Bioinformatics, 2019, 1–8
doi: 10.1093/bioinformatics/btz837
Advance Access Publication Date: 8 November 2019
Original Paper

Genome analysis

Differential privacy under dependent tuples—the case of genomic privacy

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

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

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 17, 2019; revised on November 2, 2019; editorial decision on November 4, 2019; accepted on November 6, 2019

Abstract

Motivation: The rapid progress in genome sequencing has led to high availability of genomic data. Studying these data can greatly help answer the key questions about disease associations and our evolution. However, due to growing privacy concerns about the sensitive information of participants, accessing key results and data of genomic studies (such as genome-wide association studies) is restricted to only trusted individuals. On the other hand, paving the way to biomedical breakthroughs and discoveries requires granting open access to genomic datasets. Privacy-preserving mechanisms can be a solution for granting wider access to such data while protecting their owners. In particular, there has been growing interest in applying the concept of differential privacy (DP) while sharing summary statistics about genomic data. DP provides a mathematically rigorous approach to prevent the risk of membership inference while sharing statistical information about a dataset. However, DP does not consider the dependence between tuples in the dataset, which may degrade the privacy guarantees offered by the DP.

Results: In this work, focusing on genomic datasets, we show this drawback of the DP and we propose techniques to mitigate it. First, using a real-world genomic dataset, we demonstrate the feasibility of an inference attack on differentially private query results by utilizing the correlations between the entries in the dataset. The results show the scale of vulnerability when we have dependent tuples in the dataset. We show that the adversary can infer sensitive genomic data about a user from the differentially private results of a query by exploiting the correlations between the genomes of family members. Second, we propose a mechanism for privacy-preserving sharing of statistics from genomic datasets to attain privacy guarantees while taking into consideration the dependence between tuples. By evaluating our mechanism on different genomic datasets, we empirically demonstrate that our proposed mechanism can achieve up to 50% better privacy than traditional DP-based solutions.

Availability and implementation: https://github.com/nourmadhoun/Differential-privacy-genomic-inference-attack.

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

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Today's high-throughput sequencing platforms are capable of generating a tremendous amount of sequencing data (Alser *et al.*, 2017). These technologies allow sequencing the full human genome for as little as few hundred dollars (Hert *et al.*, 2008). As a result, production of genomic information for research, clinical care and recreational purposes at a rapid pace is no longer impossible from a technical point of view (Alser *et al.*, 2019). One of the most prominent uses of genomic data is for research purposes and to make such research initiatives successful, researchers need individuals donate their genomic data. Several studies report the attitudes of public in different countries (including USA, Sweden, Japan and Singapore) toward genomic research and their willingness to donate genomic samples (Carey *et al.*, 2016; Ishiyama *et al.*, 2008; Kobayashi and Satoh, 2009; Kraft *et al.*, 2018; Nanibaa *et al.*, 2016; Pulley *et al.*, 2008; Rahm *et al.*, 2013; Storr *et al.*, 2014). Although the majority of respondents show positive attitude toward genomic research and participating in such studies, the overwhelming majority of them have ranked privacy of sensitive information as one of their top concerns. Therefore, proper and privacy-preserving management of the personal information is necessary in order to attain public support for genomic studies. In addition, transparency of the research aim and proper management of genomic data utilization should be also maintained in order to not utilize the data beyond the donor's intention (Alser *et al.*, 2015).

The availability of human genomic banks provides an adequate basis for several important applications and studies

(Commission *et al.*, 2003). Genome-wide association study (GWAS) is considered as one of the most widely conducted genomic studies. These studies help scientists uncover associations between differences in the human genomes called single nucleotide polymorphisms (SNPs) and disorders that are passed from one generation to the next. We provide a brief background on genomics in Supplementary Section S1.1. Since the first GWAS in 2005 (DeWan *et al.*, 2006), researchers have assumed that it is safe to publish aggregate statistics about the SNPs that they found relevant to particular diseases and its associated phenotypes. A typical GWAS compares the genomes of individuals that carry a disease (cases) with genomes of healthy individuals (controls). Because the reported aggregate statistics were pooled from thousands of individuals, researchers believed that their release would not compromise the participants' privacy. However, such belief was challenged when Homer *et al.* (2008) demonstrated that, under certain conditions, given an individual's genotype, one only needs the minor allele frequencies of the SNPs used in the study and other publicly available information in order to determine whether the individual is in the case group of a GWAS. After this attack, the NIH restricted the access to key results and data of GWAS to only trusted individuals.

The purpose of this access policy is mainly due to the growing privacy concern about the participants in any genomic studies and their sensitive information, such as their health status. However, accelerating the pace of biomedical breakthroughs and discoveries necessitates not only collecting millions of genomic samples, but also granting an open access to the genomic banks and datasets (Galperin *et al.*, 2015).

There has been a growing interest in applying different privacy-preserving techniques to the GWAS results in order to grant access to genomic datasets. Many works in the literature propose utilizing the differential privacy (DP) notion (Dwork, 2008) to provide formal privacy guarantees for the participants of genomic studies. In a nutshell, DP guarantees that the distribution of query results change only slightly with the addition or removal of a single individual's data in the dataset. Although DP mechanism provides formal guarantees to preserve privacy (Dwork, 2008), it does not consider the dependency of the data tuples in the dataset. In reality, data from different users in the datasets may be dependent according to social, behavioral and genomic interactions between them (Liu *et al.*, 2016; Lv and Zhu, 2019; Zhao *et al.*, 2017). For example, in social network datasets, 'friendship' relation may imply similar interests (Chaabane *et al.*, 2012). Moreover, one can infer the locations of an individual from the friends' locations since they are likely to visit the same places (Liu *et al.*, 2016; Olteanu *et al.*, 2017). Similarly, in medical studies, an adversary may infer the susceptibility of an individual to a contagious disease by using the correlation between genomes of family members (Humbert *et al.*, 2013; Kifer and Machanavajjhala, 2011). These facts about the effect of correlation between tuples and data privacy was first observed by Kifer and Machanavajjhala (2011). Later, other researchers (Liu *et al.*, 2016; Song *et al.*, 2017; Zhao *et al.*, 2017) show that one can take advantage from dependencies between users to predict the users' sensitive information from the differentially private query results.

