Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2017 July 21.

Published in final edited form as:

Cell Syst. 2016 July ; 3(1): 54–61. doi:10.1016/j.cels.2016.04.013.

Enabling Privacy-Preserving GWAS in Heterogeneous Human Populations

Sean Simmons[1,2,3], **Cenk Sahinalp**[3,4], and **Bonnie Berger**[1,2,*]

[1]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA

[3]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada

[4]School of Informatics and Computing, Indiana University, Bloomington, IN

Summary

The proliferation of large genomic databases offers the potential to perform increasingly larger-scale genome-wide association studies (GWAS). 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 GWAS that adapts principles of differential privacy—a cryptographic theory that facilitates secure analysis of sensitive data—to, for the first time, 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.

Graphical abstract

*to whom correspondence should be addressed: bab@mit.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

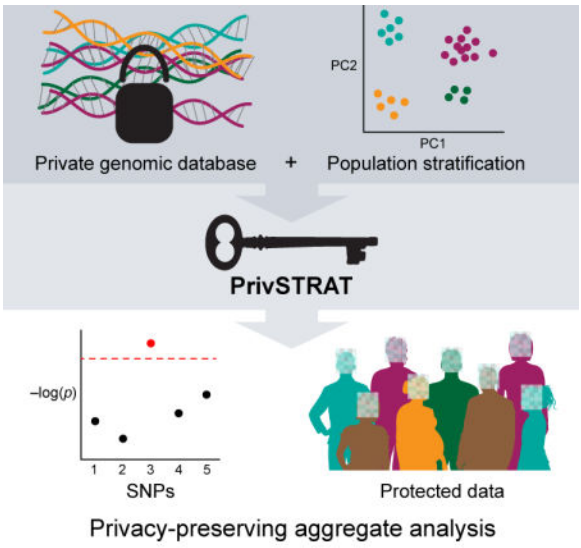
Author Contributions

S.S. and B.B. developed the algorithms, S.S. performed the experiments, and both evaluated the results. B.B. and C.S. supervised the research. S.S., B.B., and C.S. all contributed to writing the manuscript.

Supplementary Information

We have included supplementary methods detailing some technical details of our methods and proofs of correctness. We have also included a zipped file containing data and code (Data S1).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Introduction

We are experiencing an 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 (Murphy et al., 2011). 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 instead may benefit from the development of cryptographic tools that allow secure access while also ensuring accurate analyses. In particular, there has been increased interest in the usefulness of a cryptographic technique known as differential privacy (Dwork, 2011) to allow researchers access to genomic data (Jiang et al., 2014; Johnson and Shmatikov, 2013; Uhler et al., 2013; Yu et al., 2014; Yu and Ji, 2014; Chen et al., 2014; Zhao et al, 2015; Zhang et al, 2013) while preserving every patient's privacy. 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 (GWAS) (Weber et al., 2009; Lowe et al., 2009), which are commonly used to identify single nucleotide polymorphisms (SNPs) associated with a given disease. Numerous works (Homer et al., 2008; Lumley and Rice, 2010; Im et al., 2012; Zhou et al., 2011.; Sankararaman et al., 2009) have shown that aggregate genomic data, including GWAS statistics, can leak private information about participants. These findings have led the NIH, among others, to place much of their aggregate genomic data into repositories and require researchers to apply for access (Erlich and Narayanan, 2014). Recent work (Shringarpure and Bustamante, 2015) has also shown that a popular method for sharing genomic data, "Genomic Data-Sharing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Beacons," actually leaks potentially private information about participants. Their 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., 2014; Johnson and Shmatikov, 2013; Uhler 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 GWAS (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 type of genomic analyses that can be performed accurately and efficiently (Fredrikson et al, 2014).

