Protecting Genomic Data Privacy with Probabilistic Modeling

Sean Simmons

Stanley Center, Broad Institute,
Cambriadge, MA 02142, USA
E-mail: ssimmons@broadinstitute.org

Bonnie Berger *

CSAIL and Department of Mathematics, MIT
Cambriadge, MA 02142, USA
E-mail: bab@csail.mit.edu

Cenk Sahinalp *

Department of Computer Science, Indiana University,
Bloomington, Indiana 47405, USA
E-mail: cenksahi@indiana.edu

The proliferation of sequencing technologies in biomedical research has raised many new privacy concerns. These include concerns over the publication of aggregate data at a genomic scale (e.g. minor allele frequencies, regression coefficients). Methods such as differential privacy can overcome these concerns by providing strong privacy guarantees, but come at the cost of greatly perturbing the results of the analysis of interest. Here we investigate an alternative approach for achieving privacy-preserving aggregate genomic data sharing without the high cost to accuracy of differentially private methods. In particular, we demonstrate how other ideas from the statistical disclosure control literature (in particular, the idea of disclosure risk) can be applied to aggregate data to help ensure privacy. This is achieved by combining minimal amounts of perturbation with Bayesian statistics and Markov Chain Monte Carlo techniques. We test our technique on a GWAS dataset to demonstrate its utility in practice. An implementation is available at https://github.com/seanken/PrivMCMC.

Keywords: Genomic Privacy; GWAS; MCMC

1. Introduction

There is a tension in modern human genomics between data sharing and privacy concerns.[1] On the one hand, genomic data holds the promise of greatly improving human health, so the ability to share it is paramount. On the other hand, our genomes are some of the most private pieces of information we have, and the risks of sharing it openly are far from understood. Even releasing aggregate genomic data (statistics calculated on an entire group of individuals, such as odds ratios or minor allele frequencies - MAF) can raise privacy concerns.[2–5]

*Corresponding Authors.

Numerous approaches have been suggested for enabling the sharing of aggregate genomic data while respecting participants privacy.[1,6] In practice, most data is either open access (posted online with minimal privacy considerations) or controlled access (only shared with trusted individuals). Recently, there has been a push to develop alternative methods for sharing this data publicly while still preserving privacy.[3,7,8] Most of these methods rely on strong assumptions about the background population (such as independent SNPs, lack of stratification, etc). As such, it is unclear how accurate a measure they provide on real world populations. Moreover, it is unclear how to extend them to more general classes of statistics (beyond MAF, etc). Alternatively, there have been methods suggested that give strong privacy guarantees with little to no assumptions about the underlying data (namely differential privacy, see section 2.5 for a definition).[9–15] Though these methods are effective at sharing small amounts of data while preserving strong privacy guarantees, current methods become inaccurate when scaled to more than a few genomic loci.[14,15]

Here we introduce a method for preserving privacy that begins to address both concerns. We build upon ideas from the statistical disclosure literature. In particular, our approach is based off of measuring the risk of reidentification using Bayesian approaches.[16] Our method aims to protect private disease status information for participants in GWAS studies, while making minimal assumptions about how the genomic data was generated (in particular, we assume that the individual trying to learn private information, known as the adversary, does not have any information allowing them to distinguish cases from controls a priori), and allowing release of more accurate statistics than that achieved by current differentially private methods. Moreover, unlike differential privacy, it is straightforward to apply our approach to almost any statistic of interest.

2. Methods

2.1. *The Model*

In the model underlying our method, we are given two pieces of information: the genotype data of each individual, and their disease status. Let $D = \{d_1, \cdots, d_n\}$, where $d_i \in \{0, 1, 2\}^m$ for $i = 1, \ldots, n$, be the genotype data of all individuals in our study. Let $y = (y_1, \cdots, y_n) \in \{0, 1\}^n$ be the vector of disease statuses ($y_i = 1$ if individual i has the disease, $y_i = 0$ otherwise). Let n_1 be the number of times a 1 occurs in y (number of cases), n_0 the number of times 0 occurs (number of controls), n the total number of individuals, and m the number of SNPs we want to share aggregate data about.

Let Y be a random variable which takes values in $\{0, 1\}^n$. This variable represents the adversaries prior belief about how likely each individual in the study is to be a case or control.