In this work, we formalize the DP concept to handle probabilistic dependence relationships between tuples in genomic datasets. We develop an effective perturbation mechanism to achieve the privacy guarantees in DP for datasets with dependent tuples. Our mechanism uses a carefully computed dependence coefficient that quantifies the probabilistic dependence between tuples in a fine-grained manner. The contributions of our paper are as follows:

- We demonstrate the feasibility of an inference attack on differentially private query results by exploiting the dependence between tuples in a real-world genomic dataset. We assume that the goal of the adversary is to infer the genomic data of a target individual using query results from a statistical genomic dataset. We also assume that the dataset includes correlated individuals (i.e. family members of the target individual). We show that the adversary can infer significantly more genomic data about the target from the results of queries by only exploiting the correlations between the genomes of family members. Moreover, we show that a stronger adversary with partial prior information about the genomic data of family members can infer even more sensitive data.
- We formalize the notion of ϵ-DP for genomic datasets with dependent tuples to avoid inference of sensitive information by any adversary with prior knowledge about the dependency between tuples. Our proposed mechanism computes the 'adjusted' ϵ value that provides privacy guarantees in the existence of dependent tuples in the dataset. That is, according to the number of dependent tuples in the dataset and their relationships, our mechanism allows accurate computation of the ϵ values for dependent data to preserve the privacy of the dataset participants while maintaining the utility of the data.
- We evaluate our mechanism over two different real-world genomic datasets. We demonstrate that it can be applied to any genomic statistics dataset with dependent tuples. Applying the proposed mechanism can provide better privacy and utility guarantees compared to other state-of-the-art DP-based mechanisms.

2 Related work

In this section, we will summarize the existing work on DP and genomic privacy in general. We will also highlight the differences of this paper from the existing work.

2.1 Privacy of genomic data

Privacy of genomic data has recently been a trending research topic (Erlich and Narayanan, 2014). There has been also a growing interest in applying the concept of DP to different genomic studies (Johnson and Shmatikov, 2013; Uhlerop *et al.*, 2013; Yu *et al.*, 2014). Existing work mainly consider DP as a protective measure against the inference attack discovered by Homer *et al.* (2008). Uhlerop *et al.* (2013) and Yu *et al.* (2014) developed many differentially private algorithms that can be applied to release the statistical results genomic studies, such as GWAS. For instance, according to Uhlerop *et al.* (2013) and Yu *et al.* (2014), Laplace noise with scale $2/\epsilon$ can be applied in order to get differentially private cell counts from genomic datasets. In general, these works develop algorithms try to achieve DP when releasing statistics about genomic datasets or studies. However, they do not consider the correlation between the dataset tuples, and hence their privacy guarantees weaken when such correlations exists within the dataset.

2.2 Differential privacy

Many techniques are proposed to achieve DP for many data types (Dwork, 2008).

Inference attacks against DP. The auxiliary information the adversary may learn from other channels is a big challenge. For instance, Fredrikson *et al.* (2014) use differentially private query results to infer a patient's genomic marker by utilizing additional information about the patient demographic information.

The strong dependence between the tuples in the real-world datasets introduces many privacy inference attacks. Kifer and Machanavajjhala (2011) were the first to criticize the independent tuples assumption of DP. Liu *et al.* (2016) consider predicting the user location from the differentially private clustering query results by utilizing pairwise dependencies between users using Gowalla dataset (Liu *et al.*, 2016).

Handling dependent tuples for DP. Handling dependent tuples is a significant challenge to guarantee privacy. Kifer and Machanavajjhala (2012) propose the Pufferfish framework (Kifer and Machanavajjhala, 2012) as a generalization of DP to provide rigorous privacy guarantees against adversaries with access to any auxiliary background information and have a belief about the relationships between data tuples.

However, no perturbation algorithm is proposed to handle the tuple dependencies. Blowfish (He *et al.*, 2014) is a subclass of Pufferfish, considering the data correlations and adversarial prior knowledge specified by the users in the form of deterministic constraints. He *et al.* (2014) provide perturbation mechanisms to handle these constraints. Chen *et al.* (2014) handle the correlation in network data using DP by multiplying the original sensitivity of the query with the number of correlated records. This approach results in deteriorating the utility performance of the shared query results since an excessive amount of noise is added to the dataset. Bayesian DP (Yang *et al.*, 2015) uses a modification of Pufferfish. Yang *et al.* (2015) propose a perturbation mechanism which considers the adversary's prior information and the correlations between data tuples. They only focus on the data correlations which can be modeled by Gaussian Markov Random Fields. To quantify the privacy loss when applying the traditional DP for continuous aggregate data release, Cao *et al.* (2017) consider the temporal correlation which can be also modeled by a Markov Chain. Liu *et al.* (2016) define the dependent differential privacy (DDP) to protect the privacy of an individual's location information in a correlated dataset. They propose a Laplace mechanism to tackle the pairwise correlations in the dataset by computing the distance between any two tuples. Recently, Song *et al.* (2017) concretized the Pufferfish privacy. They propose the Wasserstein mechanism. The definition of the ϵ-DP for correlated data in Song *et al.* (2017) is the same as in Liu *et al.* (2016). To satisfy that definition, the Wasserstein mechanism offers a weaker privacy budget. Zhao *et al.* (2017) improve the prior work of Liu *et al.* (2016) by presenting a new definition of DDP. The privacy guarantees of DDP address any adversary with arbitrary correlation knowledge. They propose using the Laplace mechanism to handle the numeric queries and exponential mechanism to handle the non-numeric ones. However, these studies (Liu *et al.*, 2016; Song *et al.*, 2017; Zhao *et al.*, 2017) provide less privacy and utility than our mechanism as we show in Section 6.3.

