)(\boldsymbol{\mu_i} \cdot \boldsymbol{y}^*)^2}{|\boldsymbol{y}^*|^2} = \frac{(\boldsymbol{n}-\boldsymbol{k}-1)(\boldsymbol{\mu_i} \cdot \boldsymbol{y})^2}{|\boldsymbol{y}^*|^2}.$$

Therefore, it suffices to get estimates of both $\boldsymbol{\mu_i} \cdot \boldsymbol{y}$ and $|\boldsymbol{y}^*|$ that are $(\epsilon/2)$-phenotypic differential privacy and combine the results ($\mu_i \cdot y = \mu_i \cdot y^*$ follows, because y^* is the projection of y onto a linear subspace containing μ_i). This can be done easily using the Laplacian mechanism (Dwork, 2011). More details are in Supplemental Experimental Procedures.

Selecting High-Scoring SNPs

Another task we consider is returning a list of the top m_{ret}-scoring SNPs for some user-defined parameter m_{ret} while achieving ϵ-phenotypic differential privacy—in other words, we want to return the locations of the m_{ret} SNPs with the largest $\boldsymbol{\chi}^2$ values. This is equivalent to picking the m_{ret} SNPs with the largest $|\boldsymbol{\mu_i} \cdots \mathrm{y}|$ values. To do this in a privacy-preserving way, we use a modified version of the approach known as the distance-based method (Johnson and Shmatikov, 2013). This works as follows: the user chooses a threshold c > 0. The i^th SNP is considered significant if $|\boldsymbol{\mu_i} \cdot \boldsymbol{y}| > \mathrm{c}$ and not significant otherwise. (For example, c might correspond to a p value of 0.05 or 10^{-8}. In practice, instead of having the user choose c, we use a previous approach to automatically choose c [Simmons and Berger, 2016]. Details are given in Supplemental Experimental Procedures.) The i^th SNP, denoted b_i, is the minimum number of individuals whose phenotypes need to be changed to change SNP i from significant to not, or vice versa. Formally,

$$b_i = b_i(c) = min_{y' \in [0,1]^n,\ c = \left| \mu_i \cdot y' \right|} |y - y'|_0,$$

where $|v|_0$ denotes the number of nonzero entries in the vector v, and $b_i = \min\{d_i(c), d_i(-c)\}$, where

$$d_i(c) = min_{y' \in [0,1]^n,\ c = \mu_i \cdot y'} |y - y'|_0.$$

To use this neighbor distance to select high-scoring SNPs, we let $d_i^* = b_i$ for significant SNPs and $d_i^* = 1 - b_i$ for all other SNPs. The distance-based method picks m_{ret} SNPs without repetition, where the probability of picking the i^th SNP is proportional to exp $(d_i^* / 2\epsilon m_{ret})$. It is easy to see from previous work (Johnson and Shmatikov, 2013) that this mechanism is ϵ-phenotypic differential privacy. The difficult part is calculating $d_i(c)$. Our significant algorithmic development is to show that this can be done using the greedy algorithm presented in Algorithm 1.

Data

We test PrivSTRAT and PrivLMM on a rheumatoid arthritis dataset, NARAC-1 (Plenge et al., 2007). After quality control filtering, the dataset contained 893 cases, 1,243 controls, and 67,623 SNPs. This dataset includes some closely related individuals. Although LMM can handle such cryptic relatedness, EIGENSTRAT is not designed to do so. Therefore, before applying PrivSTRAT to this dataset, we used PLINK to remove relatives with an estimated identity by descent greater than 0.2. Thus, the final dataset contained 885 cases and 1,230 controls. Because this dataset has relatively little population stratification, we also used PrivSTRAT on a simulated dataset with two subpopulations. This dataset and the associated code (based on PLINK tools) (Purcell et al., 2007) are available online.