Privacy concerns are not the only hurdles facing modern GWAS. Such analyses are complicated by systematic differences between 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). In order 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), genomic control (GC) (Devlin and Roeder, 1999), etc. In recent years, there has been a growing interest in using LMMs for this task, thanks to improved algorithms (Kang et al, 2010; Lippert et al., 2011; Tucker et al., 2014; Loh et al., 2015; Furlotte and Eskin, 2015). Even still, EIGENSTRAT remains a common approach for correcting for population stratification. However, previous works on differentially private GWAS 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: (i) GWAS statistics at SNPs of interest, and (ii) 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 GWAS. We demonstrate this approach on two state-of-the-art statistics, EIGENSTRAT (Price et al., 2006) and linear mixed model (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 both in terms of accuracy and runtime; indeed their accuracy improves as sample sizes increase.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 (dbGaP) or 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 also 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. In our case, since 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 the 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 one 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). Note that 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

As privacy methods do not produce correct results on typical GWAS studies without population stratification (Supplemental Experimental Procedures, "No Stratification"), we start by looking at the most basic queries in GWAS, 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 (Supplemental Experimental Procedures) while preserving privacy. As detailed in the Experimental Procedures, this is achieved using a modified version of the Laplacian mechanism (Dwork, 2011).

We studied the tradeoff between privacy (the ε parameter) and accuracy on real GWAS data for a rheumatoid arthritis dataset (Plenge et al., 2007) (Figure 1A), as well as on simulated data with two subpopulations (Figure 1B) (Experimental Procedures). As expected, our method quickly increases in accuracy as privacy decreases (i.e, as ε increases). Note that, for reasonable values of ε (ε around 1 or 2 [Vinterbo et al., 2012]) the median error introduced by our method is around .1 or .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 (Case Study).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Note that we can also calculate a phenotypically differentially private p-value for a given SNP using this returned statistic (Supplementary Experimental Procedures.)

Privately Selecting Highly Associated SNPs

Besides calculating χ^2 statistics, users may also be interested in determining which SNPs are most highly correlated with a given disease. A simple way to do this identification would be to use the above method to estimate the EIGENSTRAT statistic for all SNPs and returning the highest-scoring SNPs. The large number of queries required to do this, however, necessitates a small ε parameter for each query, resulting in poor accuracy. Much more accurate methods have been proposed for identification of high-scoring SNPs while also preserving privacy (Uhler et al., 2013; Johnson and Shmatikov, 2013; Yu et al., 2014). 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 the Supplemental Experimental Procedure, 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 the Pearson or allelic test statistics. Here we show that, with a few algorithmic insights, we are able modify this approach to work for the much more complicated EIGENSTRAT statistic. This result is notable, given that it was only recently shown that this approach could be made computationally tractable even for the relatively simple allelic test statistic (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 (m_{ret}). The algorithm returns 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 $\exp(\varepsilon)$) if we changed one individuals disease status in our dataset (which we denote as ε - phenotypic differential privacy, see Box 1 and the Experimental Procedure), 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 m_{ret} SNPs, where $m_{ret} \in \{3, 5\}$ (this choice is based off previous work (Yu et al., 2014). Other values are explored in the Supplemental Experimental Procedures), for various values of the privacy parameter, ε. 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., 2014). For both the rheumatoid arthritis (Figure 2A) and simulated (Figure 2B) data sets, 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 Procedure). 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,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

but this is consistent with the decrease in sample size (Supplemental Experimental Procedures, “Sample Size vs Accuracy”)

Our method requires the user to specify the number of returned SNPs ahead of time (the 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). We do not, however, envision this as the main use of our method–it is not meant to return all possible hits, but only the most promising handful (see the Case Study).

By running our algorithm on subsets of the rheumatoid arthritis dataset (with n=200, 400, 600, 800, and 1000 participants), we see that accuracy increases as sample size increases (Table S2), which 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 RA dataset described in the Experimental Procedure with $m_{ret} = 3$, and measured the amount of time taken by each step of the algorithm: calculate the SVD (using either the smartpca algorithm included in EIGENSTRAT (134.16 seconds) or an approximate method (14.37 seconds), see the Supplemental Experimental Procedures), calculating the μ vectors (8.6 seconds), calculating the neighbor distance (26.23 seconds), and picking the SNPs (.25 seconds). 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 2 faster than the slowest step in the privacy-preserving algorithm. Since calculating the SVD is required by the standard EIGENSTRAT statistic, the overhead required to preserve privacy is minimal.