2.3 Contribution of this work

In this work, we demonstrate an inference attack using real-life genomic data on sensitive differentially private queries considering not only pairwise correlation as in Liu *et al.* (2016), but also interdependent data tuples in the dataset. We propose an effective Laplace mechanism to achieve DP for any genomic dataset with correlated tuples. Our mechanism is computationally efficient and it outperforms existing work in Chen *et al.* (2014); Liu *et al.* (2016); Song *et al.* (2017) and Zhao *et al.* (2017) both in terms of privacy and data utility (as shown in Section 6.3).

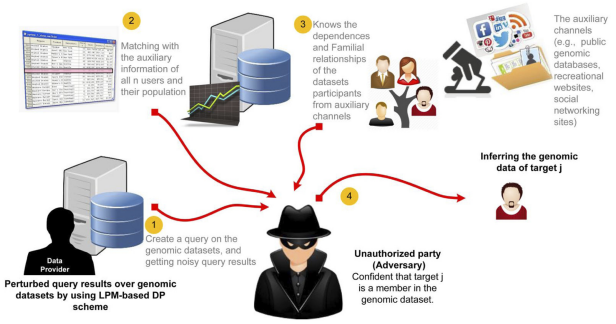

Fig. 1. The threat model. The adversary does not have any prior knowledge about the genomic data of target j, but it may have partial prior knowledge K for other members' genomic data. First, the adversary sends a query to the data provider. The data provider sends back the results with added noise using LPM. Second, the adversary identifies the individuals that are used to generate the query result using the metadata that is released along with the dataset (e.g. population). That is, the adversary identifies how many of the target's family members and unrelated individuals are used to generate the query result. Next, the adversary uses other auxiliary channels to learn the familial relationship of target j with his family members that are (i) in the dataset and (ii) used to generate the query result. Finally, using the noisy query results along with the auxiliary information and the probabilistic dependence between tuples, the adversary infers the genomic record of target j

3 Threat model

Based on the noise added to the query results, the DP mechanism probabilistically guarantees that users' sensitive data are protected regardless of adversary's prior knowledge about the dataset. However, the privacy guarantees provided by the existing DP mechanisms do not account for the dependence between the data tuples. They assume that the dataset tuples are independent. In fact, this assumption can degrade the privacy of the data from different users as they can be dependent due to various interactions.

An adversary can use auxiliary information channels to learn about such dependencies in the dataset and exploit the vulnerabilities in DP mechanisms as illustrated by Liu *et al.* (2016). Two major threats against statistical datasets are membership inference and attribute inference. In this work, we do not consider membership inference attacks, and we focus on the attribute inference attacks. The goal of the adversary in our model is to infer genomic data of a target individual.

We follow the same attack model in Liu *et al.* (2016). We assume that the adversary has access to the membership of all participants in the dataset of n individuals. This may be possible by using the metadata that is released along with the dataset (e.g. in 1000genome phases, metadata includes the populations of the dataset members). However, the adversary in our threat model is more powerful tzhan the DP adversary since he/she can also access auxiliary channels to estimate the relationship (or dependency) between tuples. To attain his goal, the adversary in our model will exploit the presence of target's family members in the same dataset and apply Mendelian inheritance rules to estimate the SNP values of the target. For all Mendelian inheritance probabilities see Supplementary Figure S1 in Supplementary Section S1.1. With this adversary model, we first perform an inference attack on the Laplace perturbation mechanism (LPM)-based differentially private data release to demonstrate that a powerful adversary can extract more information than that guaranteed by DP.

In our attack scenario (Fig. 1), the adversary is confident that the target j is a member of the dataset and some of his family members are also in the dataset. Also, the adversary may have some prior knowledge about the genomic data of target's family members. We represent the amount of such information as (i.e. K represents the fraction of prior information of the adversary about the genomic data of target's family members). The adversary combines the released noisy query results (that are compliant with DP) with knowledge of the existing dependence relations to infer the genomic data of the target (which is not available to the adversary before the attack).

4 Dataset description

For the evaluation, we use the genomic data of the family members from two datasets. Then, to get the unrelated members' genomic data, we use another dataset. Finally, we combined the family genomic data with the others genomic data. Hence, our final two datasets contain the partial DNA sequences from three sources:

- 1000Genome phase 3 data
- CEPH/Utah Pedigree 1463
- Manuel Corpas (MC) Family Pedigree

4.1 1000Genome phase 3 data

We use data from 1000Genome phase 3 with 2504 individuals from 26 populations. We extract the genotypes from chromosome 1 and chromosome 22 using the Beagle genetic analysis package (Browning *et al.*, 2018) to convert the values of genotypes to 0, 1 or 2 according to the minor alleles on each SNP. We use this data to include more participants to our dataset from the same or different population of the target and his family members. The main objective here is to test if the adversary can infer more sensitive information about the target even if the query results contain more unrelated participants.

4.2 CEPH/Utah Pedigree 1463

We use CEPH/Utah Pedigree 1463 with the partial DNA sequences of 10 family members (Drmanac *et al.*, 2010). In our inference attack, we consider the parent to be our target (Par 1 in Supplementary Fig. S2 in Supplementary Section S2.1). We only focus on first-degree relatives, and hence we use the genomic records of one parent, two grandparents and seven children (the original CEPH/Utah Pedigree 1463 includes data for 11 children, we randomly select 7 of them for our evaluation). We obtain the SNPs data for 10 individuals from the variant call format file. We select 100 common SNPs between 1000Genome members and UTAH family members to apply our inference algorithm. More details about the family structures are discussed in Supplementary Section S2.1.

