Bioinformatics, 36, 2020, i136–i145
doi: 10.1093/bioinformatics/btaa475
ISMB 2020

Inference attacks against differentially private query results from genomic datasets including dependent tuples

Nour Almadhoun[1], Erman Ayday[1,2,*] and Özgür Ulusoy[1,*]

[1]Computer Engineering Department, Bilkent University, Bilkent, Ankara 06800, Turkey and [2]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106, USA

*To whom correspondence should be addressed.

Abstract

Motivation: The rapid decrease in the sequencing technology costs leads to a revolution in medical research and clinical care. Today, researchers have access to large genomic datasets to study associations between variants and complex traits. However, availability of such genomic datasets also results in new privacy concerns about personal information of the participants in genomic studies. Differential privacy (DP) is one of the rigorous privacy concepts, which received widespread interest for sharing summary statistics from genomic datasets while protecting the privacy of participants against inference attacks. However, DP has a known drawback as it does not consider the correlation between dataset tuples. Therefore, privacy guarantees of DP-based mechanisms may degrade if the dataset includes dependent tuples, which is a common situation for genomic datasets due to the inherent correlations between genomes of family members.

Results: In this article, using two real-life genomic datasets, we show that exploiting the correlation between the dataset participants results in significant information leak from differentially private results of complex queries. We formulate this as an attribute inference attack and show the privacy loss in minor allele frequency (MAF) and chi-square queries. Our results show that using the results of differentially private MAF queries and utilizing the dependency between tuples, an adversary can reveal up to 50% more sensitive information about the genome of a target (compared to original privacy guarantees of standard DP-based mechanisms), while differentially privacy chi-square queries can reveal up to 40% more sensitive information. Furthermore, we show that the adversary can use the inferred genomic data obtained from the attribute inference attack to infer the membership of a target in another genomic dataset (e.g. associated with a sensitive trait). Using a log-likelihood-ratio test, our results also show that the inference power of the adversary can be significantly high in such an attack even using inferred (and hence partially incorrect) genomes.

Availability and implementation: https://github.com/nourmadhoun/Inference-Attacks-Differential-Privacy

Contact: exa208@case.edu or oulusoy@cs.bilkent.edu.tr

1 Introduction

Thanks to decreasing costs in sequencing technology, today, there is a rapid increase in the availability of genomic data and such data is used in many different areas. Recently, genomic data are utilized the most in research settings. Researchers need large genomic datasets to study the origins of individuals and identify associations between traits and specific parts of DNA. However, as shown by earlier work, public release of genomic data for research (even in anonymized form) causes serious privacy concerns. In particular, researchers have shown how such (anonymized) data can be linked back to its owner using metadata or genotype–phenotype associations (Gymrek *et al.*, 2013; Humbert *et al.*, 2015). Therefore, public availability of genomic datasets for research is currently a privacy challenge.

In contrast, as discussed, public availability of such datasets are prominent for researchers. Thus, many institutions (i.e. data owners that collect and store genomic data), rather than publicly releasing their genomic datasets, provide access to them through queries. Such queries are typically about statistical information about the dataset (referred to as a 'statistical dataset'). They are formed and sent by the researchers, computed at the data owner institution and only the results are shared with the researchers. Although this approach provides stronger privacy guarantees for the dataset participants, previous work has shown that such statistical datasets are prone to membership inference attacks, in which an adversary, using the results of the queries, can infer the membership of an individual to the corresponding dataset (Homer *et al.*, 2008). This attack is considered as serious because in most cases, dataset participants are associated with a known sensitive trait (e.g. cancer).

One way to mitigate such membership inference attacks is via the differential privacy (DP) concept (Dwork, 2008). DP-based solutions rely on adding some controlled noise to the query results in order to minimize the probability of membership inference attacks (Johnson *et al.*, 2013; Uhler *et al.*, 2013; Yu *et al.*, 2014). However, privacy guarantees of DP-based solutions are based on the assumption that all tuples in the dataset are independent. Existing works have shown how this dependency between dataset tuples reduce the privacy guarantees of DP (Liu *et al.*, 2016; Song *et al.*, 2017; Zhao *et al.*, 2017) and they have proposed general mechanisms to tackle this problem. In previous work, focusing on statistical genomic datasets, we have shown how this dependency between the tuples (i.e. dependency between the genomes of individuals from the same family) results in additional information leak from the differentially private query results (Almadhoun *et al.*, 2020). Focusing on a simple count (sum) query, we have analyzed how the privacy guarantees of DP-based solutions degrade when there are dependent tuples in a genomic dataset. Furthermore, we have proposed a mitigation technique that provides both stronger privacy and higher data utility compared to existing techniques for genomic data sharing.

In this article, we extend our work and first, we show this privacy risk on statistical genomic datasets by focusing on complex and more practical query types (that are issued in real-life), such as minor allele frequency (MAF) and chi-square value of a point mutation (single nucleotide polymorphism—SNP). In particular, we show the increased attribute inference risk (i.e. inferring SNPs of a target) when the family members of the target are also in the dataset. We formulate the attack by integrating the Mendel's law (it is out of our scope to consider specific inheritance patterns) considering several different relationship between the target and his family members in the dataset with the query of the adversary. Thus, we show how the adversary can utilize the Mendel's law and the results of real-life queries to accurately infer the genomic data of a target individual. For this article, attribute inference not only may result in genomic discrimination, but it may also be utilized in membership inference attacks. Thus, next, we also show how the outcome of the attribute inference can be utilized in a membership inference attack. For this, using a log-likelihood-ratio (LLR) test (Neyman *et al.*, 1933), we show the effectiveness of the membership inference attack on a statistical genomic dataset (e.g. that is associated with a sensitive phenotype) when the adversary uses the inferred SNPs of a target as a result of the attribute inference attack.

Our results show that an adversary can infer up to 50% more correctly leaked SNPs about the target (compared to original privacy guarantees of DP-based solutions, in which the dataset participants are assumed to be independent) about the genome of a target when it considers the dependency of the tuples in the dataset for MAF queries. This number becomes 40% for chi-square queries. We also show that the adversary achieves a high power in the membership inference attack even when it uses the inferred genome of the target (as a result of the attribute inference attack).

2 Related works

Here, we present the existing work on genomic privacy and DP, and how it relates to our proposed work in this article.

2.1 Inference attacks against statistical genomic datasets

Possibility of membership inference attacks against genomic datasets was first shown by Homer *et al.* (2008). Homer *et al.* (2008) show that using the distances between the MAF values of SNPs [released as a result of a genomic study, such as genome-wide association studies (GWAS)] and an individual's genotype, one can infer the involvement of an individual in the corresponding study. Later, Wang *et al.* (2009) exploit the correlations between SNPs to perform the membership inference and showed that such an approach needs significantly less MAF values compared to Homer *et al.* (2008). Zhou *et al.* (2011) analyzed the theoretical complexity of membership inference attacks on genomic datasets. Recently, Backes *et al.* (2016) showed the membership inference risk for datasets including miRNA expression data.

Several solutions have been proposed to protect the privacy of statistical genomic datasets considering the identified vulnerabilities (Naveed *et al.*, 2015). DP concept (Dwork, 2008) has been widely applied for privacy-preserving release of statistical summaries from various genomic studies, such as GWAS (Johnson *et al.*, 2013; Uhler *et al.*, 2013; Yu *et al.*, 2014). However, DP does not consider the dependency between dataset tuples and the aforementioned studies consider the standard DP mechanism. Therefore, privacy guarantees of DP-based techniques may degrade if a genomic dataset includes dependent tuples (e.g. individuals from the same family).

2.2 Inference attacks against DP-based mechanisms

Using differentially private query results along with auxiliary information results in inferring sensitive information, as shown by Fredrikson *et al.* (2014), in which authors show how a patient's demographic information helps to reveal the patient's genomic markers. Moreover, statistical relationships between the tuples in real-life datasets is considered as a vulnerability for standard DP-based mechanisms. The first attack performed to prove such limitations of DP-based mechanisms was in the study by Kifer *et al.* (2011). Later, for location datasets, Liu *et al.* (2016) showed how an adversary with the knowledge of pairwise dependencies between tuples can predict users' locations (Liu *et al.*, 2016). Recently, Almadhoun *et al.* (2020) analyzed the decrease in the privacy guarantees of DP-based mechanisms when there are dependent tuples in a statistical genomic dataset.