Extending to LMMs: PrivLMM

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

To demonstrate this application, we tested PrivLMM for returning high-scoring SNPs on the rheumatoid arthritis GWAS dataset. We used the same set-up that we used for PrivSTRAT. As expected, privacy decreases as accuracy increases (Figure 3). High accuracy can be obtained for reasonable levels of privacy ($\boldsymbol{\varepsilon}$ around 1 or 2) (Figure 3). Note that we used values for the variance parameters ($\boldsymbol{\sigma_e}$ and $\boldsymbol{\sigma_g}$) calculated using the FaST-LMM (Lippert et al., 2011) software. 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 study on a group of individuals. The study will result in a list of high-scoring SNPs, many of which are significant. There may, however, be many SNPs that are 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 concern particularly in smaller 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, however, 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 cut off of $\mathbf{10^{-6}}$ (corresponding to a Bonferroni corrected p-value of .05). Two additional SNPs, rs498422 (P= $\mathbf{2.10 \times 10^{-6}}$) and rs9419011 (P= $\mathbf{1.19 \times 10^{-6}}$), do not quite reach the p-value cutoff. Since they are close to reaching statistical significance, we would like to test them both on dataset 2 to see if they might be worth further follow-up studies. Due to privacy concerns, however, we are not given direct access to this database. We can, however, apply PrivSTRAT with a total privacy budget of $\boldsymbol{\varepsilon} = 2$. This gives us an estimate of the EIGENSTRAT statistic for both of these SNPs, with scores of 25.79 and 1.66 (estimated 95% CI 4.21–95.60 and 0–13.75), respectively (corresponding to p-values of $\mathbf{\sim 5 \times 10^{-7}}$ and ~0.19).

Since 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. Note that this result is consistent with previous findings on rheumatoid arthritis: rs498422 is close to the HLA locus in the human genome, a region known to be highly associated with rheumatoid arthritis risk. We see that, on the other hand, this result does not support further study of rs9419011, saving us possible wasted time and effort. Note that, in both cases, the PrivSTRAT statistics are very close to the actual EIGENSTRAT statistics (which equal 26.18 and 1.69, respectively). Note that we obtain similar accuracy when repeating this experiment (Supplemental Experimental Procedure, “Validation”).

Discussion

Here we introduce three key 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 also accurately answering various common queries on genomic data. Notably, this definition does not guarantee that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

information about whether or not 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 used to compromise a patient's genotype data using disease information. With electronic health record (EHR) 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 also 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 since analogous methods for much simpler statistics have only recently been devised (Johnson and Shmatikov, 2013; Simmons and Berger, 2016). Combined, our tools demonstrate that it is possible to correct for population stratification while also 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 instead from the original statistics (calculating the eigenvectors in EIGENSTRAT or calculating the variance parameters in the LMM-based statistic). 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 the computational advances introduced recently (Loh et al., 2015) for LMM-based association. (See Supplementary Materials for other potential directions."

Note that we are not advocating privacy-preserving methods for all situations in which one might want to conduct a GWAS, but only when 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 much less accurate on small databases (discussed in Experimental Procedure). 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with the disease in the study is much 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 even 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, however, 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 below, 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.

GWAS Revisited

The aim of genome-wide association studies (GWAS) 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 either 0, 1, or 2). Details on how to handle missing genotypes are provided in Supplemental Experimental Procedure. 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, $y_j = 0$ otherwise.

Given X and y, we would like to determine which SNPs are associated with the disease phenotype. Here we will mainly 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

Below 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 in a 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 $\varepsilon > 0$ if, for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 that

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

Note that this definition of privacy can be viewed as a specific instantiation of both induced differential privacy (Kifer and Machanavajjhala, 2011) and of the Blowfish framework (Kifer and Machanavajjhala, 2014; He 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 mainly on the first two (the third is addressed in the Supplemental Experimental Procedure).

