Bioinformatics, 33(23), 2017, 3716–3725
doi: 10.1093/bioinformatics/btx470
Advance Access Publication Date: 21 July 2017
Original Paper

OXFORD

Genome analysis

Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies

Meng Wang[1,*], Zhanglong Ji[2], Shuang Wang[2], Jihoon Kim[2], Hai Yang[2], Xiaoqian Jiang[2] and Lucila Ohno-Machado[2]

[1]Department of Genetics, Stanford University, Stanford, CA 94305, USA and [2]Department of Biomedical Informatics, UC San Diego, La Jolla, CA 92093, USA

Associate Editor: Bonnie Berger

Received on October 26, 2016; revised on May 29, 2017; editorial decision on July 14, 2017; accepted on July 20, 2017

Abstract

Motivation: Inappropriate disclosure of human genomes may put the privacy of study subjects and of their family members at risk. Existing privacy-preserving mechanisms for Genome-Wide Association Studies (GWAS) mainly focus on protecting individual information in case–control studies. Protecting privacy in family-based studies is more difficult. The transmission disequilibrium test (TDT) is a powerful family-based association test employed in many rare disease studies. It gathers information about families (most frequently involving parents, affected children and their siblings). It is important to develop privacy-preserving approaches to disclose TDT statistics with a guarantee that the risk of family 're-identification' stays below a pre-specified risk threshold. 're-identification' in this context means that an attacker can infer that the presence of a family in a study.

Methods: In the context of protecting family-level privacy, we developed and evaluated a suite of differentially private (DP) mechanisms for TDT. They include Laplace mechanisms based on the TDT test statistic, *P*-values, projected *P*-values and exponential mechanisms based on the TDT test statistic and the shortest Hamming distance (SHD) score.

Results: Using simulation studies with a small cohort and a large one, we showed that that the exponential mechanism based on the SHD score preserves the highest utility and privacy among all proposed DP methods. We provide a guideline on applying our DP TDT in a real dataset in analyzing Kawasaki disease with 187 families and 906 SNPs. There are some limitations, including: (1) the performance of our implementation is slow for real-time results generation and (2) handling missing data is still challenging.

Availability and implementation: The software dpTDT is available in https://github.com/mwgrassgreen/dpTDT.

Contact: mengw1@stanford.edu

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

1 Introduction

The continuous progress in genome technologies is rapidly reducing the cost of human genome sequencing (online source), and has produced massive human genome data (Clarke, 2012; Church, 2005). For example, the US government's Precision Medicine Initiative (Collins and Varmus, 2015) aims at developing treatments tailored to a person's genetic profile from a cohort of one million volunteers. Extensive collection of genome data and the development of advanced analysis techniques increase the probability of using human genome data for clinical diagnosis and treatments, but it also

leads to many privacy challenges. Many studies (Craig 2011; Gymrek, 2013; Homer, 2008; Gutmann, *et al.*, 2012, Wang, 2009) demonstrated the vulnerability of exposing human genome data without protection. For example, Lin *et al.* (2004) shows that as few as 75 independent SNPs can be used to uniquely identify an individual. Given genome data without explicit identifiers, it is possible to infer personal information (Gymrek, 2013; Sweeney, 2013). Even aggregated genome statistics can be used to recover sensitive personal information (Homer, 2008; Wang, 2009) and many recent attack models use allele frequencies (or absence/presence information such as particular variants) to reveal an individual's membership in case or control groups (Clayton, 2010; Homer, 2008; Jacobs, 2009; Sankararaman, 2009; Visscher and Hill, 2009). That is, an attacker can discover that a particular person of interest, from whom the attacker has gathered genetic information, participated in a study. This is problematic when the participants do not want to disclose their participation in a study. For example, in certain studies in which the 'controls' are individuals with a particular diagnosis receiving standard-of-care and the 'cases' are individuals with the same diagnosis receiving a new type of treatment, it is possible to infer whether a person of interest participated in the study (i.e. the person has the diagnosis) and whether she was in the case or control group when the attacker gets information about allele frequencies. For this reason it is important to develop secure and privacy-preserving methods (Chen, 2017; Wang, 2016; Zhang, 2015) to protect genome studies against emerging attacks (Humbert, 2013; Malin and Sweeney, 2001, 2004).

Differential privacy (DP) (Dwork, 2006) provides a rigorous framework to protect the genome database in the information disclosure phase. They introduce perturbations (i.e. noise) that make re-identification of targeted individuals as difficult as disclosers want (the higher the risk, the less noise is introduced and vice-versa). DP (Dwork, 2006) provides provable quantification of the privacy risk associated with disclosed information. Yu (2010) studied a DP-based logistic regression method to detect SNP associations in GWAS. A DP-based genomic data dissemination model was proposed in Wang (2014). In Johnson and Shmatikov (2013), a DP mechanism was developed to protect the outcome of chi-square test outcomes in GWAS. Uhler (2013) improved previous methodology to allow privacy-protecting release of the K most relevant SNPs. Yu (2014) demonstrated better performance in terms of privacy and utility tradeoffs, and presented formal proofs (Yu, 2014; Yu and Ji, 2014).