2.3 Contribution of this work

In this work, we demonstrate the scale of attribute inference attacks using differentially private results of two complex and real-life queries over statistical genomic datasets [compared to the simple sum query considered in the study by Almadhoun *et al.* (2020)]. As opposed to Liu *et al.* (2016), which only considers pairwise correlation between the tuples, we consider interdependent correlations between dataset participants. We also show how an adversary performs a successful membership inference attack using the inferred genomic data as a result of the attribute inference attacks.

3 Background

In this section, we provide brief background information about genomics and DP.

3.1 Genome-wide association studies

GWAS is the general name for case–control studies that focus on identifying genomic variations that are associated with a particular phenotype. On a broad scale, these studies help scientists uncover associations between individual SNPs and disorders that are passed from one generation to the next. A typical study compares the genomes of individuals that carry a disease or phenotype (cases) with the ones of healthy individuals (controls) to identify the functional impacts of certain SNPs on the corresponding disease. The SNP is causative or associated with the phenotype if there is a positive or negative correlation. To summarize the association information for each SNP, a 2×3 or 2×2 contingency table (as shown in Table 1) is used to show the number of cases and controls having a particular SNP with different values. The output of GWAS studies often consist of chi-square statistic, *P*-values, or MAFs for the most significant SNPs.

3.2 Differential privacy

DP provides formal guarantees that the distribution of query results changes only slightly with the addition or removal of a single tuple in the dataset (Dwork *et al.*, 2006). In other words, for any two neighboring input datasets D and D', using a probabilistic

Table 1. GWAS genotype distribution for a 2×3 contingency table (left) and a 2×2 contingency table (right)

	Genotype					Genotype		
	0	1	2	Total		0	1	Total
Case	S_0	S_1	S_2	S	Case	S_0	S_1+S_2	S
Control	C_0	C_1	C_2	C	Control	C_0	C_1+C_2	C
Total	n_0	n_1	n_2	n	Total	n_0	n_1+n_2	n

mechanism $\mathbb{A}$ will induce output distributions $\mathbb{A}(D)$ and $\mathbb{A}(D)'$ with probabilities differing by a bounded multiplicative factor e^{ϵ}.

To achieve DP, there are different general approaches. The most common three approaches are: First, by adding Laplace noise proportional to the query's global sensitivity using a Laplace perturbation mechanism (LPM) (Nissim *et al.*, 2007). Second, by adding noise related to the smooth bound of the query's local sensitivity (McSherry *et al.*, 2007). Finally, using the exponential mechanism to select a result among all possible results (Drmanac *et al.*, 2010).

3.3 DP for privacy-preserving release of GWAS results

Uhler *et al.* (2013), Yu *et al.* (2014) and Johnson *et al.* (2013) developed differentially private algorithms that release MAF and χ^2 results for genomic studies, such as GWAS. In a case–control dataset with n individuals and m SNPs, under the assumption of equal number of cases $s = n/2$ and controls $c = n/2$, Uhler *et al.* (2013) computed the sensitivity for privacy-preserving release of MAFs as $2m/n$ and χ^2 statistics as $4n/(n+2)$ (based on 2×3 contingency tables). Johnson *et al.* (2013) claimed that adding Laplace noise with scale $2/\epsilon$ to the counts of any 2×2 contingency tables results in accurate χ^2 statistics or P-values. Yu *et al.* (2014) assumed that the adversary can know the complete information of the individuals in the control group using the publicly available datasets. Hence, the sensitivity for privacy-preserving release of the χ^2 statistics is computed as $\frac{n^2}{SC}\frac{C_{max}}{(C_{max}+1)}$, where $C_{max} = \max(C_0, C_1, C_2)$ (based on 2×3 contingency tables).

4 Threat model

We consider two major threats against genomic datasets: attribute inference and membership inference. We notably aim to show how the outcome of attribute inference attack (that includes complex queries to a dataset) can be utilized in a membership inference attack. Hence, in our scenario, the goals of the adversary are (i) to infer sensitive genomic information about a target (e.g. target's rare SNPs) by sending queries to a dataset (for which the adversary knows the membership of a target and his family members); and then (ii) to infer the target's membership to another genomic dataset (for which the membership information is not publicly available). The membership of an individual in a dataset means that the corresponding individual is included in the dataset. We also show this threat model in Figure 1.

DP mechanism provides strong guarantees for protecting sensitive data of dataset members even if the adversary has prior information about the dataset. The amount of random perturbation added to the aggregate query results determines the privacy and accuracy levels. However, standard DP mechanisms do not consider the inherent correlations (or dependency) between the data tuples (e.g. correlations between genomes of family members in a genomic dataset), and hence their privacy guarantees degrade if such correlations are used by the adversary. An adversary can use auxiliary channels to learn about such dependencies in the dataset and exploit this vulnerability in DP mechanism as shown by Liu *et al.* (2016) and Almadhoun *et al.* (2020).

For the attribute inference, we follow the same attack model by Almadhoun *et al.* (2020). We assume a stronger adversary than the standard DP adversary with the following assumptions: (i) the adversary has access to the membership information of all n members in the dataset. This is possible using the publicly available metadata associated to the dataset (e.g. population information about dataset members are published along with the 1000Genomes dataset). (ii) The adversary can estimate the dependency between the dataset tuples (e.g. familial relationships) using auxiliary channels. Using prior information about the familial relationships between a target j and his family members in a genomic dataset along with the released (noisy) query results, the adversary uses the Mendelian inheritance rules to infer the genomic records of target j (X_j). These Mendelian inheritance probabilities are shown in Table 2. We study the attribute inference attack on the LPM-based differentially private query results including two queries: MAF and chi-square (χ^2).

For the membership inference, we assume that the adversary uses the outcome of the attribute inference attack (in which it infers SNP data X'_j about a target j) and tries to infer the membership of the target to another statistical case–control dataset (e.g. that may be associated with a sensitive phenotype). This is a relatively harder task compared to existing membership inference attacks against statistical genomic datasets, as here, the adversary uses the inferred (and hence partially incorrect) data about the target for its attack. The adversary has access to the MAF values of SNPs of individuals in the control group (M_C) and the MAF values of the SNPs for the entire dataset population (M_P). We assume the query type to be MAF for membership inference, and hence the adversary sends its queries to a case–control dataset asking about the MAF values for the SNPs of individuals in the case group (M_S). Using its prior information (M_C and M_P) along with the released LPM-based noisy query results about the values in M_S and the inferred genomic record X'_j of target j, the adversary's goal is to infer the membership of the target to the statistical genomic dataset. We quantify the success of this attack using a LLR test.

5 Dataset description

To evaluate the identified vulnerability, we use (and customize) real genomic datasets from two sources: (i) Manuel Corpas (MC) Family Pedigree and (ii) 1000Genome phase 3 data.

5.1 MC family pedigree

With the launch of direct-to-consumer genomic services (e.g. 23andMe), an Anglo-Spanish biologist named Manuel Corpas chose to have his and four of his family members' genomes sequenced to understand the information contained in the family personal genomics tests (Corpas, 2013). Using 23andMe services (Stoeklé *et al.*, 2016) and myKaryoView tool (Jimenez *et al.*, 2011), the DNA records of Manuel Corpas, his father, mother, sister and aunt are released in variant call format (VCF). For the considered attribute inference attack, the goal of the adversary is to infer Manuel Corpas's genomic data using the genomic records of up to his four family members (in set **F**) in the dataset. The set **F** may include (depending on the particular scenario) the target's first-degree relatives: parents and sibling, and the target's second-degree relative: aunt. We extracted the common SNPs in chromosome 22 between these 5 family members.

5.2 1000genome phase 3 data

To study the effect of unrelated individuals on the identified vulnerabilities and to study the membership inference attack, we also use the 1000Genome dataset. The 1000Genome phase 3 dataset includes partial genomic records of 2504 individuals from 26 populations. Among these, we extracted chromosome 22 genotypes for the European population using the Beagle genetic analysis package (Browning *et al.*, 2018). We identified the common SNPs between the extracted genomic data of the 1000Genome participants and the MC family members to build datasets that include members from MC family and other unrelated participants.