In particular, we will define it so that $Pr(Y = y')$ is equal for all $y' \in \{0, 1\}^n$ such that y' with exactly n_1 ones, and 0 otherwise; i.e.:

$$Pr(Y = y') = \frac{1}{\binom{n}{n_0}}$$

This represents the prior probability of each individual in the study being either a case or control, assuming all such assignments are equally likely. In essence, this model is meant to represent an adversary who knows everything about the study (genotypes, etc) except has no idea who is a case and who is a control.

2.2. *The Privacy Approach*

We want to release some statistics based on y and D. In order to protect participants privacy, however, we add a small amount of noise to them.

More formally, consider a statistic X that takes in both genotype data and disease status information, and outputs a vector of statistical information in $\mathbb{R}^k$ for some integer k. We want to release $X(y, D)$ while preserving privacy. In order to do this, we instead release $X + \epsilon$, where $\epsilon = (\epsilon_1, \cdots, \epsilon_k)$ is a random noise term.

For our purposes, we will assume that each ϵ_i is either a Laplacian random variable or a truncated Laplacian random variable. This choice is so as to be consistent with the Laplacian mechanism, a standard diferentially private technique.[17] In particular, for given parameter λ and bound δ, we have that:

$$Pr(\epsilon_i = z) \propto \begin{cases} exp(-\frac{|z|}{\lambda}), & -\delta < z < \delta \\ 0 & \text{otherwise} \end{cases}$$

Here λ controls the variance, and δ the maximum/ minimum amount of noise added. When δ is set to infinity, we get a standard (unbounded) Laplacian random variable, represented by $Lap(0, \lambda)$.

2.3. *The Privacy Measure*

Having specified a method for releasing privacy-preserving statistics, we want to be able to measure how much privacy is lost upon releasing them. Instead of using differential privacy based measures, however, we suggest an alternative approach based on prior probability, specified by the random variable Y. The measure of privacy we use is based on the assumption that anyone looking at the statistics does not know which participants are in the case versus the control cohort. In particular, we consider all possible permutations of participants disease status, and assume that all such assignments are equally likely from an outsiders point of view. This probabilistic model is inspired by the model used to justify k-anonymity (a standard technique in the statistical disclosure literature) and related techniques.[16,18,19]

For a given statistic X, genetic dataset D, and a disease status vector y, we want to release χ, a noisy version of X, defined as:

$$\chi(y, D) = X(y, D) + \epsilon$$

In order to measure the disclosure risk of releasing this data, we consider, for the ith individual, the probability

$$Pr(Y_i = 1 | \chi(Y, D) = \chi(y, D))$$

This can be seen as measuring the probability that the adversary believes the ith individual has the disease based on the perturbed statistic $\chi(y, D)$. Our goal is to keep this quantity as small as possible, particularly when $y_i = 1$ (when the individual has the disease). It is worth noting that, for a randomly chosen i, this has an expected value of $\frac{n_1}{n}$. Note that we do not consider the probability that $Y_i = 0$, since in general revealing that a given individual does not have a disease is not considered a privacy breach. This decision is consistent with the membership privacy idea used in previous work,[10,20] though our method can be easily modified to consider the probability $Y_i = 0$ as well.

2.4. *Estimating the Posterior*

In theory, we would like to have an exact estimate of $Pr(Y_i = 1|\chi(Y, D) = \chi(y, D))$. In practice, however, there does not seem to be an easy way to do this. Short of brute force, there does not seem to be a general method that works for more than a handful of statistics. As such, we use a form of Markov Chain Monte Carlo (MCMC) known as the Metropolis-Hastings algorithm[21] to estimate this probability.

In order to achieve this, we first draw $y' \sim Pr(Y = y'|\chi(Y, D) = \chi(y, D))$ using a two step process.

(1) Pick $y' \sim Pr(Y = y'|X(Y, D) + Lap(0, \lambda) = \chi(y, D))$ using Metropolis-Hastings, where $Lap(0, \lambda)$ is a k-dimensional unbounded Laplacian variable. The proposal distribution, q, we use to do this is chosen so that $q(y_1, y_2) \propto 1$ if $|y_1 - y_2|_1 = 2$ and equals 0 otherwise.

(2) If $\max_{\forall i} |X_i(y', D) - \chi_i(y, D)| < \delta$ return y', else go back to the previous step

Here, the proposal distribution dictates the probability of each step in the random walk used for MCMC. Our choice ensures each such jump corresponds to swapping one case and one control.