4.3 MC family Pedigree

A scientist named MC (Corpas, 2013) decided to release his family DNA dataset for research purposes. The dataset contains the DNA sequences in variant call format for the father, mother, son (MC), daughter and aunt. We choose the son to be our target and we used the genomic records of his first-degree family members (father, mother and sister). Similar to the Utah family dataset, we extract the common 100 SNPs in all MC family members' 1000Genome members for the evaluation of our inference algorithm. More details about the family structure are discussed in Supplementary Section S2.2.

5 DP under dependent tuples

As we discussed in Section 3, DP mechanism does not account for the dependency of the data tuples in the dataset. On the other hand, family members' genotypes are inherently correlated and this correlation is stronger between close family members. Thus, existence of individuals from a target individual's family may provide an important source for an adversary to infer the target's genomic data even though their genomic data are not known by the adversary. This privacy breach has been proven by Humbert *et al.* (2013). In our scenario, the adversary sent his query asking about the total number of a specific SNP i for participants sharing the same demographic data, such as location or age.

The adversary gets the noisy result of his query, $\tilde{T}^i_{pj} = (T^i_p + T^i_j) + \delta$, for (p) participants included in the query results and individual j. δ represents the added Laplace noise with parameter $2/\epsilon$, T^i_j represents the SNP value for individual j, and T^i_p is the sum of the SNP values for other (p) participants. According to the query statement, the query results may include only the target j's related family members or also other unrelated individuals. Hence, the probabilistic dependence can be considered as:

$$T^i_p = T^i_j + Dy, \tag{1}$$

where $D = p$ if $p \leq 2$ and $D = 2p$ if $p > 2$ (p is the number of all individuals included in the query result except the target j). Also, y is a kinship coefficient that satisfies the Mendel's law. y is in $[-1, 1]$ for $p \leq 1$, and y is in $[0, 1]$ for $p > 1$.

5.1 Inference evaluation algorithm

We assume that the adversary can query the dataset based on the demographics of the dataset participants. As a result of his query, the adversary obtains the differentially private sum of genotype values $\tilde{T}^i_{pj} = (T^i_p + T^i_j)$ for different cases, e.g.:

- Total value of a SNP for people from same location area, or address.
- Total value of a SNP for people with the same age.

The adversary has access to auxiliary information about the membership of each participant including the target j, and also to the familial relationship between the target and other individuals in the dataset. Hence, the adversary can infer the value of T^i_j for target j using the number of dependent people related to that member in the dataset. We use two metrics to quantify the success of the attacks: correctness and leaked information. Correctness quantifies the distance *Dist* between the true value of the SNP and the inferred value by the adversary. The leaked information quantifies the change in the adversary's prior information after the inference attack. To measure the correctness we use the expected estimation error as follows:

$$E = \sum_{i=1}^{m} P(x_{ij}|T^i_j)|Dist(x_{ij}, x'_{ij})|. \tag{2}$$

To measure the leaked information we use the following equation:

$$L = \sum_{i=1}^{m} 1 - |sgn(Dist(x_{ij}, x'_{ij}))|, \tag{3}$$

where m is the number of targeted SNPs, and *sgn* denotes the sign function, which extracts the sign of any real number. *sgn* gives the value of 1 for all positive real numbers, 0 for number 0 and -1 for all negative real numbers. Hence in Equation (3), if there is any difference between x_{ij} which is the true value of SNP i for the target individual i and x'_{ij} which is the estimated value of SNP i for the target individual i, it means the adversary could not infer the correct value of the SNP and the SNP information is not leaked. We use Algorithm 1 in Supplementary Section S4 for evaluating the correctness and leaked information.

5.2 Evaluation

As discussed in Section 4, we use two datasets to evaluate the proposed attack model. We define both datasets as T that include n individuals (n= 2514 for the first dataset and n= 2508 for the second one). S is the set of SNP IDs on chromosome 1 and chromosome 22, and m is the number of SNPs for each individual ($m = 100$ for each dataset). To infer the values of these m SNPs, 100 queries are performed for each dataset. T^i_j represents the value of a SNP i ($i \in S$) for individual j ($j \in T$).

In the proposed inference attacks we assume the differentially private query results that are computed including individuals for different cases as follows:

Case 1: individual j with a direct family member.

Case 2: individual j with multiple family members.

Case 3: individual j with multiple family members, and other unrelated individuals.

We evaluate the performance of the attack for these cases considering two different types of attacks: (i) the adversary assumes that there is no correlation between individuals, and (ii) the adversary utilizes the genomic association between individuals to do genome reconstruction and infer genomic data. We use the algorithm described in Section 5.1 to quantify the success of the attacks by evaluating the two metrics: correctness and leaked information.

5.2.1 Experimental results

The adversary aims to construct individual j's genomic record, while the adversary only knows the membership of the individual and his family members in the dataset. We compare the dependent and independent assumptions to show the vulnerability of independent assumption and to come up with countermeasures for dependent cases. In Figure 2, we examine the effect of the number of relatives and non-relatives included in the result of a query to the target dataset on adversary's success in terms of his correctness in inferring SNPs of individual j.