Eventually, we created three different datasets using 1000Genome phase 3 data: (i) a statistical dataset $D1$ (to evaluate the attribute inference attack using MAF queries), (ii) a case–control (defined in Section 3.1) dataset $D2$ (to evaluate attribute inference

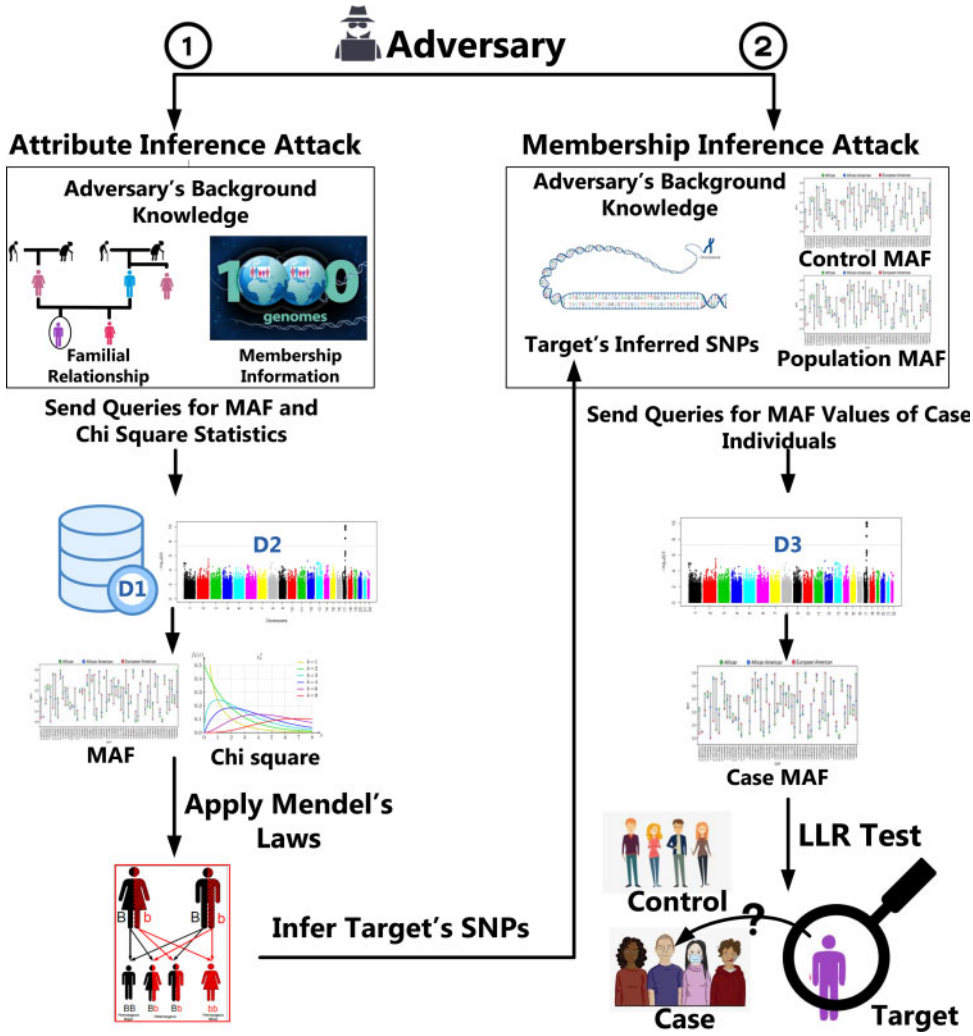

Fig. 1. Considered threat model. The adversary first runs the attribute inference attack against the target by using (i) results of differentially private MAF queries (from dataset $D1$) or chi-square statistics (from dataset $D2$), (ii) dependency between the target and target's family members that are in the dataset and (iii) Mendel's laws. Using the inferred SNPs of the target, the adversary runs a membership inference attack to infer the membership of the target in the case population in dataset $D3$ (via a LLR test)

Table 2. Mendelian inheritance probabilities for a child's SNP value given his/her parents' genotypes (left), and Mendelian inheritance probabilities for a father's SNP value given the genotypes of mother and child (right)

	Father			Son		
Mother	RR	Rr	rr	RR	Rr	Rr
RR	RR: 1 Rr: 0 rr: 0	RR: 0.5 Rr: 0.5 rr: 0	RR: 0 Rr: 1 rr: 0	RR: 0.5 Rr: 0.5 rr: 0	RR: 0 Rr: 0.5 rr: 0.5	N/A
Rr	RR: 0.5 Rr: 0.5 rr: 0	RR: 0.25 Rr: 0.5 rr: 0.25	RR: 0 Rr: 0.5 rr: 0.5	RR: 0.5 Rr: 0.5 rr: 0	RR: 0.33 Rr: 0.33 rr: 0.33	RR: 0 Rr: 0.5 rr: 0.5
rr	RR: 0 Rr: 1 rr: 0	RR: 0 Rr: 0.5 rr: 0.5	RR: 0 Rr: 0 rr: 1	N/A	RR: 0.5 Rr: 0.5 rr: 0	RR: 0 Rr: 0.5 rr: 0.5

Note: 'R' represents a major allele and 'r' represents a minor allele.

attack using χ^2 queries) and (iii) a case–control dataset $D3$ (to evaluate membership inference attack).

6 Attribute inference attack for complex queries

In the study by Almadhoun *et al.* (2020), we showed the attribute inference attack for datasets with dependent tuples by considering a simple count (sum) query, in which the adversary asks about total number of a specific SNP i's value among a subset of dataset participants (determined based on demographic data, such as location or age). Here, we analyze the attribute inference risk for more complex and realistic query types that are used in actual studies (e.g. GWAS). In particular, we consider queries that compute MAF and chi-square (χ^2) statistics of SNPs.

For a statistical dataset $D1$, we represent a SNP i's (noisy or differentially private) MAF value that is computed over target j and other dataset members in set $\mathbf{O}$ ($|\mathbf{O}| = o$) sharing some demographic information with j as $\frac{\tilde{Q}^i_o}{t} = \frac{Q^i_o}{t_1} + \frac{x^i_j}{2} + \delta$. Here, δ is the added Laplace noise suggested by Uhler *et al.* (2013), and x^i_j is the value of SNP i for target j. Q^i_o/t_1 is the MAF value of SNP i due to individuals in $\mathbf{O}$, where Q^i_o represents the number of minor alleles in $\mathbf{O}$ and t_1 is the total number of alleles in $\mathbf{O}$ for SNP i (as each SNP carries 2 alleles, $t_1 = 2o$). Also, t is the total number of alleles for all participants included in the query result ($t = 2o + 2$). Individuals in set $\mathbf{O}$ (except for the target) can be (i) the family members of target j, which we represent as a set $\mathbf{F}$ ($|\mathbf{F}| = f$) or (ii) other unrelated individuals in set $\mathbf{U}$ ($|\mathbf{U}| = u$). Hence, the probabilistic dependence for the MAF statistics can be shown as:

$$M_o^i = M_j^i + (o+1)y, \quad (1)$$

where M_o^i and M_j^i are the MAF values due to individuals in **O** and target j, respectively. $(o+1)$ is the number of all participants included in the query result (**O** and target j). Also, y is a kinship coefficient that satisfies the Mendel's law. y is in $\left[\frac{-2}{t}, \frac{2}{t}\right]$.

For a case–control dataset $D2$, (noisy) χ^2 statistics for a SNP i when the query results include target j and other dataset members in set **O** sharing some demographic data with j is represented as $\chi_i^2 = \chi_i^2 + \delta$. Here, δ represents the added Laplace noise suggested by Uhler *et al.* (2013), Yu *et al.* (2014) and Johnson *et al.* (2013). Then, the probabilistic dependence for the χ^2 statistics can be considered as:

$$\chi_i^2 = Q_{oj}^i + 2(o+1)y, \quad (2)$$

where Q_{oj}^i is the sum of the SNP values for $(o+1)$ participants included in the query (dataset members in set **O** and target j). Similar to the MAF case, y is a kinship coefficient that satisfies the Mendel's law. Thus, y is in $[-2, 2]$.