However, these current privacy-preserving mechanisms in GWAS only focus on case–control studies. As far as we know, none of the work in GWAS considers protecting privacy in family-based studies. Since the data are recorded and structured in the unit of a family in these studies, it is possible that an attacker may infringe the privacy of the whole family instead of that of just one individual. Hence, to avoid this, we need a reliable method to protect privacy at the family level. This paper focuses on protecting family privacy for transmission disequilibrium test (TDT) in GWAS.

The TDT is primarily designed to account for population stratification and detect potential Mendelian inconsistencies. It measures the over-transmission of an allele from heterozygous parents to an affected offspring (Spielman, 1993). It takes into account the family structure allowing for investigating heritability questions especially powerfully in the rare diseases such as Kawasaki disease, which cannot be addressed in the case–control studies (Ott, 2011). TDT has been successfully applied to whole exome sequencing studies in autism (Levin-Decanini, 2013) and Kashin-Beck disease (Yang, 2014).

To protect privacy in TDT, we consider that an attacker might violate the privacy of the entire family to figure out whether one family participates in the study, and not just an individual of the family. To protect family level privacy, we develop and analyze differentially private (DP) mechanisms for TDT based on test statistics, P-values and the shortest Hamming distance (SHD) scores. This is a non-trivial extension from individual privacy-protection methods because we need to preserve the family structure while introducing the perturbation. Under the DP framework, we define the neighbor datasets by replacing single family with other valid types to protect the privacy of trios (mother, father and one child) used in TDT. We carefully manipulate the sensitivity of the statistics to guarantee privacy while maximizing the utility. We develop both exact and approximate algorithms to get SHD scores for TDT. To maintain both privacy and utility (the usefulness of the method and a former definition is in the section 'Results'), our DP methods on TDT is preferable to be applied in a large cohort with sufficient number of families. We give an illustration to apply our methods in a real GWAS study dataset for the Kawasaki disease. On a proper cohort size, our proposed DP mechanisms based on SHD score can preserve higher privacy and better utility than other proposed methods for TDT in GWAS.

1.1 Transmission disequilibrium test

TDT is a family-based association test for linkage disequilibrium (LD). It tests at a marker locus for the LD in two alleles and as well as more alleles from the genotypes of trio families (two parents with one affected child) or other types of families (Spielman, 1993). In GWAS, we do not apply TDT only in one locus but in hundreds, thousands even hundreds of thousands loci. The process might reveal sensitive patient information and therefore we are proposing a private version of TDT to protect family information. Our analysis focuses on applying TDT to SNPs (considering two alleles) from trio families. Other extension work can be found in the 'Discussion' section.

In on SNP, TDT determines whether two alleles, W and w, have equal probability of being transmitted from the parents to their child. Let b be the number of the Ws transmitted when w is not transmitted and c be the number of w's transmitted when W is not transmitted. For example, in a family with genotype configuration $WW \times Ww \rightarrow WW$, one parent WW transmits one W to the child ($b = 1$) while w is non-transmitted from neither of the parents ($c = 0$). All possible parent genotypes are enumerated in the Appendix (Supplementary Table S1). Suppose there are N independent trio-families in our study. In Table 1 we gather the allele transmission (and non-transmission) information from these N families and give the numbers of families in each type of (b, c).

From Table 1, we get in one SNP from N trio-families,

$$b = n_1 \times 1 + n_2 \times 0 + n_3 \times 1 + n_4 \times 2 + n_5 \times 0 + n_6 \times 0;$$
$$c = n_1 \times 0 + n_2 \times 1 + n_3 \times 1 + n_4 \times 0 + n_5 \times 2 + n_6 \times 0;$$
$$N = n_1 + n_2 + n_3 + n_4 + n_5 + n_6.$$

Table 2 summarizes allele transmission in N trio-families. The TDT statistic in one SNP is based on

$$T := T(b,c) = \frac{(b-s)^2}{s} + \frac{(c-s)^2}{s} = \frac{(b-c)^2}{b+c}, \text{ where } s = \frac{b+c}{2}. \quad (1)$$

In the concern of privacy, we include the families with homozygous parents $b = c = 0$. We define $T = 0/0 = 0$. Under the null hypothesis that there is no linkage disequilibrium and the independence

Table 1. Number of Trio-families in one SNP

(b,c) in one family	$(1,0)$	$(0,1)$	$(1,1)$	$(2,0)$	$(0,2)$	$(0,0)$
# of families	n_1	n_2	n_3	n_4	n_5	n_6