Note that, if the noise is not truncated, then step 1 suffices. We can use the above algorithm to generate a series of samples which can be used to estimate $Pr(Y_i = 1|\chi(Y, D) = \chi(y, D))$. It can be shown that this approach results in a correct asymptotic estimate of the probabilities of interest. Note that we use 100,000 steps as burn-in with 10,000 steps between samples in the Metropolis-Hastings algorithm.

2.5. *Comparison to differential privacy*

Differential privacy[17] is a common definition of privacy in the cryptographic literature. Formally:

Definition 1. *A random function F is ϵ-differentially private, if for all datasets D and D' that differ in exactly one entry, and for all sets S, we have that*

$$Pr(F(D) \in S) \leq exp(\epsilon)Pr(F(D') \in S)$$

Note that it is hard to directly compare the privacy guarantees of differential privacy to the privacy guarantees provided by risk based methods, since there is no clear correspondence (they have different assumptions, one gives a risk bound and one a risk estimate, etc). In an attempt to overcome this qualitative difference, we note that, under reasonable assumptions (namely that the distribution is a mutually independent distribution,[10] assumptions we believe reasonable to assume in our setting), differential privacy can be thought of as ensuring that, for any dataset D_1 and any $d \in D_1$

$$log(Pr(d \in D_1|F(D_1))/Pr(d \in D_1)) \leq \epsilon$$

In our setting, if we take D_1 to be the case cohort and our statistic of interest to be the MAF (see Section 2.8), this corresponds to

$$log(Pr(Y_i = 1|\chi(Y, D) = \chi(y, D))/Pr(Y_i = 1)) \leq \epsilon$$

Therefore, in order to compare differential privacy— in particular a well-known differentially private mechanism known as the Laplacian mechanism[17]— to our method, we compare this upper bound to the maximum value of our MCMC based estimate of $log(Pr(Y_i = 1|\chi(Y, D) = \chi(y, D))/Pr(Y_i = 1))$ taken over all individuals in the case cohort. The smaller this quantity, the smaller the risk relative to the background risk. Though not a perfect comparison, it gives us some idea of how our method compares to differential privacy. To vary epsilon in our analysis the number of SNPs is varied from 10 to 50 SNPs, using noise parameter $\lambda = .01$, and the log probability ratios are calculated with both our approach and the Laplacian mechanism.

2.6. *Error in MCMC*

To measure the error in our MCMC approach, we consider a dataset with 1000 individuals, 50 cases and 950 controls, each with 20 SNPs. By error, we mean the difference between our estimated probabilities and the theoretical ones. For each SNP, the controls have 0 copies of the minor allele, and the cases have 2 copies. For this dataset, it is easy to calculate the marginals of interest using simple combinatorial and probabilistic arguments. As such, we are able to compare the exact marginals on this dataset with the marginals estimated by our method.

2.7. *The Data*

In order to test our method, we use genotype data from Plenge et al.[22] This data is from a rheumatoid arthritis dataset. In most of our tests, we used 50 random cases and 950 random controls. Note that we did not use the full dataset, since our method requires that the controls out number the cases by a large margin—in particular, it is aimed at datasets without ascertainment biases (e.g. studies that have the same percentage cases as the background population). Otherwise, simply knowing someone is in the dataset reveals that they have a greater risk of having the disease being studied than someone in the general population. Though historically GWAS have been enriched for cases compared to the background population, recent population level datasets (such as the UK biobank, direct to consumer studies, etc) have started to change this, a trend likely to continue. Note that this dataset was used in all the results below, except for Fig 2 and Fig 4 where we use data from a GWAS of bladder cancer,[23] choosing 50 cases and 950 controls, and in Section 3.4 where simulated data was used. For all results we used randomly chosen SNPs with MAF greater than .05 and no missing values.

2.8. *Statistics of Interest*

We test our method on the MAF of the case cohort and the log odds ratio from the entire dataset. The MAF is defined as:

$$maf(y, D) = \frac{1}{2n_1} \sum_i y_i d_i$$

while the log odds ratio is defined, for the jth snp, as:

$$logOdds_j(y, D) = log \left(\frac{a_j(1 - b_j)}{b_j(1 - a_j)} \right)$$

where $a_j = maf_j(y, D)$ and $b_j = maf_j(\bar{1} - y, D)$.