6.1 Inference evaluation algorithm

The adversary generates its queries that include the members of the same family (e.g. by forming a query based on age—location—street level—city level—state level, etc.) and receives the differentially private MAF ($\tilde{M}_{oj}^i$) or the differentially private chi-square ($\tilde{\chi}_i^2$) values. As we discussed in Section 4, the adversary has full knowledge about the membership of the dataset participants using auxiliary channels. The adversary is also aware about (i) the dependency between target j and other family members (in set **F**) that are also in the dataset and (ii) Mendel's law to formulate this dependency. Hence, using the (noisy) query result about a SNP i along with the knowledge of familial relationships, the adversary can infer the value of x_j^i, which represents the value of SNP i for target j. More specifically, from the MAF query results, the adversary can estimate the total number of minor alleles Q_{oj}^i for $(o+1)$ individuals in the query results. Then, it uses the coin change algorithm (D'Errico, 2018) to obtain all the possible partitions of Q_{oj}^i (each partition will include $\leq (o+1)$ individuals). Next, for each valid partition (validity of the partition is determined using Mendel's law when family members of the target are in the query results), the adversary computes its probability using Mendel's law by considering potential values of SNP i (0, 1 and 2) for target j. For the χ_i^2 query results, the adversary utilizes the valid partitions for different Q_{oj}^i values to guess the actual value of χ_i^2 (before the added noise by the DP-based mechanism) for different number of cases and controls. Then, the adversary compares $\tilde{\chi}_i^2$ with χ_i^2 for the same number of individuals included in the query result.

We use two metrics as in the study by Almadhoun *et al.* (2020) to quantify the success of the identified attacks: correctness and leaked information. We define the correctness of the adversary in terms of its estimation error, which is the distance between the true value of the target's actual SNP x_j^i and the value inferred by the adversary x_{ij}^i. We compute the estimation error for all m-targeted SNPs of the target as follows:

$$E = \sum_{i=1}^{m} P(x_j^i | X_j) |\text{Dist}(x_j^i, x_{ij}^i)| \quad (3)$$

Thus, correctness of the adversary can be expressed as one minus its estimation error. To quantify the difference between the adversary's prior and posterior information after the attack for m SNPs of the target, we express the leaked information as follows:

$$L = \sum_{i=1}^{m} 1 - |\text{sgn}(\text{Dist}(x_j^i, x_{ij}^i))| \quad (4)$$

6.2 Evaluation

As discussed in Section 5, we use two different datasets ($D1$ and $D2$) for each considered query type. To evaluate the attribute inference attack for MAF queries, we perform the inference attack over a statistical dataset ($D1$) with n individuals ($n = 164$) from European population. To evaluate the attribute inference attack for χ^2 queries, we create a case–control dataset $D2$. Dataset D_2 includes ($s = n/2 = 82$) cases and ($c = n/2 = 82$) controls. **A** is the set of SNP IDs for target j. For both queries, the adversary aims to infer m SNPs $\in$ set **A** on chromosome 22 for target j using $m = 100$ queries over datasets $D1$ or D_2.

To study different cases for individuals that are included in the differentially private query results, we consider the following two scenarios: (i) query results are computed over target j and multiple first and second-degree family members in **F**; and (ii) results are computed over target j, multiple family members in **F** and multiple other unrelated members (nonrelatives) in **U**. Furthermore, to show the vulnerability due to considering dependence between tuples in the query results, we evaluate the attribute inference attack considering two types of adversaries. The first adversary exploits the familial relationships (i.e. dependency) between the dataset members to reconstruct target j's genomic record. The second adversary considers that there is no dependency between the dataset members. Using the correctness and leaked information metrics described in Section 6.1, we evaluate the success of these attacks. Any extra leaked information the first adversary can infer by considering the familial relationships (dependence between the tuples) is considered as leakage that violates the standard DP guarantees.

6.3 Experimental results

Here, we show the results of the attribute inference attack for the MAF and chi-square queries.

6.3.1 MAF queries

In Figure 2, we show the estimation error of the adversary in inferring target j's m SNPs ($m = 100$). Figure 2(a) shows the effect of different sets of family members to the estimation error of the adversary. We start including one first-degree relative (which can be the father or the mother of the target) to the query results. Second, we include both the father and the mother in the query results. Third, we include the sister of target j together with his father and mother. Finally, we consider a second-degree relative (aunt of target j) in the query results along with the rest of the family members (we follow the same strategy for the other MAF experiments as well). Furthermore, Figure 2(b) shows the effect of the number of nonrelatives included in the result of a query on adversary's success in terms of its correctness in inferring SNPs of target j.

Using the results of MAF queries over the statistical dataset $D1$, we make the following key observations: (i) In Figure 2, the most accurate inference of the adversary is when the query computation includes two first-degree family members along with target j. This is unlike the results in the study by Almadhoun *et al.* (2020), in which the adversary obtains more accurate results as the number of family members included in the query computation increases. This is because here, we consider both the first-degree and second-degree relatives in a different familial dataset; and the query types we consider are more complex than the one in the study by Almadhoun *et al.* (2020). Including a second- or third-degree family member can enlarge the range of possible SNP values for the target, and hence make it more difficult to accurately infer the correct SNP value with a high probability. (ii) The estimation error of the adversary to infer the actual values of the targeted SNPs decreases (i.e. its correctness increases) considerably (by 70% with considering data dependency and by 40% without considering data dependency) as the budget privacy, ϵ, increases from 0.1 to 5, as shown in Figure 2(a). In a similar trend, in Figure 2(b), the probability of inferring the correct values for the targeted SNPs slightly increases as the value of the privacy budget, ϵ, increases from 0.1 to 5. (iii) The estimation error of the adversary with the knowledge of the data dependency is about 30% less compared to the case in which the adversary do not consider the data dependency in the query results [Fig. 2(a)]. (iv) In accordance with the results in the study by Almadhoun *et al.* (2020), including more nonrelatives in the query results (e.g. increasing the

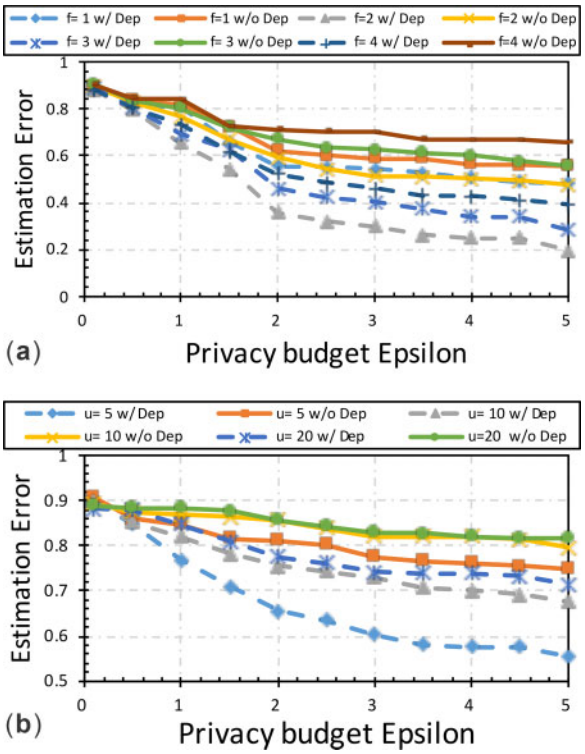

Fig. 2. The effect of different values of the privacy budget, ϵ and the number of (a) family members in set **F** ($|\mathbf{F}| = f$) and (b) two first-degree relatives (father and mother) in set **F** along with different numbers of nonrelatives in set **U** ($|\mathbf{U}| = u$) on adversary's correctness (1—estimation error) in inferring the targeted SNPs from the noisy results of MAF statistics. (w/ Dep) represents the scenario in which the adversary considers the data dependency and (w/o Dep) represents the opposite