We make three key observations: (i) in Figure 2a, the adversary is able to infer the targeted SNPs ($m = 100$) more accurately as the number of family members included in the query computation increases. We start with one first-degree relative who can be the father or the mother. Then, we gradually include the sons of individual j together with his father and mother. We observe that if the query results include data for more than four first-degree relatives of the same family, then the correctness of the adversary converges for $\epsilon \leq 2$. (ii) In Figure 2a, based on the correctness metric, we observe that if the adversary has the knowledge that the data

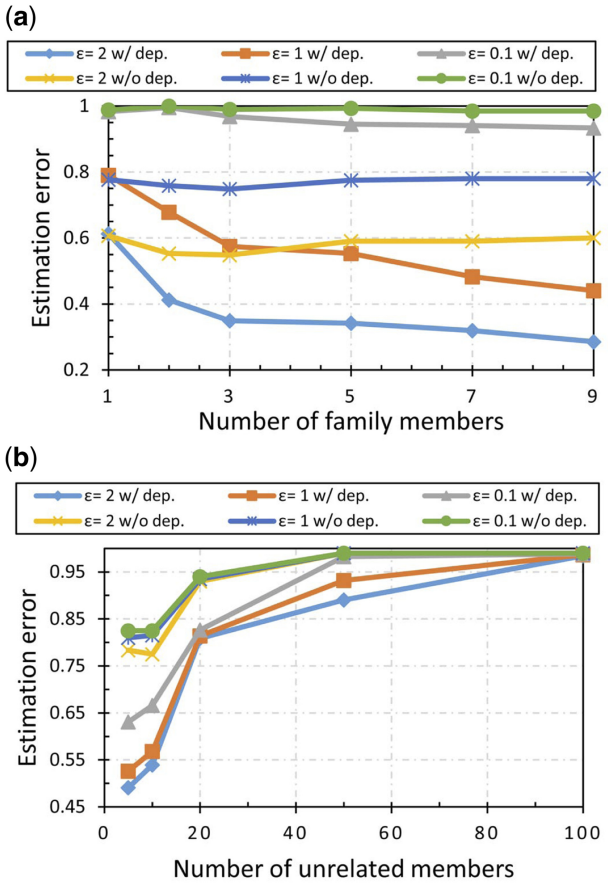

Fig. 2. The effect of (**a**) including only first-degree relatives and (**b**) including nine first-degree relatives with different numbers of non-relatives in the query results, on the probability of the adversary's correctness in inferring the targeted SNPs

of relatives (i.e. dependent tuples) exist in the target dataset, then the adversary's observation of the targeted SNPs is up to two times (depending on the value of ϵ) more accurate compared to not having this knowledge. As expected, we also observe that the difference between the correctness of the inferred SNPs with and without the knowledge of the data dependency increases as the value of the privacy budget, ϵ, increases. (iii) In Figure 2b, we observe that including the nine first-degree relatives and increasing the number of non-relatives included in results of the queries from 5 to 100 decreases significantly the ability of the adversary to infer the actual value of the targeted SNPs by about 20–50%, even if the adversary has the knowledge of the data dependency. Increasing the number of non-relatives beyond 100 members leads to the mitigation (with a probability of 0.99) of the leakage of SNP information of the participants. These heuristic results show the estimated scale of vulnerability that occurs when we have dependent tuples in a dataset that responds to queries based on DP.

Next, we evaluate the effect of different values of the privacy budget, ϵ, on the adversary's correctness in inferring the targeted SNPs. We show the results in Supplementary Section S5. The observations we make from these results are in accordance with our previous observations in Figure 2. After that, we evaluate the leaked information with different numbers of relatives and non-relatives included in the query results. We do not show the experimental results for leaked information metric due to space constraints. The results we obtain are compatible with the results of the correctness. That is, the adversary with the knowledge that the target dataset has dependent tuples can infer more SNPs as the number of family members included in the query results increases from 1 member to 9 members. Moreover, increasing the number of non-relatives in the query results decreases the number of leaked SNPs. The full details are provided in Supplementary Section S6 (Supplementary Figs S5 and S6).

Finally, we consider a stronger adversary who has access to partial information, e.g. $K = 50\%$ of other d family members' genomes included in the query results (as discussed in Section 3). The results are provided in Supplementary Sections S7 and S8. The results show that the adversary who considers the familial relationship between tuples in any genomic dataset can infer more information than the DP adversary. If the query results include many uncorrelated individuals with the target's relatives, it is more difficult for the adversary to infer the genomic record of the target. Moreover, for any adversary with prior partial information $K = 50\%$, the correctness of the targeted SNPs is considerably less (about 50%) for any adversary without any prior information $K = 0\%$. We further discuss the experimental results in Supplementary Section S9.

6 Countermeasures

As shown in Section 5, a genomic dataset with dependent tuples requires a stronger privacy notion than the existing DP mechanisms to get the same level of privacy guarantees. According to the evaluation results, using the adversary model described in Section 3, in this section, we formalize the notion of ϵ-DP for genomic datasets with dependent tuples to avoid inferring more sensitive information by an adversary with prior knowledge about the dependency between tuples. For any dataset T, we denote the number of dependent tuples in T as d (there may be different sets of dependent tuples in the dataset, we just focus on the largest set of dependent tuples with size d). We run the attack on a victim among these d dependent tuples. We define the dependence relationship between two tuples j and h as $R_{j,h}$, where R represents the familial relationships in real-world genomic datasets. In Section 5, we show an instance of R, where the dependence R can be known through the online information about the participants of the genomic studies and it can be formulated using the probabilistic dependence $T_p^i = T_j^i + Dy$. Like DP, ϵ-DP for genomic datasets with dependent tuples uses the notion of neighboring datasets, which can be defined as follows:

Definition 6.1. The datasets T and T' with d dependent tuples, which is the largest number of dependent tuples having probabilistic genomic relationship R, are neighboring dependent datasets if the change of one tuple value in T causes change of at most d-1 tuple values in T'.

Accordingly, we define the ϵ-DP for genomic datasets with dependent tuples as follows:

Definition 6.2. A randomized algorithm A satisfies ϵ-DP if for any pair of neighboring datasets T and T' with d dependent tuples, and for any $O \subseteq Range(A)$,

$$Pr[A(T) \in O] \leq e^{\epsilon} Pr[A(T') \in O].$$

Note that when R represents no dependency between tuples ($R = 0$), our privacy model is equivalent to DP mechanism. In order to restrict an adversary from inferring more sensitive information of an individual, we compute the value of the privacy budget (ϵ) for datasets that include dependent tuples so that the privacy guarantee will be the same as the datasets with independent tuples.