3. Results

3.1. *The Privacy Cost of Releasing More Data*

We first apply our method to the minor allele frequency (MAF) of the case cohort. In particular, we use a set of 50 cases and 950 controls from a Rheumatoid Arthritis GWAS (see

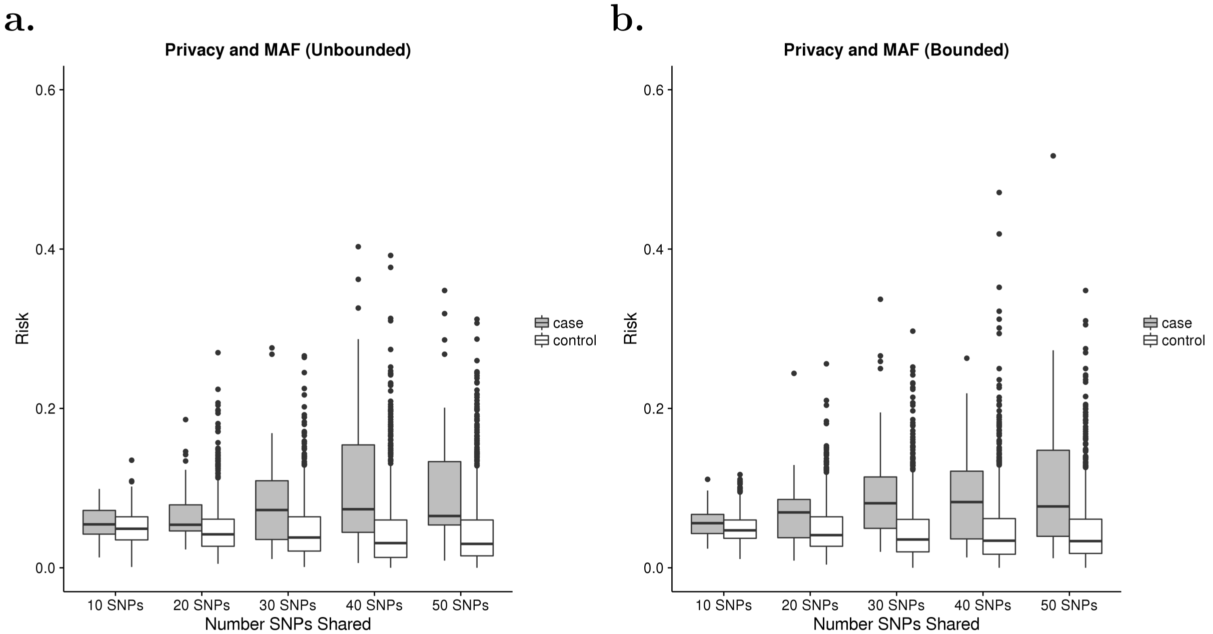

Fig. 1. Number of SNPs versus privacy. We compare the disclosure risk versus the number of SNPs whose MAF data we release from a rheumatoid arthritis GWAS, with both (a) unbounded and (b) bounded noise. This demonstrates that, unsurprisingly, privacy is greatly affected by the amount of data released. Less intuitively, we see most of this privacy loss is suffered by a few individuals, rather than being evenly shared between all individuals in the cohort. More importantly, it demonstrates the utility of disclosure risk to measure the level of privacy concerns. The risk is calculated on a dataset of 50 cases, 950 controls, with the number of SNPs released varying between 10 and 50 SNPs, with noise $\lambda = .01$. The bounded noise is bounded by $\delta = .05$.

Methods). In order to ensure privacy, we add Laplacian noise with parameter $\lambda = .01$ to the output MAF for each SNP (see Methods). This corresponds to an expected error in the returned statistic of .01.

In this setting, we used our method to measure the amount of privacy lost when releasing the MAF from various numbers of SNPS between 10 and 50 (Fig 1a). We look at randomly chosen SNPs, though one could use similar techniques to look at SNPs of particular interest (such as those with low p-values). Unsurprisingly, we see that, as the number of SNPs increases, the disclosure risk (that is to say the amount of privacy loss) increases for individuals in the case cohort. Less intuitively, we see that, though the average risk for cases increases slowly as more data is released, there are a few outliers with much higher risk. This suggests the possibility of removing these individuals in an attempt to lower the risk of releasing the data (assuming that doing so does not introduce bias into the results). We also tested our method on another GWAS of bladder cancer patients and found similar results (Fig 2).