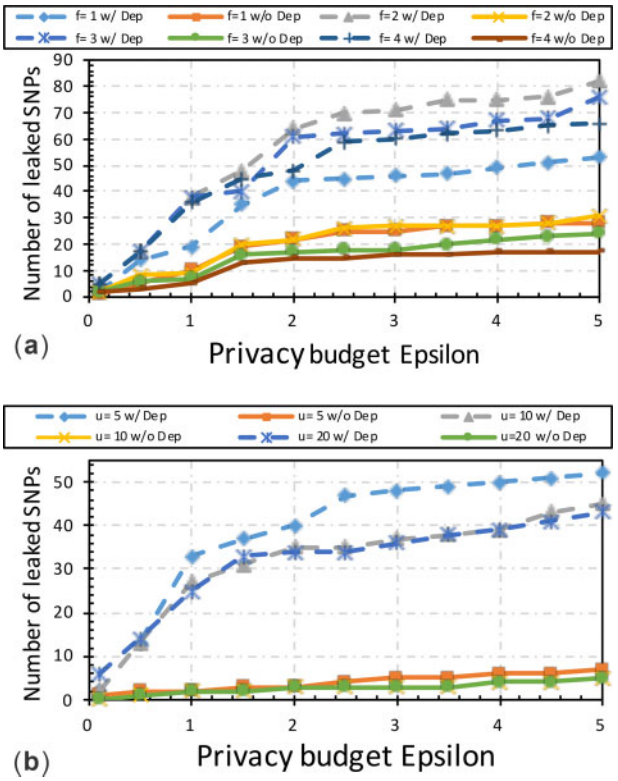

Fig. 3. The effect of including (a) only family members in **F** ($|\mathbf{F}| = f$) and (b) two first-degree relatives (father and mother) with different numbers of nonrelatives in **U** ($|\mathbf{U}| = u$) on the leaked information (i.e. number of leaked SNPs of target j) using the noisy results of MAF statistics. (w/ Dep) represents the scenario in which the adversary considers the data dependency and (w/o Dep) represents the opposite

size of **U** from 5 to 20) results in a significant increase in the estimation error of the adversary (even if the adversary has the knowledge of the data dependency), as shown Figure 2(b). Moreover, increasing the number of nonrelatives in the query results beyond 20 leads to a natural countermeasure (with a probability of 0.9) against the leakage of SNP information of the target.

Next, in Figure 3, we evaluate the effect of different values of the privacy budget, ϵ, on the number of the target's leaked SNPs (defined in Section 6.1) with different numbers of relatives (in set **F**) and nonrelatives (in set **U**) included in the query results. The results we obtain are consistent with the results of the correctness (in Fig. 2). We make four key observations: (i) The adversary, using the knowledge of the data dependency, can infer up to 50% more SNPs of the target compared to the case in which it do not consider the data dependency in the query results. (ii) Increasing the privacy budget ϵ from 0.1 to 5 results in 2–8 times (depending on the number of relatives in **F** and whether or not the adversary considers data dependency) more SNPs to be inferred by the adversary, as shown in Figure 3(a). The adversary infers the maximum number of SNPs when we include two first-degree family members in **F** and the adversary considers the dependency between tuples. We observe that the inference power of the adversary decreases when we include a second-degree relative (along with target's first-degree family members) in **F**. We also observe that if the adversary does not consider the data dependency, varying the number of family members in **F** from one to four has a slight effect on the number of leaked SNPs, as expected [Fig. 3(a)]. (iii) Increasing the privacy budget (ϵ) from 0.1 to 5 increases the number of leaked SNPs by only up to three times if the adversary does not consider the data dependency and regardless of the number of the nonrelatives, as we show in Figure 3(b). If the adversary considers the data dependency, then increasing the privacy budget from 0.1 to 5 increases the number of leaked SNPs significantly [up to 6 times in Fig. 3(b)]. (iv) When the query results include only family members in **F**, the adversary can infer about two times more SNPs when it considers the dependency compared to not considering dependency [Fig. 3(a)]. However, this increase in the adversary's inference power is four times when the query results include both family members and also nonrelated individuals [Fig. 3(b)].

Finally, in Figure 4, based on the leaked information metric, we evaluate the effect of different values of the privacy budget, ϵ, on the number of the target's leaked rare SNPs. A SNP is considered rare when it carries an allele that has low frequencies in the population. Hence, rare SNPs provide sensitive information about predispositions of individuals for complex diseases (Goldstein *et al.*, 2013). Here, out of 100 targeted SNPs, we identified 11 rare variants, for which MAF < 0.05. Out of these 11 rare variants, the results show that the adversary that considers the dependency between the tuples can infer a significant portion of target's sensitive information.

6.3.2 Chi-square queries

To compute the χ^2 statistics, we use dataset $D2$, in which data are represented as 2×3 and 2×2 contingency tables for each SNP. We use the techniques proposed by Uhler *et al.* (2013), Yu *et al.* (2014) and Johnson *et al.* (2013) for differentially private release of χ^2 statistics over D_2.

Similar to before, our goal is to show the how much the adversary's inference power increases when the dependencies in the dataset are utilized in the attack. As Yu *et al.* (2014) assume that the adversary knows the complete information of the individuals in the control group, when using the technique by Yu *et al.* (2014), we consider all the family members to be in the case group. Therefore, we use the technique (Yu *et al.*, 2014) only for the second scenario (described in Section 6.2), in which the query results are computed over target j, multiple family members and multiple other unrelated members (nonrelatives).

In Figure 5, we evaluate the effect of different values of the privacy budget, ϵ, on the adversary's correctness in inferring the targeted

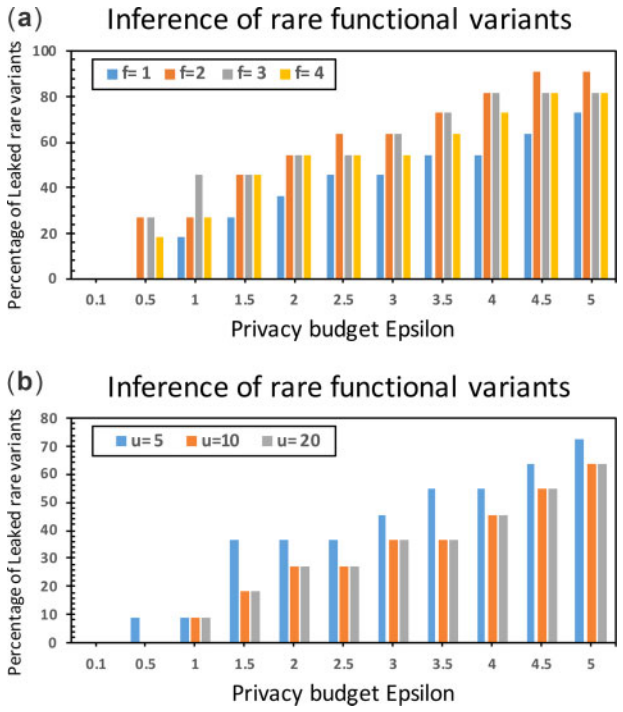

Fig. 4. The effect of including (a) only family members in **F** ($|\mathbf{F}| = f$) and (b) two first-degree relatives (father and mother) with different numbers of nonrelatives in **U** ($|\mathbf{U}| = u$) on the leaked information by only considering the rare SNPs of the target for MAF queries

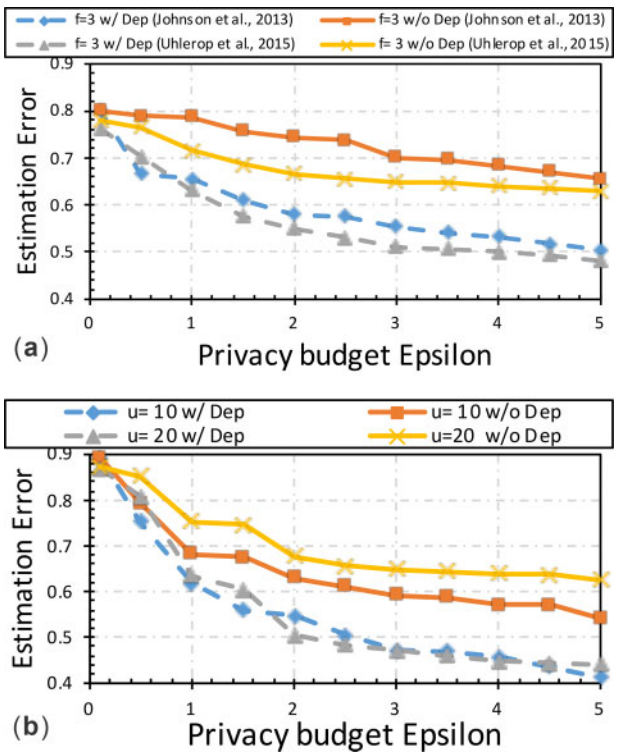

Fig. 5. The effect of different values of the privacy budget, ϵ and the number of (a) family members in **F** ($|\mathbf{F}| = f$) and (b) 2 first-degree relatives (father and mother) with different numbers of nonrelatives in **U** ($|\mathbf{U}| = u$) on adversary's correctness (1—estimation error) in inferring the targeted SNPs, using the noisy results of χ^2 statistics. (w/ Dep) represents the scenario in which the adversary considers the data dependency and (w/o Dep) represents the opposite