Analyzing LPM: recall the results we got for the threat model discussed in Section 3. We have a tuple T_j^i that has a probabilistic dependence relationship R with T_p^i as $T_p^i = T_j^i + Dy$, considering the result of a sum query where $Q(T) = (T_p^i + T_j^i)$. To achieve DP, we add Laplace noise with parameter $2/\epsilon$ for the sum query. We analyze the LPM-based DP mechanism while considering two assumptions:

- Independent tuples
- Dependent tuples

From the results, we have the following observations:

1. For any dataset with independent tuples, the noisy query output guarantees achieving DP with the same budget of ϵ value.
2. We need a smaller ϵ value to achieve DP for any dataset with dependent tuples. In other words, reducing the ϵ value used to achieve DP causes the Laplace noise to be augmented.
3. There may be different sets of dependent tuples in the dataset, according to the size of the largest dependent tuples, the added noise will be determined.

From our observations, we analyze the results of different queries for different number of dependent tuples. We compare the leaked information in a dataset assuming dependent and independent tuples in order to compute the DP sensitivity for different dependency size in a genomic dataset. The sensitivity can be defined as follows:

Definition 6.3. The dependent sensitivity for publishing the results of any query Q over a genomic dataset with correlated tuples is

$$\varsigma = \sigma \Delta Q, \tag{4}$$

where σ is the variable used to obtain the new value of ϵ. We describe computation of σ later. Also, ΔQ is the query Q's global sensitivity, which is the maximum difference in the query's result on any two neighboring datasets. Therefore, to achieve privacy guarantees in a genomic dataset, we formalize the mechanism to get ϵ-DP for genomic datasets with dependent tuples as follows:

Theorem 6.1. Let A be a randomized algorithm. Then, for a dataset T with d genomic dependent tuples, $A(T)$ provides ϵ-DP for a query Q with global sensitivity ς, if $A(T) = Q(T) + LAP(\varsigma/\epsilon)$, where ς is computed as in Equation (4).

Proof. We provide the proof in Supplementary Section S12.

Consider the leaked information an adversary can get without dependency assumption to be L_0, the leaked information any adversary can get with dependency assumption to be L_1 and the set of different ϵ values used in the query results over the dataset T to be ν. The following equation gives the value of σ:

$$\sigma = \nu / \left(\sum_{\epsilon \in \nu}^{|\nu|} L_0 / L_1 \right). \tag{5}$$

From our results, we calculate σ which allows accurate computation of the sensitivity for dependent data using 12 different ϵ values ($|\nu| = 12$).

6.1 Methodology for countermeasures

The data provider gets the query and identifies the largest set of dependent tuples in the query results. There may be more than one dependent tuples set (i.e. different sets of families) that are included in the calculation of the query results. We provide two practical strategies for computing the sensitivity. Based on the size of the dependent tuples the data provider can compute the value of σ. The data provider can select a proper model according to the query of the querier.

1. Using the query results over only the dependent tuples in the dataset: the data provider receives a query and observes the size of the largest dependent tuples set on it. He/she assumes that the querier has a complete knowledge about the correlation between tuples and the query results will only contain information from these dependent tuples.
   For example, in our evaluation scenario in which the maximum number of first-degree relatives that can be included in the same dataset together is nine. Hence, the data provider can compute the value of σ directly from the size of correlated tuples d as:
   $$\sigma = 0.219 \ln(d) + 1.4056. \tag{6}$$
   We show how Equation (6) is derived in Supplementary Figure S11a in Supplementary Section S10.
2. Using the query results over the dependent tuples and unrelated tuples in the dataset: The data provider assumes that the querier has a complete knowledge about the correlation between tuples, but the query results will contain information from these dependent tuples and other unrelated tuples. Here, we compute the σ value for six different values of unrelated members included in the results of query over the dataset T. The number of unrelated other members starts from 5 and gradually increases to 500. The data provider can compute the value of σ directly from the size of unrelated tuples u as:
   $$\sigma = -0.038 \ln(u) + 0.3337. \tag{7}$$
3. We show how Equation (7) is derived for this scenario in Supplementary Figure S11b in Supplementary Section S10.

6.2 Evaluation of countermeasures

In this section, we evaluate the performance of the proposed countermeasures to release the query results of dependent data over two real genomic datasets. We apply our algorithms over two datasets containing genomic data from (i) 1000Genome phase 3 data and CEPH/Utah Pedigree 1463 and (ii) 1000Genome phase 3 and MC Family Pedigree. We use 100 SNPs from chromosome 1 and chromosome 22 to analyze the resistance of our privacy mechanism to the threat model presented in Section 3.

Consider the first scenario of our privacy model in which the data provider publishes the perturbed genomic data of $T_p^i + T_j^i$ where the query results only contain information from these dependent tuples. We use the empirically determined values in Equation (6) to compute σ for different number of dependent tuples d. Then, according to Equation (4), we compute the dependent sensitivity ς. Figure 3 analyzes the amount of leaked information for individual j, an adversary can reconstruct from the perturbed query results assuming two cases for the dependent tuple size d. As before, we assume the adversary target to be the son of MC family. The first query results include the data of the target and his father. We can see under the same privacy budget, ϵ, our privacy model has much lower leaked information than the DP approach except for $\epsilon = 3$