Estimating χ^2

We would like to estimate the χ^2 -statistic from EIGENSTRAT (Supplemental Experimental Procedure). In particular, assume we want an estimate of the EIGENSTRAT statistic, χ_i^2, for a given SNP i. In order to do this, note that if we let $\mu_i = x_i^* / |x_i^*|$, then:

$$\chi_i^2 = \frac{(n-k-1)(\mu_i \cdot y^*)^2}{|y^*|^2} = \frac{(n-k-1)(\mu_i \cdot y)^2}{|y^*|^2}$$

so it suffices to get estimates of both $\mu_i \cdot y$ and $|y^*|$ that are $\frac{\varepsilon}{2}$-phynotypic differential privacy and combine the results (the fact that $\mu_i \cdot y = \mu_i \cdot y^*$ follows since y^* is the projection of y onto a linear subspace containing μ_i). This can be done easily, however, using the Laplacian mechanism (Dwork, 2011). More details are in the Supplementary Experimental Procedure.

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 largest χ^2 values. This is equivalent to picking the m_{ret} SNPs with largest $|\mu_i \cdot y|$ values. In order 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 ith SNP is considered significant if $|\mu_i \cdot y| > c$, not significant otherwise (for example, c might correspond to a p-value of .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 the Supplemental Experimental Procedure). The neighbor distance for the ith 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:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

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

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

In order 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 ith SNP is proportional to $\exp(\frac{d_i^*}{2\varepsilon m_{ret}})$ It is easy to see from previous work (Johnson and Shmatikov, 2013) that this mechanism is $\boldsymbol{\varepsilon}$–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.

Algorithm 1: Calculates the neighbor distance

Require: y; μ_i; c

Ensure: Returns the neighbor distance, $d_i(c)$.

Let $i_1, \ldots, i_n$ be a permutation on 1, …, n such that, if

$$\boldsymbol{u_r = max}\left\{\mu_{i,i_r}(\mathbf{1} - \boldsymbol{y}_{i_r}), \mu_{i,i_r}(\mathbf{0} - \boldsymbol{y}_{i_r})\right\}$$

then $u_1 \geq u_2 \geq \cdots \geq u_n$.

Let $j_1, \ldots, j_n$ be a permutation on 1, …, n such that, if

$$\boldsymbol{l_r = min}\left\{\mu_{i,j_r}(\mathbf{1} - \boldsymbol{y}_{j_r}), \mu_{i,j_r}(\mathbf{0} - \boldsymbol{y}_{j_r})\right\}$$

then $l_1 \leq l_2 \leq \cdots \leq l_n$.

Let $U_r = \sum_{j=1}^{k} u_j$ and $L_r = \sum_{j=1}^{k} l_j$ for $k = 1, \cdots, n$

Return r such that $c - \mu_i \cdot y \in [L_{r+1}, L_r) \cup (U_r, U_{r+1}]$

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 and 1,243 controls, and a total of 67,623 SNPs. Note that this dataset includes some closely related individuals. Although LMM can handle such cryptic relatedness, EIGENSTRAT is not designed to do so. As such, before applying PrivSTRAT to this dataset, we used PLINK to remove relatives with estimated IBD greater than 0.2. Thus, the final dataset contained 885 cases and 1,230 controls. Since 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

S.S. and B.B. are partly supported by NIH GM108348; S.S. and C.S. by NSERC Discovery Frontiers Program project: The Cancer Genome Collaboratory. We would also like to thank J. Bienkowska for providing us with data access, and the NARAC for performing the initial data collection. Also, N. Daniels, J. Peng, and other members of the Berger 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.

References

Abowd J, Schneider M, Vilhuber L. Differential privacy applications to bayesian and linear mixed model estimation. Journal of Privacy and Condentiality. 2013; 5:73–105.