Algorithm 1: Calculates the Neighbor Distance

```
Require: y; $\mu_i$; c.
Ensure: Returns the neighbor distance $d_i(c)$.
    Let $i_1, \ldots, i_n$ be a permutation on $1, \ldots, n$ such that, if
        $\boldsymbol{u_r} = \boldsymbol{max}\{\boldsymbol{\mu_{i,i_r}}(1 - \boldsymbol{y_{i_r}}), \boldsymbol{\mu_{i,i_r}}(0 - \boldsymbol{y_{i_r}})\},$
    then $u_1 \geq u_2 \geq \cdots \geq u_n$.
    Let $j_1, \ldots, j_n$ be a permutation on $1, \ldots, n$ such that, if
        $\boldsymbol{l_r} = \boldsymbol{min}\left\{\boldsymbol{\mu_{i,j_r}}\left(1 - \boldsymbol{y_{j_r}}\right), \boldsymbol{\mu_{i,j_r}}\left(0 - \boldsymbol{y_{j_r}}\right)\right\},$
    then $l_1 \leq l_2 \leq \ldots \leq l_n$.
    Let $U_r = \Sigma_{j=1}^{k} u_j$ and $L_r = \Sigma_{j=1}^{k} l_j$ for $k = 1, \cdots, n$.
    Return r such that $c - \mu_i \cdot y \in [L_{r+1}, L_r) \cup (U_r, U_{r+1}]$.
```

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, two tables, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.04.013.

AUTHOR CONTRIBUTIONS

S.S. developed the algorithms with help from B.B., S.S. performed the experiments, and S.S. and B.B. evaluated the results. B.B. and C.S. supervised the research. S.S., B.B., and C.S. contributed to writing the manuscript.

ACKNOWLEDGMENTS

S.S. and B.B. are partly supported by NIH GM108348; S.S. and C.S. are supported by the NSERC Discovery Frontiers Program project: The Cancer Genome Collaboratory. We thank J. Bienkowska for providing us with data access, and the NARAC for performing the initial data collection. We also thank N. Daniels, J. Peng, and other members of the B.B. lab for useful discussions. We are grateful to the reviewers and editors for many helpful comments, R. Daniels for invaluable advice on the manuscript, and L. Gaffney for design of the graphical abstract. An early version of this paper was submitted to and peer reviewed at the 2016 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at *Cell Systems*.

Received: February 5, 2016
Revised: April 8, 2016
Accepted: April 17, 2016
Published: July 21, 2016

REFERENCES

Abowd, J.M., Schneider, M.J., and Vilhuber, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. J. Priv. Confid. *5*, 73–105.

Chen, R., Peng, Y., Choi, B., Xu, J., and Hu, H. (2014). A private DNA motif finding algorithm. J. Biomed. Inform. *50*, 122–132.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.

Dwork, C. (2011). Differential privacy. In Encyclopedia of Cryptography and Security, H.C.A. van Tilborg and S. Jajodia, eds. (Springer), pp. 338–340.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography, S. Halevi and T. Rabin, eds. Lecture Notes in Computer Science, Volume *3876* (Springer), pp. 265–284.

Erlich, Y., and Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. *15*, 409–421.

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. (2014). Privacy in pharmacogenetics: an end-to-end case study of personalized Warfarin dosing. Proceedings of the 23rd USENIX Security Symposium, 17–32.

Furlotte, N.A., and Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. Genetics *200*, 59–68.

Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., and Erlich, Y. (2013). Identifying personal genomes by surname inference. Science *339*, 321–324.

Hi, X., Machanavajjhala, A., and Ding, B. (2014). Blowfish privacy: tuning privacy-utility trade-offs using policies. Proceedings of the ACM SIGMOD International Conference on Management of Data, 1447–1458.

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 Genet. *4*, e1000167.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B., and Roth, A. (2014). Differential privacy: an economic method for choosing epsilon. IEEE 27th Computer Security Foundations Symposium, 398–410.

Im, H.K., Gamazon, E.R., Nicolae, D.L., and Cox, N.J. (2012). On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am. J. Hum. Genet. *90*, 591–598.

Ioannidis, J.P. (2005). Why most published research findings are false. PLoS Med. *2*, e124.

Jiang, X., Zhao, Y., Wang, X., Malin, B., Wang, S., Ohno-Machado, L., and Tang, H. (2014). A community assessment of privacy preserving techniques for human genomes. BMC Med. Inform. Decis. Mak. *14* (*Suppl 1*), S1.

Johnson, A., and Shmatikov, V. (2013). Privacy-preserving data exploration in genome-wide association studies. KDD *2013*, 1079–1087.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.

Kifer, D., and Machanavajjhala, A. (2011). No free lunch in data privacy. Proceedings of the ACM SIGMOD International Conference on Management of Data, 193–204.

Kifer, D., and Machanavajjhala, A. (2014). Pufferish: a framework for mathematical privacy definitions. ACM Transactions on Database Systems *39*, 1–36.