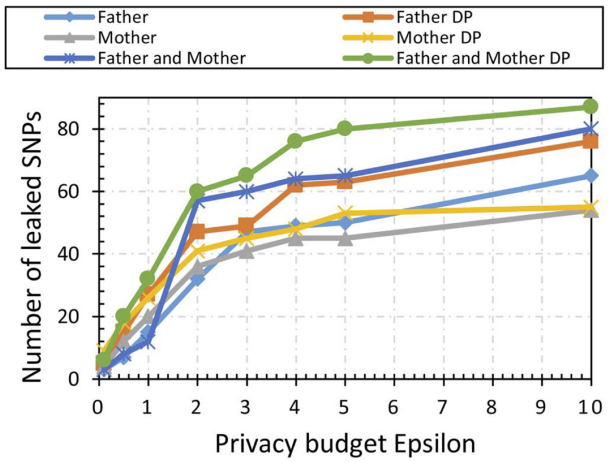

Fig. 3. The effect of applying our proposed countermeasure for different values of the privacy budget, ϵ. 'DP' lines stand for applying DP mechanism (over three different sets of family members in the dataset) and the other three lines show the leaked SNPs when our proposed mechanism is applied

where we get almost the same number of leaked SNPs. The second query results include the data of the target and his mother. Similarly, our privacy model achieves better privacy for various privacy budgets. In the third query results, the dependent tuples size in the dataset increased to three; we have the target, his father and mother included in the query results. As illustrated in Figure 3, we can confirm that our privacy model provides better privacy performance in terms of leaked information metric. In all different values of ϵ, using leaked information and correctness metrics our privacy model provides better privacy and it can reach up to 63% less leaked information in the case of $d = 3$ and $\epsilon = 1$. Therefore, our privacy model achieves better privacy guarantees than the existing approaches of DP for genomic studies, and this advantage increases for smaller ϵ values.

Next, we apply our mechanism to the second dataset, in which we target the par 1 in CEPH/Utah dataset and try to protect him against any inference attack aiming to detect his genomic data exploiting that his family members are included in the same genomic dataset. We assume eight cases for different numbers of correlated tuples d starting from two dependent tuples (the target j and one first-degree family member) and gradually increase until eight dependent tuples d (the target j and seven first-degree family members) in the dataset. Our model decreases the leaked information better than DP in all the eight cases. Hence, we increase the correctness of the adversary and decrease the leaked information about the genomic data of the target j. The results are shown in Supplementary Figure S12 in Supplementary Section S11.

6.3 Comparison with existing work

In the following, we compare our mechanism with the most similar existing work (Liu *et al.*, 2016; Zhao *et al.*, 2017) using a sum query over a dataset with $n = 1000$ tuples. Since Zhao *et al.* (2017) and Liu *et al.* (2016) consider Markov chain-based correlations, in their models, all 1000 tuples are correlated. Thus, for this comparison, we also report the results of our scheme for 1000 dependent tuples.

Figure 4 compares our mechanism with Zhao *et al.* (2017) and Liu *et al.* (2016) in terms of privacy (Fig. 4a) and utility (Fig. 4b).

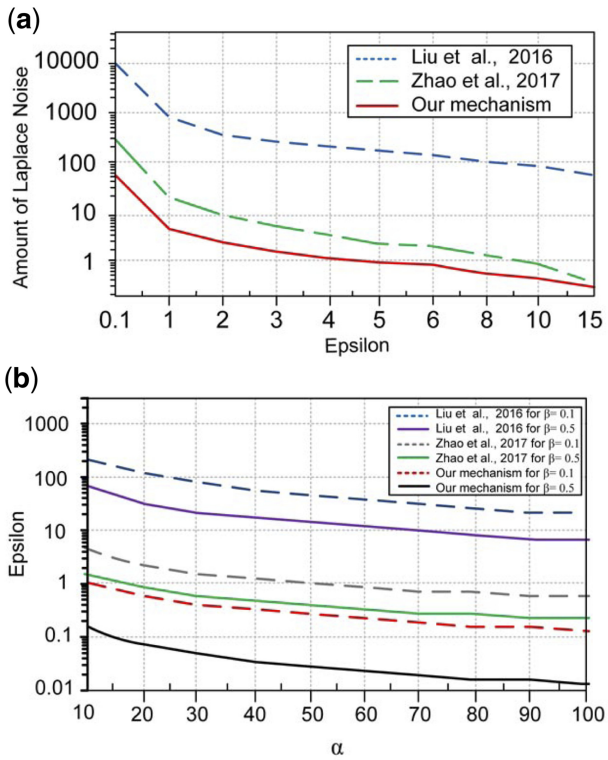

Fig. 4. (**a**) The amount of Laplace noise added for different values of privacy budget ϵ. (**b**) The privacy performance of different mechanisms which guarantee the (α, β)-usefulness. Here, the noisy output of the query should deviate by at most α from the real value (in terms of L1-norm) with probability $(1 - \beta)$

Figure 4a shows the amount of noise added to achieve ϵ-DP by considering the dependence between tuples. Here, we can see that for all ϵ values, our proposed scheme adds significantly smaller amount of noise, and hence provides better utility. For example, when $\epsilon = 0.1$, the amount of noise added in our scheme is 0.58% of the noise added by Liu *et al.* (2016) and 17.32% of the noise added by Zhao *et al.* (2017). Figure 4b shows the $(\alpha, \beta)-$ usefulness defined by Blum *et al.* (2013) which is commonly used for evaluating the utility guarantees for privacy mechanisms. It means the noisy output of the query should deviate by at most α from the real value (in terms of L1-norm) with probability $(1 - \beta)$. Figure 4b shows the smallest privacy parameter (ϵ) for different α values. For instance, to have $\alpha = 10$ and $\beta = 0.1$ (i.e. deviate by at most 10 from the original query result with a probability of 0.9), our proposed scheme requires a privacy budget of $\epsilon = 1.34$. To achieve the same $(\alpha, \beta)-$ usefulness, $\epsilon = 230$ shall be used for required of Liu *et al.* (2016) and $\epsilon = 3.8$ shall be used for required of Zhao *et al.* (2017). Thus, compared to existing work, for all α values, our mechanism requires a significantly smaller ϵ, and hence better privacy guarantees.