Bakker P, Yelensky R, Pe'er I, Gabriel S, Daly M, Altshuler D. Efficiency and power in genetic association studies. Nature Genetics. 2005; 37:1217–1223. [PubMed: 16244653]

Chen R, Pen Y, Choi B, Xu J, Hua H. A private DNA motif finding algorithm. JBI. 2014; 50:122–132.

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]

Dwork C. Differential privacy. Encyclopedia of Cryptography and Security. 2011:338–340.

Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. The reusable holdout: preserving validity in adaptive data analysis. Science. 2015; 349:636–638. [PubMed: 26250683]

Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. TCC '06. 2006:265–284.

Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nature Reviews Genetics. 2014; 15:409–421.

Fredrikson M, Lantz E, Jha E. Privacy in pharmacogenetics: an end-to-end case study of personalized Warfarin dosing. USENIX. 2014; 23:17–32.

Furlotte N, Eskin E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. Genetics. 2015; 200:59–68. [PubMed: 25724382]

Galinsky K, Bhatia G, Loh P, Georgiev S, Mukherjee S, Patterson N, Price A. Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. BioRxiv. 2015

Gymrek M, McGuire A, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013; 339:321–324. [PubMed: 23329047]

Gymrek M, Willems T, Zeng H, Markus B, Daly M, Price A, Pritchard J, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature Genetics. 2016; 48:22–29. [PubMed: 26642241]

He D, Furlotte N, Hormozdiari F, Joo J, Wadia A, Ostrovsky R, Sahai1 A, Eskin E. Identifying genetic relatives without compromising privacy. Genome Research. 2014; 24:664–672. [PubMed: 24614977]

Hi X, Machanavajjhala A, Ding B. Blowfish privacy: tuning privacy-utility trade-offs using policies. SIGMOD '14. 2014:1447–1458.

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J, Stephan D, Nelson S, Craig D. Resolving individual's contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008:4.

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

Im H, Gamazon E, Nicolae D, Cox N. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet. 2012; 90:591–598. [PubMed: 22463877]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Ioannidis J. Why Most Published Research Findings Are False. PLOS Medicine. 2005; 2:e124.doi: 10.1371/journal.pmed.0020124 [PubMed: 16060722]

Jiang X, Zhao Y, Wang X, Malin B, Wang S, Ohno-Machado L, Tang H. A community assessment of privacy preserving techniques for human genomes. BMC Medical Informatics and Decision Making. 2014; 14:S1. [PubMed: 25521230]

Johnson A, Shmatikov V. Privacy-preserving data exploration in genome-wide association studies. KDD '13. 2013:1079–1087.

Kang H, Sul J, Service S, Zaitlen N, Kong S, Freimer N, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–54. [PubMed: 20208533]

Kifer D, Machanavajjhala A. No free lunch in data privacy. SIGMOD '11. 2011:193–204.

Kifer D, Machanavajjhala A. Pufferish: A framework for mathematical privacy definitions. ACM Transactions on Database Systems. 2014; 39(1):3:1–3:36.

Kohane I, Hsing M, Kong S. Taxonomizing, sizing, and overcoming the incidentalome. Genetics in Medicine. 2012; 14:399–404. [PubMed: 22323072]

Lippert C, Listgarten J, Liu Y, Kadie C, Davidson R, Heckerman D. FaST linear mixed models for genome-wide association studies. Nature Methods. 2011; 8:833–835. [PubMed: 21892150]

Loh P, et al. Efficient Bayesian mixed model analysis increases association power in large cohorts. Nat Genet. 2015; 47:284–290. [PubMed: 25642633]

Lowe H, Ferris T, Hern Pez, Webe S. STRIDE - An Integrated Standards-Based Translational Research Informatics Platform. AMIA Annu Symp Proc. 2009:391–395. [PubMed: 20351886]

Lumley T, Rice K. Potential for revealing individual-level information in genome-wide association studies. J Am Med Assoc. 2010; 303:659–660.

Malin B, Emam K, O'Keefe C. Biomedical data privacy: problems perspectives and recent advances. JAMIA. 2013; 1:2–6. [PubMed: 23221359]

Marchini J, Cardon L, Phillips M, Donnelly P. The effects of human population structure on large genetic association studies. Nature Genetics. 2004; 36:512–517. [PubMed: 15052271]

Murphy S, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in I2B2. JAMIA. 2011; 18:103–108.