Kohane, I.S., Hsing, M., and Kong, S.W. (2012). Taxonomizing, sizing, and overcoming the incidentalome. Genet. Med. *14*, 399–404.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.

Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.

Lowe, H.J., Ferris, T.A., Hernandez, P.M., and Weber, S.C. (2009). STRIDE—an integrated standards-based translational research informatics platform. AMIA Annu. Symp. Proc. *2009*, 391–395.

Lumley, T., and Rice, K. (2010). Potential for revealing individual-level information in genome-wide association studies. JAMA *303*, 659–660.

Malin, B.A., Emam, K.E., and O'Keefe, C.M. (2013). Biomedical data privacy: problems, perspectives, and recent advances. J. Am. Med. Inform. Assoc. *20*, 2–6.

Marchini, J., Cardon, L.R., Phillips, M.S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. Nat. Genet. *36*, 512–517.

Murphy, S.N., Gainer, V., Mendis, M., Churchill, S., and Kohane, I. (2011). Strategies for maintaining patient privacy in i2b2. J. Am. Med. Inform. Assoc. *18* (*Suppl 1*), i103–i108.

Nyholt, D.R., Yu, C.E., and Visscher, P.M. (2009). On Jim Watson's APOE status: genetic information is hard to hide. Eur. J. Hum. Genet. *17*, 147–149.

Plenge, R.M., Seielstad, M., Padyukov, L., Lee, A.T., Remmers, E.F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L.R., et al. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. N. Engl. J. Med. *357*, 1199–1209.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Sankararaman, S., Obozinski, G., Jordan, M.I., and Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. Nat. Genet. *41*, 965–967.

Shringarpure, S.S., and Bustamante, C.D. (2015). Privacy risks from genomic data-sharing beacons. Am. J. Hum. Genet. *97*, 631–646.

Simmons, S., and Berger, B. (2015). One size doesn't fit all: measuring individual privacy in aggregate genomic data. IEEE Security and Privacy Workshops, 41–49.

Simmons, S., and Berger, B. (2016). Realizing privacy preserving genome-wide association studies. Bioinformatics, Published online January 14, 2014. http://dx.doi.org/10.1186/1472-6947-15-S5-S2.

Tramer, F., Huang, Z., Habeaux, J.P., and Ayday, E. (2015). Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1286–1297.

Tucker, G., Price, A.L., and Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. Genetics *197*, 1045–1049.

Uhlerop, C., Slavković, A., and Fienberg, S.E. (2013). Privacy-preserving data sharing for genome-wide association studies. J. Priv. Confid. *5*, 137–166.

Vinterbo, S.A., Sarwate, A.D., and Boxwala, A.A. (2012). Protecting count queries in study design. J. Am. Med. Inform. Assoc. *19*, 750–757.

Wasserman, L., and Zhou, S. (2010). A statistical framework for differential privacy. J. Am. Stat. Assoc. *105*, 375–389.

Weber, G.M., Murphy, S.N., McMurry, A.J., Macfadden, D., Nigrin, D.J., Churchill, S., and Kohane, I.S. (2009). The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J. Am. Med. Inform. Assoc. *16*, 624–630.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.

Yu, F., and Ji, Z. (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. BMC Med. Inform. Decis. Mak. *14* (*Suppl 1*), S3.

Yu, F., Fienberg, S.E., Slavković, A.B., and Uhler, C. (2014a). Scalable privacy-preserving data sharing methodology for genome-wide association studies. J. Biomed. Inform. *50*, 133–141.

Yu, F., Rybar, M., Uhler, C., and Fienberg, S.E. (2014b). Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In Privacy in Statistical Databases, J. Domingo-Ferrer, ed. Lecture Notes in Computer Science, Volume 8744 (Springer), pp. 170–184.

Zhang, J., Xiao, X., Yang, Y., Zhang, Z., and Winslett, M. (2013). PrivGene: differentially private model fitting using genetic algorithms. Proceedings of the ACM SIGMOD International Conference on Management of Data, 665–676.

Zhao, Y., Wang, X., Jiang, X., Ohno-Machado, L., and Tang, H. (2015). Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery. J. Am. Med. Inform. Assoc. *22*, 100–108.

Zhou, X., Peng, B., Li, Y.F., Chen, Y., Tang, H., and Wang, X. (2011). To release or not to release: evaluating information leaks in aggregate human-genome data. Proceedings of the 16th European Conference on Research in Computer Security, 607–627.