Many practitioners are uncomfortable with the idea of adding unbounded noise to a statistic, even in the name of ensuring privacy. Unlike differential privacy, our method is flexible enough to allow us to bound the error of our output. As such, we considered adding bounded Laplacian noise to the MAF, to see how bounding the noise effects privacy (see Methods). This ensures that the released statistic is within a window of length .1 centered around the true MAF. Using this noise, we ran the same experiment that was run for Laplacian noise above (Figure 1b). Again, we see that disclosure risk increases as the number of SNPs released increases, most notably in a few outliers.

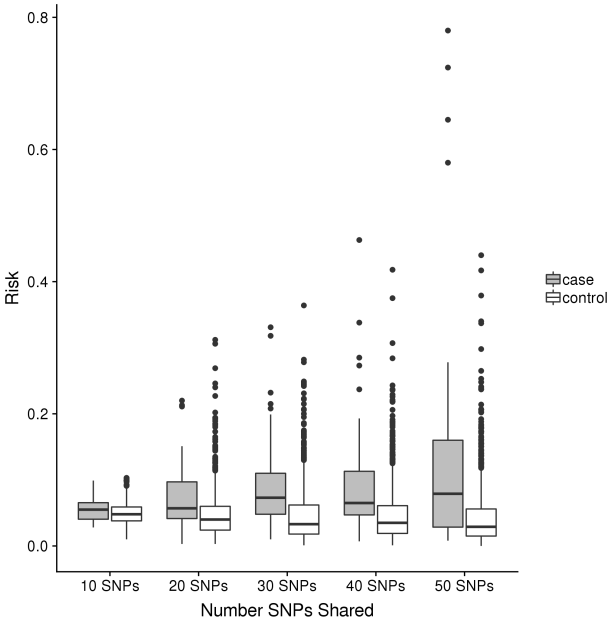

Fig. 2. Number of SNPs versus privacy. We compare the disclosure risk versus the number of SNPs whose MAF data we release, with unbounded noise on a bladder cancer GWAS dataset. The results are, unsurprisingly, similar to those in the Rheumatoid Arthritis dataset.

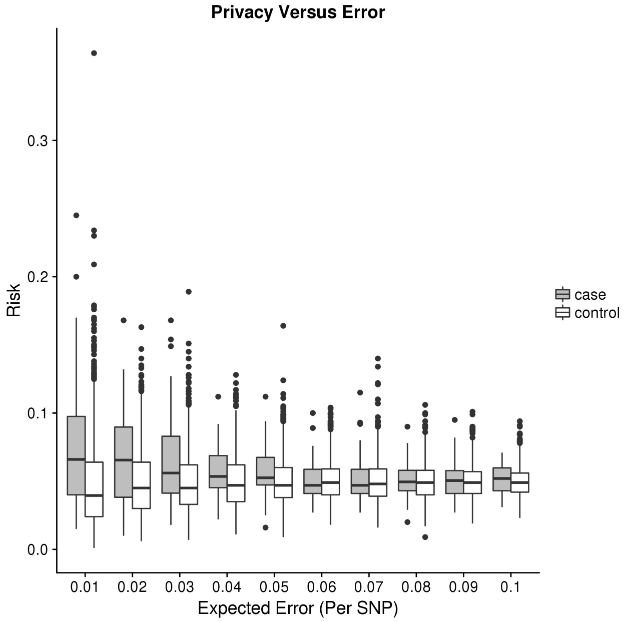

Fig. 3. Privacy versus accuracy. We compare the effect of the amount of noise versus disclosure risk when adding unbounded noise. We see that, as the noise increases, disclosure risk decreases, so privacy level increases. This increase in privacy, however, sees diminishing returns with fairly low errors, suggesting that adding large amounts of noise might not be needed. The risk is calculated on a rheumatoid arthritis GWAS dataset of 50 cases, 950 controls, with 25 SNPs released, and noise varying between $\lambda = .01$ and $\lambda = .1$.