Table 2. TDT contingency table from *N* trio-families in one SNP

	Non-transmitted allele		
Transmitted allele	W	w	Total
W	a	b	a + b
w	c	d	c + d
Total	a + c	b + d	2N

assumption among families, the counts b and c have binomial distribution with size b+c and probability 0.5 and thus the test statistic T in (1) approximately obeys a χ^2-distribution with one degree of freedom (Spielman, 1993). Under the alternative hypothesis, the counts b and c become unbalanced so the hypothesis is rejected for large values of T. Since both $(n_1,\ldots,n_6)$ and (b,c) are both sufficient statistics for TDT, we can represent the dataset by either one of them. The exact algorithm in the Section 2.2 is based on the representation of $(n_1,\ldots,n_6)$. In the other parts of this article we consider the dataset based on (b,c). We define the dataset in one SNP by $\mathcal{D} := \{(b,c) : b+c \leq 2N, b,c \in \mathbb{N}\}$ where $\mathbb{N}$ is the set of nonnegative integers. In the studying of family-based association in GWAS, we apply TDT for each SNP. For M SNPs, we represent the dataset D by $((b_1\ ,\ c_1),\ \cdots,\ (b_M\ ,\ c_M))$ and the data space is $\mathcal{D}^M$.

Now the family information is summarized in (b,c) pairs. From one pair, one can infer possible family genotypes and thus it leaves a potential risk for an attacker to identify families. In the framework of DP, Figure 1 illustrates the workflow for private data access from DP TDT. We take the query as to get the top K significant SNPs. The data owner first calculates the statistics T_i's for each SNP and next applies a DP mechanism (Laplace or exponential mechanism) to perturb the original T_i's. Based on the perturbed responses, the data owner reports to the data user the top K most significant SNPs. Our DP TDT is developed to protect the privacy of outcomes of data analyses. Instead of exposing the data to end users, we can hide them behind a query interface (for example, by providing results of certain statistical tests). It has been shown that repeated queries can lead to information leakage (Shringarpure and Bustamante, 2015). Our model is developed to protect such information leakage using the principled differential private criteria (Dwork, 2006).

1.2 Differential privacy model

Differential privacy (Dwork, 2006) provides guarantees on the privacy of whole database against any arbitrary external attacks. In practice, there are two main mechanisms (randomized functions) satisfying the differential privacy definition (in Definition 1). One is the Laplace mechanism (Dwork, 2006), which adds Laplace noise to the original output. The Laplace parameter depends on the sensitivity of the output in Definition 2. The other one is the exponential mechanism (McSherry and Talwar, 2007), which first assigns a score to each pair (*data, its output*). Then, the mechanism samples original outputs from an exponential distribution based on the score and the sensitivity of the score function (in Definition 2).

Definition 1: (Differential Privacy (Dwork, 2006)) A randomized mechanism $\mathcal{M}$ is ε-differentially private if, for all datasets D and D'

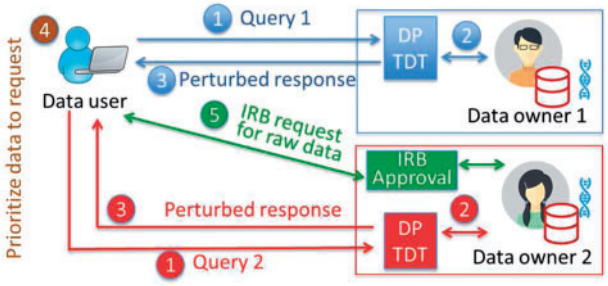

Fig. 1. A workflow for privacy-protecting data access from DP TDT. (1) A data user sends queries to different datasets. (2) Data owners compute the true results based on the query. (3) Instead of returning the true results, data owners return perturbed, DP TDT results. (4) The data user compares the utility among different datasets to prioritize requests for data. (5) The data user files a data access application to obtain participant-level data

which differ on at most one family and for any measurable subset $S \subset range(\mathcal{M})$,

$$\frac{\mathbb{P}(\mathcal{M}(D) \in S)}{\mathbb{P}(\mathcal{M}(D') \in S)} \leq e^{\varepsilon}. \tag{2}$$

In Definition 1, D and D' are called ***the neighbors*** and denoted $D \sim D'$. In our case, the dataset D gathers family information (in terms of (b,c) pairs defined in Section 1.1). Note that the exchange of two families may affect TDT on several SNPs so we protect privacy in each SNP. Parameter ε is called ***the privacy budget***. When ε is small, the probabilities in the numerator and denominator in (2) are similar and when $\varepsilon = 0$, they are identical. If a mechanism $\mathcal{M}$ controls the probability ratio in (2) under ε, we say $\mathcal{M}$ protects the privacy of D under ε-differentially privacy. ε is the budget that has to pay for the level of privacy. In other words, the privacy budget ε can be imagined as a pre-defined limit of information disclosure. We provide a short sketch on its relationship with the probabilities of type I and type II errors in any statistical test in the Appendix, within the section 'How privacy budgets affect probabilities of type I and type II errors'.