To sum up, our results demonstrate the following observations:

- Our model better minimizes the leaked information for genomic datasets compared to the state-of-the-art approaches (Chen *et al.*, 2014; Dwork, 2008; He *et al.*, 2014; Kifer and Machanavajjhala, 2012; Liben-Nowell and Kleinberg, 2007; Liu *et al.*, 2016; Zhu *et al.*, 2015). Thus, we can select an appropriate privacy budget to achieve the optimal desired privacy while maintaining utility of the data for different genomic applications.
- Our model can achieve up to 50% on average better privacy guarantees based on the estimated error and the leaked information metrics than DP approaches, based on the leaked information metric, for publishing the average number of SNP values for a group of members participating on any genomic studies.
- Our model is resistant to state-of-the-art inference attacks (Fredrikson *et al.*, 2014; Liu *et al.*, 2016). It reduces the leaked information even with a larger number of dependent tuples for various values of ϵ.

7 Conclusion

DP is considered as a concept that provides rigorous privacy guarantees. However, it suffers from weak privacy performance due to some limitations, such as ignoring the dependence between the tuples in the dataset. In this paper, we have utilized an inference attack to assess the vulnerability of the state-of-the-art DP-based approaches and we have shown the effect of data dependence on the genomic privacy. We have shown that an adversary, knowing the familial relationship between some individuals in a genomic dataset, may infer more information than what is guaranteed by traditional DP. To mitigate such privacy risks, we have introduced ϵ-DP for genomic datasets with dependent tuples that takes into consideration the probabilistic dependence relationship between data tuples and provides rigorous privacy guarantees. Furthermore, we have evaluated our perturbation mechanism over different genomic datasets. Our results show that our privacy model performs significantly better than the existing DP-based mechanisms.

Conflict of Interest: none declared.

References

Alser,M. *et al.* (2015) Can you really anonymize the donors of genomic data in today's digital world? In: *Data Privacy Management, and Security Assurance*. Springer, New York, pp. 237–244.

Alser,M. *et al.* (2019) Shouji: a fast and efficient pre-alignment filter for sequence alignment. *Bioinformatics*, **35**, 4255–4263.

Alser,M. *et al.* (2017) Gatekeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping. *Bioinformatics*, **33**, 3355–3363.

Blum,A. *et al*. (2013) A learning theory approach to noninteractive database privacy. *JACM*, **60**, 1.

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

Cao,Y. *et al*. (2017) Quantifying differential privacy under temporal correlations. In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. pp. 821–832. IEEE.

Carey,D.J. *et al*. (2016) The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. *Genet. Med*., **18**, 906.

Chaabane,A. *et al*. (2012) You are what you like! Information leakage through users' interests. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS), San Diego, California, USA*.

Chen,R. *et al*. (2014) Correlated network data publication via differential privacy *VLDB J*., **23**, 653–676.

Commission,A.L.R. *et al*. (2003) Essentially Yours–The Protection of Human Genetic Information in Australia, Vol. 1 and Vol. 2. *Report 96*.

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

DeWan,A. *et al*. (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, **314**, 989–992.

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.

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

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

Galperin,M.Y. *et al*. (2015) The 2015 nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Res*., **43**, D1–D5.

He,X. *et al*. (2014) Blowfish privacy: tuning privacy-utility trade-offs using policies. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. pp. 1447–1458. ACM.

Hert,D.G. *et al*. (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, **29**, 4618–4626.

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*. (2013) Addressing the concerns of the lacks family: quantification of kin genomic privacy. In: *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. pp. 1141–1152. ACM.

Ishiyama,I. *et al*. (2008) Relationship between public attitudes toward genomic studies related to medicine and their level of genomic literacy in Japan. *Am. J. Med. Genet. A*, **146**, 1696–1706.

Johnson,A. and Shmatikov,V. (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*. pp. 1079–1087. ACM.

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

Kifer,D. and Machanavajjhala,A. (2012) A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ACM, pp. 77–88.

Kobayashi,E. and Satoh,N. (2009) Public involvement in pharmacogenomics research: a national survey on public attitudes towards pharmacogenomics research and the willingness to donate DNA samples to a DNA bank in Japan. *Cell Tissue Bank*., **10**, 281.

Kraft,S.A. *et al*. (2018) Beyond consent: building trusting relationships with diverse populations in precision medicine research. *Am. J. Bioeth*., **18**, 3–20.

Liben-Nowell,D. and Kleinberg,J. (2007) The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Tec*., **58**, 1019–1031.

Liu,C. *et al*. (2016) Dependence makes you vulnerable: differential privacy under dependent tuples. In NDSS, Vol. **16**, pp. 21–24.

Lv,D. and Zhu,S. (2019) Achieving correlated differential privacy of big data publication. *Comput. Secur*., **82**, 184–195.

Nanibaa'A,G. *et al*. (2016) A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. *Genet. Med*., **18**, 663.

Olteanu,A.M. *et al*. (2017) Quantifying interdependent privacy risks with location data. *IEEE Trans. Mob. Comput*., **16**, 829–842.

Pulley,J.M. *et al*. (2008) Attitudes and perceptions of patients towards methods of establishing a DNA biobank. *Cell Tissue Bank*., **9**, 55–65.

Rahm,A.K. *et al*. (2013) Biobanking for research: a survey of patient population attitudes and understanding. *J. Community Genet*., **4**, 445–450.

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

Storr,C.L. *et al*. (2014) Genetic research participation in a young adult community sample. *J. Commun. Genet*. **5**, 363–375.

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

Yang,B. *et al*. (2015) Bayesian differential privacy on correlated data. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. pp. 747–762. ACM.

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

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

Zhu,T. *et al*. (2015) Correlated differential privacy: hiding information in non-IID data set. *IEEE Trans. Inf. Forensics Secur*., **10**, 229–242.