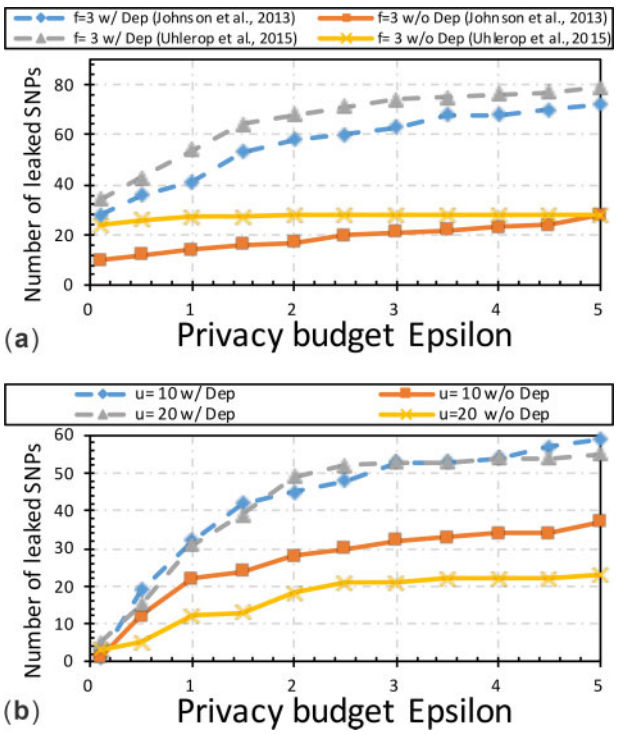

Fig. 6. The effect of including (a) only family members in **F** ($|\mathbf{F}| = f$) and (b) two first-degree relatives (father and mother) with different numbers of nonrelatives in **U** ($|\mathbf{U}| = u$) on the leaked information (i.e. number of leaked SNPs of target j) using the noisy results of χ^2 statistics. (w/ Dep) represents the scenario in which the adversary considers the data dependency and (w/o Dep) represents the opposite

SNPs using (noisy) χ^2 statistics about the SNPs. Figure 5(a) shows the estimation error when χ^2 query results only include target j and his father, mother and sister in **F** ($|\mathbf{F}| = 3$). Also, Figure 5(b) shows the estimation error when χ^2 query results include target j, both of his parents ($|\mathbf{F}| = 2$) and other unrelated individuals in **U**($|\mathbf{U}| = u$).

From these results, we make three key observations: (i) The estimation error of the adversary with the knowledge of the data dependency is up to 20% less compared to the case in which the adversary do not consider the data dependency in the query results. (ii) The estimation error for the targeted SNPs is slightly less when we apply the algorithm described by Uhler *et al.* (2013) (while releasing the noisy statistics) compared to the algorithm in the study by Johnson *et al.* (2013). (iii) In Figure 5(b), using the algorithm in the study by Yu *et al.* (2014) (for the release of noisy statistics) over (D_2), in which the data are represented as 2×3 case–control tables for each SNP, we observe that adding more nonrelatives into the query computation does not significantly affect the correctness of the adversary compared to the results in Figure 5(a) [obtained using algorithms by Johnson *et al.* (2013) and Uhler *et al.* (2013)]. This is due to the strong adversary assumption by Yu *et al.* (2014) threat model. These results show that complex queries, such as χ^2, are also vulnerable to the dependency between the tuples in the dataset. Furthermore, we observe that the results we obtain for the χ^2 query are consistent with the results of the MAF query in Section 6.3.1 and the count query in the study by Almadhoun *et al.* (2020).

Next, we evaluate the number of the target's leaked SNPs (leaked information) for different values of the privacy budget, ϵ, in Figure 6. We consider different numbers of relatives (in set **F**) and nonrelatives (in set **U**) to be included in the query results. In-line with the results of the correctness (in Fig. 5), we make the following key observations: (i) The adversary, using the knowledge of the data dependency, can infer up to 40% more SNPs of the target compared to the case in which it do not consider the data dependency in the query results. (ii) When the adversary considers the dependency in the data, increasing the privacy budget ϵ from 0.1 to 5 results in up to four times more SNPs to be inferred by the adversary, as shown in Figure 6(a). (iii) Applying different algorithms (to release the noisy statistics) has a slight effect (about 10%) on the number of

leaked SNPs. (iv) When nonrelatives from **U** are also included in the query results, the number of leaked SNPs increases by 2.5 times if the adversary considers the dependency in the data [Fig. 6(b)].

Finally, we evaluate the adversary's power to infer the target's rare variants using the leaked information metric. For different values of the privacy budget, ϵ, we compare the results of applying the algorithms proposed by Uhler *et al.* (2013), Johnson *et al.* (2013) and Yu *et al.* (2014) (to release the noisy χ^2 statistics). Similar to the results of MAF query in Section 6.3.1, out of $m = 11$ rare variants, the results show that the adversary can exploit the dependency between the tuples to infer a significant portion of the target's rare SNPs. Consistent with the previous results, we observe that the technique proposed by Johnson *et al.* (2013) provides slightly better privacy compared to the one proposed by Uhler *et al.* (2013) when the query results include three family members. Moreover, the adversary can still infer the target's rare SNPs, using the technique proposed by Yu *et al.* (2014) when the query results include nonrelatives. We do not include the plots for this experiment due to space restrictions.

6.3.3 Comparison with the sum query by Almadhoun *et al.* (2020)

In the following, to show how the vulnerability for attribute inference changes based on the type of the query, we compare the correctness of the adversary for MAF (in Section 6.3.1), χ^2 (in Section 6.3.2) and sum queries [used in the study by Almadhoun *et al.* (2020)]. Figure 7 shows the effect of different values of the privacy budget, ϵ on the adversary's estimation error (1 - correctness) in inferring the SNPs for target j, when the adversary consider the familial relationship between tuples in the dataset. First, we consider that query results include three family members in set **F** (father, mother and sister). Note that, for this case, we used two different algorithms to compute differentially private χ^2 statistics. Next, we consider that the query results include 2 family members in set **F** (father and mother) along with 10 nonrelatives in set **U** ($|\mathbf{U}| = 10$). We also observe that using the algorithm by Yu *et al.* (2014) decreases the estimation error for the targeted SNPs even if we include nonrelatives in the χ^2 query results. This is due to the strong adversary assumption by Yu *et al.* (2014) (that the adversary has full information about the controls). Thus, the algorithm by Yu *et al.* (2014) provides lower privacy compared to the one provided by Uhler *et al.* (2013) and Johnson *et al.* (2013) for releasing χ^2 statistics. However, the algorithms by Uhler *et al.* (2013) and Johnson *et al.* (2013) provide less utility for the shared query results.

Overall, our results show that the existence of dependent tuples in a genomic dataset is the main reason for the identified privacy vulnerability. Hence, an action is required to filter out or distort the dependencies between tuples in statistical genomic datasets.

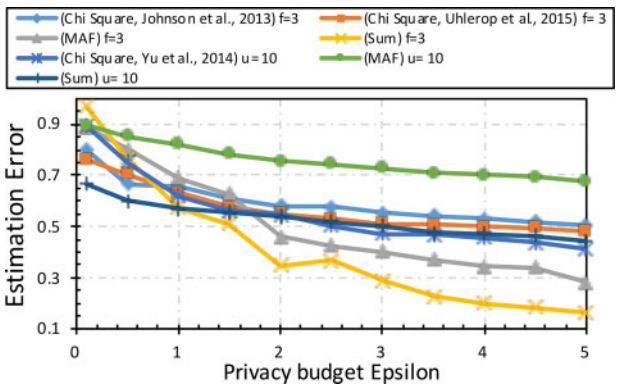

Fig. 7. The effect of different values of the privacy budget, ϵ, when the query results include (i) three family members (father, mother and sister) in set **F** and (ii) two family members (father and mother) in set **F** along with 10 nonrelatives in set **U** ($|\mathbf{U}| = 10$) on adversary's correctness (1-estimation error) in inferring the targeted SNPs. The adversary exploits the noisy results of three different queries: sum, MAF and χ^2 statistics. χ^2 statistics include three different cases as we apply three existing algorithms to generate the query results

7 Membership inference attack

In Section 6, we show the attribute inference attack for two genomic datasets with dependent tuples by considering MAF and χ^2 queries. The adversary can infer target j's genomic data Xj exploiting the probabilistic dependence between the target and his family members in **F** that are included in the query results. Here, we show how the adversary can use the outcome of the attribute inference attack to infer the membership of the target in another dataset. For the membership inference, we assume that the goal of the adversary is to determine whether target j is in the case group (including individuals with a sensitive trait) of a statistical case–control dataset ($D3$).