3.2. *Accuracy Versus Privacy*

We are also interested in exploring the trade off between accuracy and privacy. The larger the amount of noise added to our statistics, the less privacy risks are encountered. At the same time, the more noise that is added, the less accuracy that is achieved. As such, we compared the the amount of noise added to the level of risk. In particular, we considered releasing the MAF for 25 SNPs with Laplacian noise added to them. We varied the expected error per SNP between .01 and .1 (Fig 3). This was achieved by varying the λ parameter of the Laplacian distribution. We see that, as the accuracy increases, the risk of disclosure increases as well. The trade off between accuracy and privacy is important, since it can help determine which choice of noise parameter is reasonable in any particular setting. In particular, we see that the privacy gains of increasing the error level off quickly, suggesting that there is not much incentive to add large amounts of noise to the data.

It is also of interest to figure out the amount of privacy lost when publishing unperturbed statistics. Unfortunately, our method relies on the addition of noise to calculate the risk (to enable MCMC). Having said that, as the noise approaches zero, the risk should approach that of releasing the unperturbed statistics. As such, the risk we see in Fig 3 for low levels of noise should give us some idea about the risk of the unperturbed dataset.

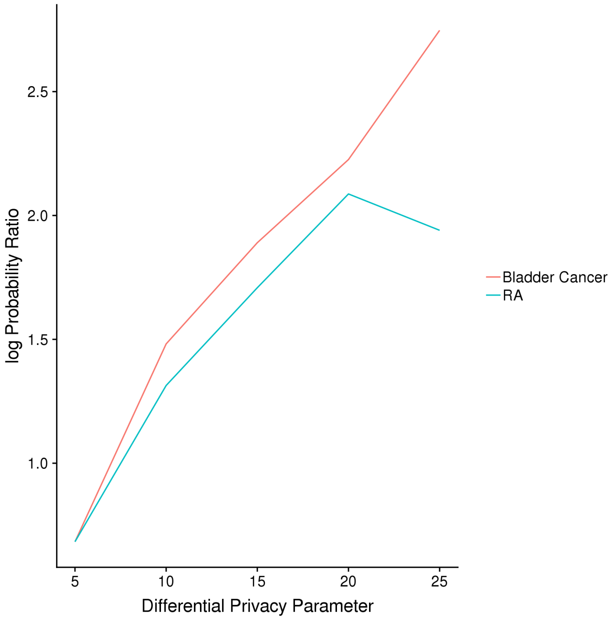

Fig. 4. Comparison to differential privacy. We compare ϵ to the log probability ratio (see methods) when adding unbounded noise for both bladder cancer (BC) and rheumatoid arthritis (RA) cohorts. We see that the log ratio for our measure is much smaller than that predicted by differentially private bounds for the Laplacian mechanism, differing by roughly a factor of 10. This shows that, under reasonable assumptions, the Laplacian mechanism greatly overestimates the level of noise required to achieve a given level of privacy. The risk is calculated on a rheumatoid arthritis dataset of 50 cases, 950 controls, with 10 to 50 SNPs released, and noise $\lambda = .01$.

3.3. *Comparison to Differential Privacy*

One of the main candidates that has been suggested for privacy preserving statistical calculations is known as differential privacy. The informal idea behind differential privacy is that, by adding noise to a released statistics, one is able to achieve a level of plausible deniability that a particular individual was in your dataset. This level of deniability is measured by a privacy parameter, ϵ. The larger the ϵ parameter, the less plausible deniability is preserved.

Under reasonable assumptions ϵ can be seen as being an upper bound on the ratio of the probability of any individual being in the dataset before and after releasing the perturbed statistic of interest (see Methods). The smaller this log ratio, the less information that is being leaked. As such, we wanted to compare this upper bound with the log probability ratio produced by our probability model (Fig 4). Note that we apply the unbounded version of our method, since standard differential privacy techniques do not allow for bounded noise. To this end, we apply a standard differentially private mechanism, known as the Laplacian mechanism. We see that the log ratio for our measure is much smaller than that predicted by the differentially private bounds from the Laplacian mechanism, differing by roughly a factor of 10— far outside the normal range for differential privacy. Importantly, these results shows that our approach allows for the release of much more data, at the cost of a slightly weaker privacy guarantee.