Nyholt D, Yu C, Visscher P. On Jim Watson's APOE status: genetic information is hard to hide. Eur J Hum Genet. 2009; 17:147–149. [PubMed: 18941475]

Plenge R, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis– a genomewide study. New England Journal of Medicine. 2007; 357:1199–1209. [PubMed: 17804836]

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 2006; 38:904–909. [PubMed: 16862161]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Bakker P, Daly M, Sham P. PLINK: a toolset for whole-genome association and population-based linkage analysis. AJHG. 2007; 81:559–575.

Sankararaman S, Obozinski G, Jordan M, Halperin E. Genomic privacy and the limits of individual detection in a pool. Nat Genet. 2009; 41:965–967. [PubMed: 19701190]

Schwartz R, Halldorsson B, Bafna V, Clark A, Istrail S. Robustness of inference of haplotype block structure. JCB. 2003; 10:13–19.

Shringarpure S, Bustamante C. Privacy risks from genomic data-sharing beacons. AJHG. 2015; 97:631–646.

Simmons S, Berger B. One size doesn't fit all: measuring individual privacy in aggregate genomic data. SPW. 2015:41–49.

Simmons S, Berger B. Realizing privacy preserving genome-wide association studies. Bioinformatics to appear. 2016; doi: 10.1186/1472-6947-15-S5-S2

Tramer F, Huang Z, Habeaux JP, Ayday E. Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. CCS '15. 2015:1286–1297.

Tucker G, Price A, Berger B. Improving power in GWAS while addressing confounding from population stratification with PC-Select. Genetics. 2014; 197:1044–1049.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Uhler C, Fienberg S, Slavkovic A. Privacy-preserving data sharing for genome-wide association studies. Journal of Privacy and Confidentiality. 2013; 5:137–166. [PubMed: 26525346]

Vinterbo S, Sarwate A, Boxwala A. Protecting count queries in study design. JAMIA. 2012; 19:750–757. [PubMed: 22511018]

Wasserman L, Zhou S. A statistical framework for differential privacy. Journal of the American Statistical Association. 2010; 105:375–389.

Weber G, Murphy S, McMurry A, MacFadden D, Nigrin D, Churchill S, Kohane I. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. JAMIA. 2009; 16:624–630. [PubMed: 19567788]

Yang J, Zaitlen N, Goddard M, Visscher P, Price A. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014; 46:100–106. [PubMed: 24473328]

Yu F, Fienbergb S, Slavkovic A, Uhler C. Scalable privacy-preserving data sharing methodology for genome-wide association studies. JBI. 2014; 50:133–141.

Yu F, Rybar M, Uhler C, Fienberg S. Differentially private logistic regression for detecting multiple-SNP association calable privacy-preserving data sharing methodology for genomin GWAS databases. Privacy in Statistical Databases. 2014; 8744:170–184.

Yu F, Ji Z. Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. BMC Medical Informatics and Decision Making. 2014; 14:S3. [PubMed: 25521367]

Zhang J, Xiao X, Yang Y, Zhang Z, Winslett M. PrivGene: differentially private model fitting using genetic algorithms. SIGMOD ‘13. 2013

Zhao Y, Wang X, Jiang X, Ohno-Machado L, Tang H. Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery. JAMIA. 2015; 22:100–108. [PubMed: 25352565]

Zhou X, Peng B, Li Y, Chen Y, Tang H, Wang X. To release or not to release: evaluating information leaks in aggregate human-genome data. ESORICS. 2011; 2011:607–627.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 whether or not any one individual participates in the study (more specifically, the probability differs by a factor of $\leq \exp(\varepsilon)$, where ε is a positive real number). This helps insure 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 $\boldsymbol{\varepsilon}$, 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. This means we would like to set $\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\varepsilon}$. Informally, taking the frequentist's perspective (Wasserman and Zhou, 2010), one can think of $\exp(\boldsymbol{\varepsilon})$ as bounding the power-to-significance ratio of any statistical test the adversary might use to determine a participant's disease status based on $\boldsymbol{\varepsilon}$-phenotypically differentially private data.