We assume D_3 includes 2×2 case–control tables for m SNPs and the adversary queries for the MAF values of the SNPs for the individuals in the case group in D_3. Similar to existing membership inference attacks on statistical genomic datasets, we assume that the adversary has access to (i) the set of MAF values of SNPs for individuals in the control group (M_C) and (ii) the set of MAF values of SNPs for a similar population as the entire dataset $D3$, including both the case and control individuals (M_P). The adversary receives the LPM-based noisy query result about the MAF value of a SNP i for individuals in the case group (**S**, where $|\mathbf{S}| = s$) as $\tilde{M}^i_S = M^i_S + \delta$, where δ represents the added Laplace noise suggested by Uhler *et al.* (2013).

7.1 Membership inference attack evaluation algorithm

To determine the membership of target j in the case group, we evaluate the success of the attack using a LLR. According to Neyman *et al.* (1933), at a given false-positive level (α) in binary hypothesis test, the maximum power β can be achieved using the exact LLR test. Sankararaman *et al.* (2009) introduced an LLR test that predicts whether an individual is in the case group using the query responses received from a statistical genomic dataset, and here, we follow the same model proposed in the study by Sankararaman *et al.* (2009). We assume that under the null hypothesis, target j is not a part of the case group and under the alternative hypothesis, target j is part of the case group **S**. Accordingly, the overall LLR test can be computed as:

$$\mathrm{LLR} = \sum_{i=1}^{m} x_j^i \log \frac{M_S}{M_C} + (1 - x_j^i) \log \frac{1 - M_S}{1 - M_C} \tag{5}$$

For a moderately large **S** ($|\mathbf{S}| = s$) and a moderately large number of SNPs with MAF > 0.05 (e.g. $m > 100$), Sankararaman *et al.* (2009) modeled the LLR as a Gaussian distribution to estimate the maximum achievable power β given the false-positive rate α as:

$$Z_\alpha + Z_{1-\beta} \simeq \sqrt{\frac{m}{s}}, \tag{6}$$

where Z_α is the $100(1-\alpha)$ percentile of the standard normal distribution. Moreover, Sankararaman *et al.* (2009) claimed that one can estimate the controls' MAF values M_C from the population MAFs M_P as follows:

$$M_P = \frac{s}{s+c} M_S + \frac{c}{s+c} M_C. \tag{7}$$

7.2 Evaluation

We use dataset ($D3$), in which the data is represented as 2×2 case–control tables for each SNP, as discussed in Section 5. To evaluate the membership inference attack, we choose n individuals ($n = 80$) from European population including target j. we construct the case group (individuals carrying a sensitive phenotype) of size $s = n/2 = 40$ and the control group of size $c = n/2 = 40$. Out of the m SNPs in set **A**, we choose $m = 250$ independent SNPs after discarding the rare variants with MAF < 0.05. We discard the rare variants

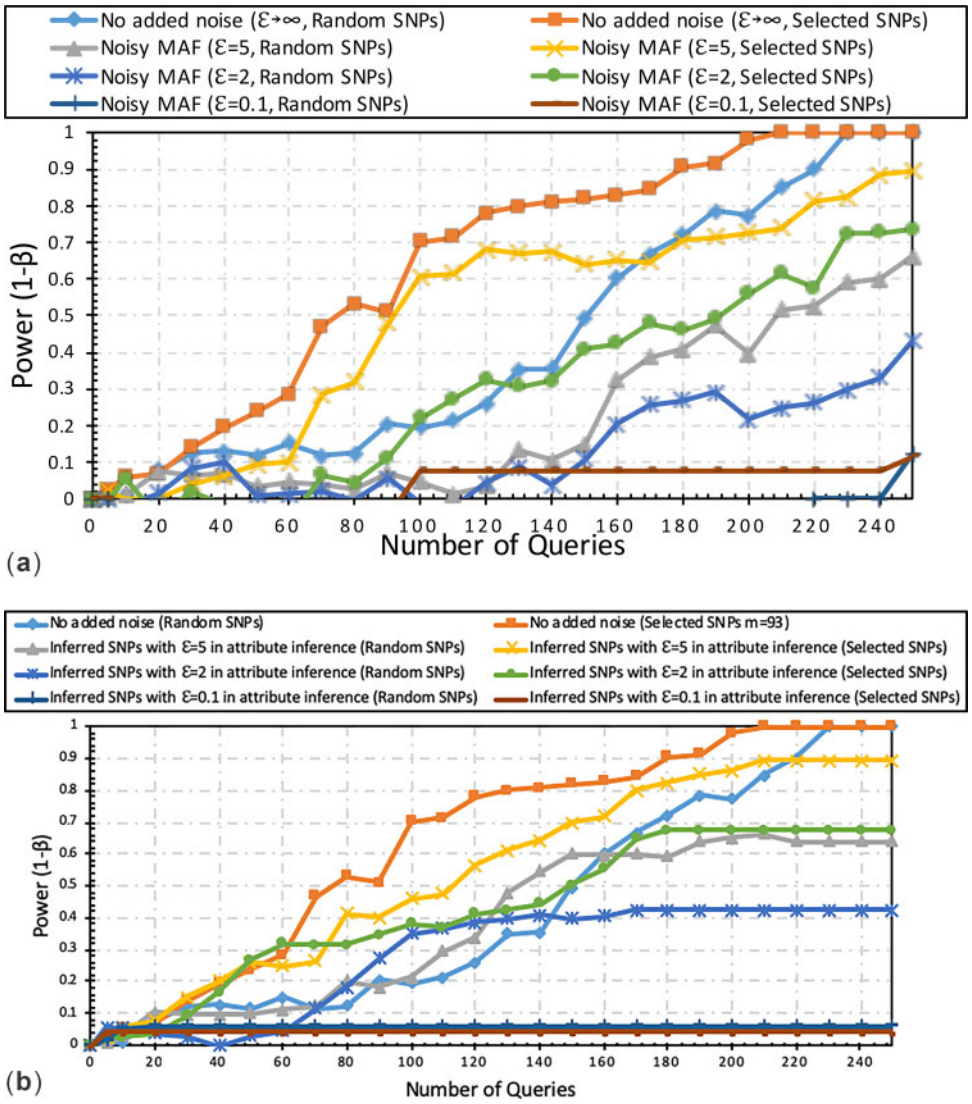

Fig. 8. Power of the adversary for the membership inference attack for different number of MAF queries over dataset $D3$. (a) The power when the adversary uses the actual (correct) SNPs of target j. (b) The power when the adversary uses the inferred SNPs of the target as a result of the attribute inference attack. In (b), ϵ values are the ones used in the attribute inference attack; for membership inference, the values in M_S are shared (in a differentially private way) with the adversary using $\epsilon = 5$ for all cases

because in the association studies, usually the exposed SNPs have a prespecified minimum MAF > 0.05. We empirically build the null hypothesis using the MAF values of the SNPs in the control group (M_C). We reject the null hypothesis and conclude that target j is more likely to be a part of the case group if LLR value is greater than a threshold value υ. We find the threshold υ from the null hypothesis with $\alpha = 0.1$. Here, α is the false-positive rate, describing when target j is not part of the case group and $LLR > \upsilon$. The power $1 - \beta$ is then the true-positive rate, describing when target j is part of the case group and $LLR > \upsilon$.