3.4. *Accuracy of MCMC*

Our method aims to estimate the true disclosure risk using a sampling based technique. Such sampling techniques introduce uncertainty in the estimated disclosure risk. In order to quantify this, we generated a dataset were we could directly calculate the probability of any particular individual being in the output (see Methods). We then compared the true probability versus the estimated probability using our MCMC based approach. We see that, for 98.5% of the simulated individuals the error is less than .05, with only one individual having an error of greater than .1. Moreover, the estimates can be improved by increasing the number of samples taken, as well as the number of MCMC iterations per sample.

3.5. *Beyond MAF: applications to log odds ratios*

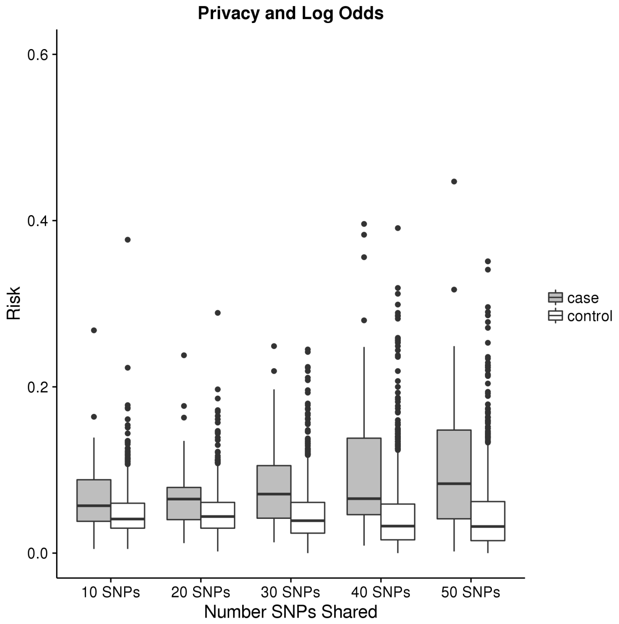

Fig. 5. Privacy and the log odds ratio. We compare the disclosure risk versus the number of SNPs whose log odds ratio data we release, with unbounded noise. The results are qualitatively very close to those we see for the MAF. More importantly, this shows that our technique can be extended to new statistics, and is not just limited to one. The risk is calculated on a rheumatoid arthritis dataset of 50 cases, 950 controls, varying the number of SNPs released from 10 and 50 SNPs, with $\lambda = .1$.

So far we have focused on using our approach to measure the privacy loss when releasing MAF for a large number of SNPs. As mentioned, however, the approach introduced here can be applied to almost any real valued statistic (or collection of statistics). To see this, we apply our method to measure the amount of privacy lost when releasing information about the odds ratio. More specifically, we release the log odds ratio for numerous SNPs.

We calculate these log odds ratios on 50 cases and 950 controls from the rheumatoid arthritis GWAS (Figure 5). We add (unbounded) Laplacian noise with parameter $\lambda = .1$ (this corresponds to an average additive error of .1 in the returned log odds ratio), and measured the privacy for various number of SNPs. We see that, in this experiment, the disclosure risk is fairly small for most individuals, with a few outliers who have greater risk. In particular, the results are comparable to what we see when releasing noisy MAF. This suggests that releasing noisy log odds ratios has a minimal effect on privacy when $\lambda = .1$.

4. Conclusion

We have introduced a novel method for measuring privacy loss in aggregate genomic data. Our method manages to avoid making the strong assumptions about the background population required by many other methods (assumptions that might not hold in practice), while still achieving better accuracy than standard differentially private methods.

The framework we introduced here can be extended in many ways, enabling more expressive analysis. For example, the current method requires adding noise to the output statistic. If this noise is small enough, the effect on accuracy is minimal, and is similar to the effect of only releasing a small number of significant digits (a common practice in most analysis). Even still, many practitioners are uncomfortable with the idea of adding noise to their data, so extending the method to unperturbed data would be of great use. For example, ideas from approximate Bayesian calculations might be used to help achieve this.[24]

Another important direction for future work is to improve runtime. This direction is of particular importance since most datasets without ascertainment bias (the type of datasets our method is meant to be applied to) are quite large. To address this, we are currently exploring approximate methods, such as variational techniques, to allow for greater scalability.[25]

Our method provides another tool to help understand the privacy risks inherent in sharing genomic data. Many of the arguments between those who want to publicly share genomic data (and health data more broadly) and those who want to keep it under lock and key revolve around the fact that we are still not certain about the real world risks posed by public disclosure of genomic data. As such, continuing to investigate the benefits and risks of sharing this data is paramount in order to be able to improve human health without negatively effecting study participants.