More formally, assume there is an adversary who would like to determine the i^{th} 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 $\boldsymbol{\varepsilon}$-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 $\exp(\boldsymbol{\varepsilon})$ 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 $\exp(\boldsymbol{\varepsilon})$ from a more 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), $\boldsymbol{\varepsilon}$-phenotypic differential privacy guarantees the probability of that negative event occurring differs by at most a factor of $\exp(\boldsymbol{\varepsilon})$, based on whether or not they have the disease.

For example, $\boldsymbol{\varepsilon} \leq 2$ 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 when the participant does not have the disease. In agreement with previous work (Vinterbo et al., 2012), we consider $\boldsymbol{\varepsilon} \leq 2$ to be a realistic level of privacy, though exact thresholds differ from application to application.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Note that the privacy guarantee decreases as the number of queries increases: If the user makes k queries, where the ith query is ε_i - phenotypic differentially private, the result is $(\varepsilon_1 + \cdots + \varepsilon_k)$- phenotypic differentially private.

In practice, anyone who wants privacy-preserving access to a database begins by signing up for an account. Note that, unlike in the controlled access model, no time consuming application is necessary. All that is needed is to ensure a given individual has not signed up for an account before (you could, for example, have a system based off academic email addresses or PubMed author IDs). Upon registering, each user is given a privacy budget, ε. They can use that privacy budget to query the database in a differentially private way, until their privacy budget is used up (for example, the user may make k queries, where each query is $\frac{\varepsilon}{k}$-differentially private). After the privacy budget is used up, the user must than apply for an increase in their privacy budget. More details are given in the Supplemental Experimental Procedure.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

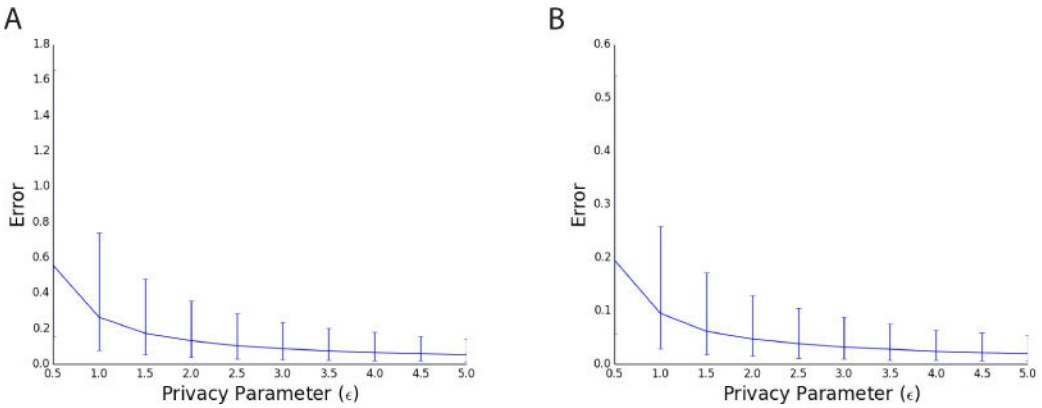


Figure 1.

The accuracy of our mechanisms for approximating the EIGENSTRAT statistic on real (a) and simulated (b) GWAS data for various privacy parameters, ε. Median error over all SNPs, with error bars representing the 25% and 75% quantiles. As expected, the accuracy increases as ε increases (a.k.a. as privacy decreases).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