Using the (noisy) differentially private MAF values obtained for the case group (M_S), the adversary can compute the LLR value for target j. To estimate the number of queries needed to achieve a high power (which represents the adversary's confidence that target j is a member of the case group), we consider two types of adversaries: (i) The first adversary uses the actual (correct) SNP set X_j for target j; and (ii) the second adversary uses the inferred SNP set X'_j as a result of the attribute inference attack using the noisy MAF queries (in Section 6). We also let each adversary follow two different methods for the attack. In the first method, the adversary sends its queries for m SNPs by randomly choosing the SNPs from set $\mathbf{A}$. In the second method, the adversary estimates the MAF values of the SNPs in the case group (M'_S) using MAF values of the population (M_P) and MAF values of the control group (M_C) by applying Equation 7 suggested by Sankararaman *et al.* (2009). Then, the adversary identifies some 'selected SNPs' to use in its queries. A SNP i is categorized as a selected SNP as follows:

$$\text{Selected SNPi} = \begin{cases} \text{True} & \text{if } M'_S \geq M_C \text{ and } x_j^i = 1 \\ \text{True} & \text{if } M'_S < M_C \text{ and } x_j^i = 0 \\ \text{False} & \text{otherwise} \end{cases}$$

The adversary first sends its queries for the selected m SNPs. Then, (if needed) the adversary sends more queries for other $|\mathbf{A}| - m$ SNPs in $\mathbf{A}$. Using the LLR metric described in Section 7.1, we evaluate the success of the membership inference attack.

7.3 Experimental results

In Figure 8, we show the power curves for the membership inference attack considering the aforementioned two adversaries, each at 10% false-positive rate (i.e. $\alpha = 0.1$). We start with the first adversary which uses the actual (correct) SNP values in set X_j for target j. Figure 8(a) shows the power for different numbers of queries over D_3 for randomly chosen SNPs and a selected set of SNPs. For each method, the values in M_S are shared (in a differentially private way)

with the adversary using four different ϵ (0.1, 2, 5 and when $\epsilon \to \infty$). Using the results of MAF queries over dataset $D3$, we make the following key observations: (i) Using the actual (correct) SNP values in set X_j for target j and the correct values in M_S (when $\epsilon \to \infty$), it is possible to correctly determine the membership of the target (with power more than 0.8) by sending either 190 queries for random SNPs or 93 queries for selected SNPs followed by 27 queries for random SNPs. (ii) The adversary can achieve a strong power to infer the membership of the target if it starts to query for the selected SNPs rather than random ones. (iii) As expected, decreasing the privacy budget (ϵ) used for releasing the differentially private values in M_S from 5 to 0.1 results in a 80% loss of power to infer the membership of target j in the case group.

Figure 8(b) shows the power for different numbers of queries over D_3, when the adversary uses the target's inferred SNPs in X'_j. The inferred SNPs are computed using four different ϵ values (0.1, 2, 5 and when $\epsilon \to \infty$) over the MAF queries (as shown in Section 6.3.1). In this experiments, we assume that the values in M_S are shared (in a differentially private way) with the adversary using $\epsilon = 5$ for all cases. We make four key observations: (i) Here, the adversary has fewer queries to send because it may not infer the whole SNPs in X_j for the target (for some received MAF responses in the attribute inference attack, the adversary cannot make any inference). For instance, when $\epsilon = 5$ in the attribute inference attack, out of 250 SNPs, the adversary can infer 220 of them. In spite of this, it can still achieve a high power in the membership inference attack ($1 - \beta = 0.8$) after sending 170 queries. (ii) First querying for the selected SNPs leads to a high power using less number of queries. (iii) The adversary can achieve a high power to detect the membership of target j in the case group even if part of the SNP values in X'_j are inferred wrong (as a result of the attribute inference attack). (iv) In-line with Figure 8(a), for smaller values of the privacy budget (ϵ) in the attribute attack, it becomes more challenging to determine the membership of target j in the case group. When the privacy budget is less than 2, power does not exceed 0.7 after 250 queries, and hence the adversary needs more inferred SNPs to infer the membership of the target.

8 Conclusion

In this article, we have demonstrated the limitations of the state-of-the-art DP-based mechanisms for genomic datasets with dependent tuples. For this, first, we have identified attribute inference attacks using two complex queries over real-life genomic datasets. We have shown how kinship relationships between individuals in a genomic dataset cause a significant reduction in the privacy guarantees of traditional DP-based mechanisms, and hence help an adversary infer sensitive genomic data of dataset participants. Furthermore, we have shown that these inferred genomic records (as a result of the attribute inference attack) can be utilized by the attacker to perform successful membership inference attacks to other statistical datasets. Our results also show the shortcomings of traditional DP-based mechanisms for privacy-preserving data sharing from statistical genomic datasets that include dependent tuples. As future work, we plan to consider more specific inheritance patterns and disease-related genomic features to study the genomic correlation probabilities between tuples.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Almadhoun,N. *et al.* (2020) Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics*, **36**, 1696–1703.

Backes,M. *et al.* (2016) Membership privacy in microrna-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Association for Computing Machinery, New York, NY, USA, pp. 319–330.

Browning,B.L. *et al.* (2018) A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, **103**, 338–348.

Corpas,M. (2013) Crowdsourcing the corpasome. *Source Code Biol. Med.*, **8**, 13.

D'Errico,J. (2018) Partitions of an integer. *MATLAB Central File Exchange*, **12009** https://ch.mathworks.com/matlabcentral/fileexchange/12009-partitions-of-an-integer (7 October 2019, date last accessed).

Drmanac,R. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.

Dwork,C. (2008) Differential privacy: a survey of results. In: *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer, Berlin, Heidelberg.

Dwork,C. *et al.* (2006) Calibrating noise to sensitivity in private data analysis. In: *Theory of Cryptography Conference*, pp. 265–284. Springer, Berlin, Heidelberg.

Fredrikson,M. *et al.* (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: *USENIX Security Symposium*, USENIX Association, Berkeley, CA, USA, pp. 17–32.

Goldstein,D.B. *et al.* (2013) Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.*, **14**, 460–470.

Gymrek,M. *et al.* (2013) Identifying personal genomes by surname inference. *Science*, **339**, 321–324.

Homer,N. *et al.* (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.*, **4**, e1000167.

Humbert,M. *et al.* (2015) De-anonymizing genomic databases using phenotypic traits. *Proc. Priv. Enhanc. Technol.*, **2015**, 99–114.

Jimenez,R.C. *et al.* (2011) myKaryoView: a light-weight client for visualization of genomic data. *PLoS One*, **6**, e26345.

Johnson,A. *et al.* (2013) Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, pp. 1079–1087.

Kifer,D. *et al.* (2011) No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, pp. 193–204.

Liu,C. *et al.* (2016) Dependence Makes You Vulnberable: Differential Privacy Under Dependent Tuples. 23rd Annual Network and Distributed System Security Symposium, (NDSS), 2016, San Diego, California, USA, February 21–24, 2016, The Internet Society, VA, USA.

McSherry,F. *et al.* (2007) Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science, 2007. FOCS'07*. IEEE, Los Alamitos, CA, USA, pp. 94–103.

Naveed,M. *et al.* (2015) Privacy in the genomic era. *ACM Comput. Surv. (CSUR)*, **48**, 1–44.

Neyman,J. *et al.* (1933) IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A*, **231**, 289–337.

Nissim,K. *et al.* (2007) Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, ACM, New York, NY, USA, pp. 75–84.

Sankararaman,S. *et al.* (2009) Genomic privacy and limits of individual detection in a pool. *Nat. Genet.*, **41**, 965–967.

Song,S. *et al.* (2017) Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, ACM, New York, NY, USA, pp. 1291–1306.

Stoeklé,H.C. *et al.* (2016) 23andMe: a new two-sided data-banking market model. *BMC Med. Ethics*, **17**, 19.

Uhler,C. *et al.* (2013) Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confid.*, **5**, 137.

Wang,R. *et al.* (2009) Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, Association for Computing Machinery, New York, NY, USA, pp. 534–544.

Yu,F. *et al.* (2014) Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inf.*, **50**, 133–141.

Zhao,J. *et al.* (2017) Dependent differential privacy for correlated data. In: *2017 IEEE Globecom Workshops (GC Wkshps)*, IEEE, Los Alamitos, CA, USA, pp. 1–7.

Zhou,X. *et al.* (2011) To release or not to release: evaluating information leaks in aggregate human-genome data. In: *European Symposium on Research in Computer Security*, Springer, Berlin, Heidelberg, pp. 607–627.