Acknowledgements

We want to acknowledge Jadwiga Bienkowska for introducing us to the data set we used, as well as Noah Daniels, Jian Peng, Hoon Cho, and other members of the Berger and Sahinalp labs for useful discussions.

B.B. and C.S. are partially supported by the US National Institutes of Health grant GM108348. C.S. and S.S. were partially funded by NSERC Discovery Frontiers Program, ”The Cancer Genome Collaboratory”. C.S. is partially funded by Indiana University Precision Health Initiative.

References

1. Y. Erlich and A. Narayanan, Routes for breaching and protecting genetic privacy, *Nature Reviews Genetics* **15**, 409 (2014).

2. N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. Pearson, D. Stephan, S. Nelson and D. Craig., Resolving individual‘s contributing trace amounts of DNA to highly complex mixtures using high-density snp genotyping microarrays, *PLoS Genet* **4** (2008).
3. X. Zhou, B. Peng, Y. Li, Y. Chen, H. Tang and X. Wang, To release or not to release: evaluating information leaks in aggregate human-genome data, in *ESORICS*, 2011.
4. E. Schadt, S. Woo and K. Hao, Bayesian method to predict individual snp genotypes from gene expression data, *Nat Genet* **44**, 603 (2012).
5. H. Im, E. Gamazon, D. Nicolae and N. Cox, 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 (2012).
6. X. Jiang, Y. Zhao, X. Wang, B. Malin, S. Wang, L. Ohno-Machado and H. Tang, A community assessment of privacy preserving techniques for human genomes, *BMC Medical Informatics and Decision Making* **14** (2014).
7. S. Sankararaman, G. Obozinski, M. Jordan and E. Halperin, Genomic privacy and the limits of individual detection in a pool, *Nat Genet* **41**, 965 (2009).
8. S. Simmons and B. Berger, One size doesn't fit all: Measuring individual privacy in aggregate genomic data, in *GenoPri*, 2015.
9. C. Uhler, S. Fienberg and A. Slavkovic, Privacy-preserving data sharing for genome-wide association studies, *Journal of Privacy and Confidentiality* **5**, 137 (2013).
10. F. Tramer, Z. Huang, J. Hubaux and E. Ayday, Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies, in *CCS*, 2015.
11. S. Wang, N. Mohammed and R. Chen, Differentially private genome data dissemination through top-down specialization, *BMC Medical Informatics and Decision Making* **14** (2014).
12. F. Yu and Z. Ji, 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* **14** (2014).
13. Y. Zhao *et al.*, Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery, *JAMIA* **22**, 100 (2015).
14. S. Simmons and B. Berger, Realizing privacy preserving genome-wide association studies, *Bioinformatics* **32**, 1293 (2015).
15. S. Simmons, C. Sahinalp and B. Berger, Enabling privacy-preserving gwass in heterogeneous human populations, *Cell Systems* **3**, 54 (2016).
16. J. Forster, Bayesian methods for disclosure risk assessment, in *Monographs of Official Statistics*, 2006.
17. C. Dwork and R. Pottenger, Towards practicing privacy, *J Am Med Inform Assoc* **20**, 102 (2013).
18. L. Sweeney, K-anonymity: a model for protecting privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**, 557 (2011).
19. K. E. Emam, E. Jonker, L. Arbuckle and B. Malin, A systematic review of re-identification attacks on health data, *PLoS ONE* **6** (2011).
20. N. Li, W. Qardaji, D. Su, Y. Wu and W. Yang, Membership privacy: a unifying framework for privacy definitions, in *SIGSAC*, 2013.
21. W. Hastings, Monte carlo sampling methods using markov chains and their applications, *Biometrika* **57**, 97 (1970).
22. R. Plenge *et al.*, Traf1-c5 as a risk locus for rheumatoid arthritis– a genomewide study, *New England Journal of Medicine* , 1199 (2007).
23. J. Figueroa *et al.*, Genome-wide association study identifies multiple loci associated with bladder cancer risk, *Hum Mol Genet* **23**, 1387 (2014).
24. M. Sunnaker, A. Busetto, E. Numminen, J. Corander, M. Foll and C. Dessimoz, Approximate bayesian computation, *Plos Computational Biology* **9**, p. e1002803 (2013).
25. C. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).