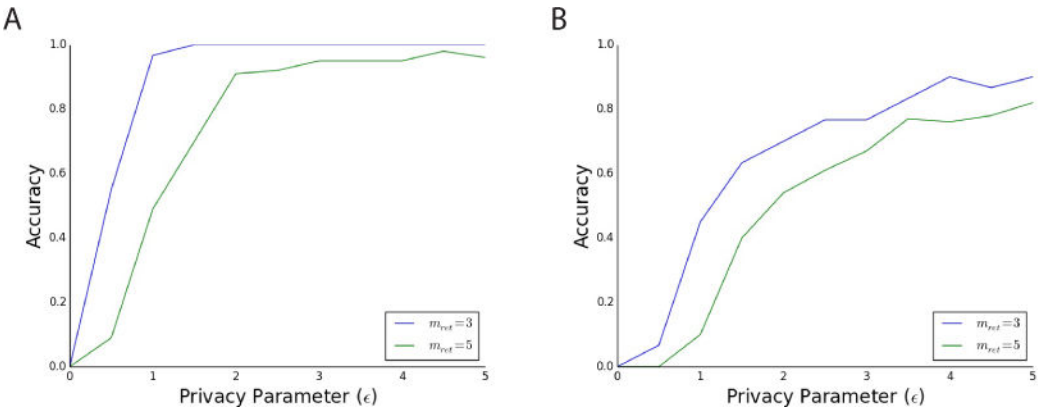


Figure 2.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

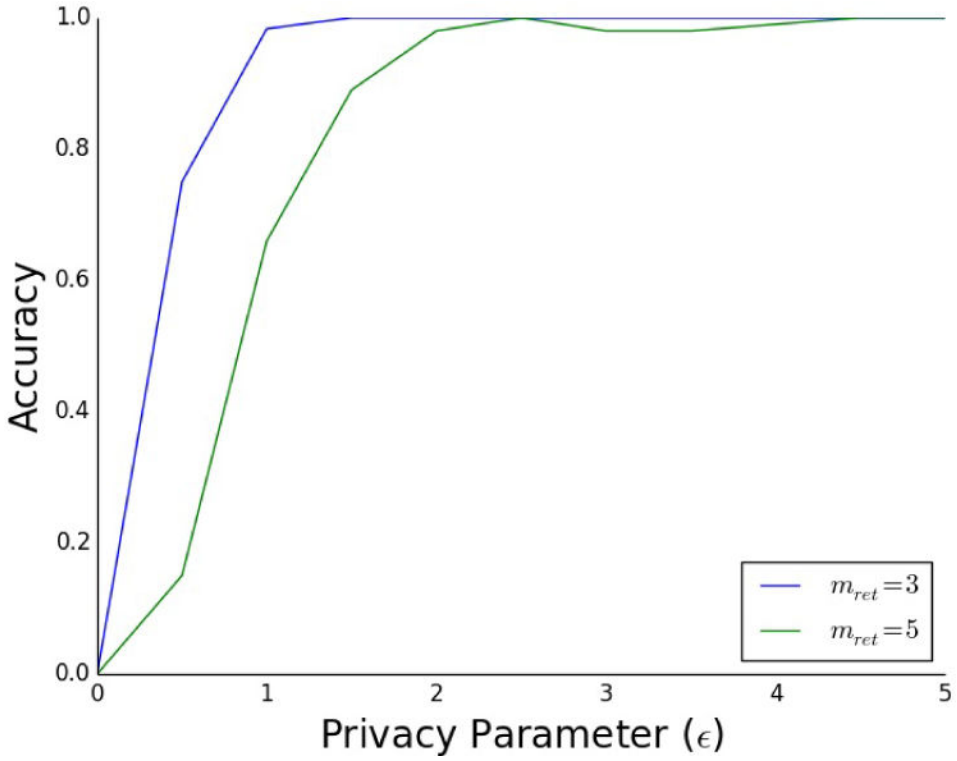


Figure 3.

The accuracy (percentage of top SNPs correctly returned) of the PrivLMM method for selection of top SNPs with m_{ret} (the number of SNPs being returned) equal to 3 and 5 for our rheumatoid arthritis GWAS dataset, with varying values of the privacy parameter, ϵ. We see that, in both cases, high accuracy is achieved for reasonable privacy levels (ϵ around 2). These results are averaged over 20 iterations.