Definition 2: (Sensitivity for the Laplace Mechanism (Dwork, 2006)) The sensitivity of a function $f : \mathcal{D}^M \to \mathbb{R}^M$ in the Laplace Mechanism is the smallest number $S(f)$ such that

$$\| f(D) - f(D') \|_1 \leq S(f),$$

for all neighbors

$$D \sim D',\ \textit{and}\ D, D' \in \mathcal{D}^M.$$

Definition 3: (Sensitivity for Exponential Mechanism (McSherry and Talwar, 2007)) The sensitivity of a score function $q : \mathcal{D}^M \times \{1, \ldots,\ M\} \to \mathbb{R}$ in the exponential mechanism is the smallest number $S(q)$ such that

$$|q(D,r) - q(D',r)| \leq S(q),$$

for all neighbors $D \sim D'$, *and* $D, D' \in \mathcal{D}^M$ and $r \in \{1,\ \ldots,\ M\}$ is the index of the SNP.

Definition 4: (Laplace Mechanism (Dwork, 2006)) Releasing ($f(D) + noise$) where *noise* has $Laplace(S(f)/\varepsilon)$ distribution is an ε-differentially private mechanism satisfying Definition 1, where f is a function of data, $Laplace(\lambda = S(f)/\varepsilon)$ has density $\frac{1}{2\lambda}\exp\left(-\frac{|x|}{\lambda}\right)$, and $S(f)$ is the sensitivity of f from Definition 2.

Definition 5: (Exponential Mechanism (McSherry and Talwar, 2007)) Let $q : \mathcal{D}^M \times \{1, \ldots, M\} \to \mathbb{R}$ be a function assigning a score to the pair (D, r) where $D \in \mathcal{D}^M$ and $r \in \{1, \ldots, M\}$ is the SNP index associated with D. Choose r from the mechanism $\mathcal{M}_q^\varepsilon$ that has distribution

$$\mathbb{P}\left(\mathcal{M}_q^\varepsilon(D) = r\right) = \frac{\exp\left(\frac{\varepsilon q(D,r)}{2S(q)}\right)}{\sum_{s\in\{1,\ \ldots,\ M\}} \exp\left(\frac{\varepsilon q(D,s)}{2S(q)}\right)},$$

where $S(q)$ is the sensitivity of q from Definition 3. Releasing $\mathcal{M}_q^\varepsilon$ is an ε-differential privacy mechanism according to Definition 1.

Different from the Laplace mechanism, which directly adds noise on the original output, the exponential mechanism perturbs the probability of the outcomes by assigning scores to the original outputs such that the ones with high scores have higher probability to be sampled.

2 Materials and methods

In GWAS, we are often interested in a small number of significant SNPs. In this study, our goal is to answer the query about top K most significant SNPs based on a privacy-preserving TDT algorithm to avoid an attacker violating family information. Releasing differentially private top K significant SNPs has been considered in Bhaskar (2010). In GWAS case–control studies, there are several differentially private approaches. Uhler (2013) analyzed private allelic test and develop Laplace and exponential mechanisms based on the test statistics and P-values. Johnson and Shmatikov (2013) took another approach based on Shortest Hamming Distance (SHD). Yu (2014), Yu and Ji (2014) and Simmons and Berger (2016) extended and improved the SHD method in the allelic test. These previous works are in the case–control studies to protect individuals' privacy. To the best of our knowledge, we are the first to analyze differentially private TDT to protect the privacy of families.

Different from protecing individual information, to protect family privacy in TDT, the algorithms and computation are more involved. Under the framework of DP, we start our analysis from defining neighbor families. Suppose there are N trio-families and M SNPs. We define two neighbor datasets D, D′ by exchanging the genotypes of exact one trio-family. For example, suppose in one SNP, in D there is one family with $\mathrm{WW} \times \mathrm{Ww} \to \mathrm{WW}$ $((\mathrm{b}, \mathrm{c}) = (1, 0))$ and now in D′ it changes to $\mathrm{WW} \times \mathrm{Ww} \to \mathrm{Ww}$ $((\mathrm{b}', \mathrm{c}') = (0, 1))$ while D and D′ are consistent in all other families. In this case, changing from D to D′ gives a moving direction $\vec{\mathrm{v}} := (\Delta\mathrm{b}, \Delta\mathrm{c}) = (\mathrm{b}' - \mathrm{b}, \mathrm{c}' - \mathrm{c}) = (-1, 1)$. Considering all the family genotypes, we obtain 18 valid moving directions between two neighbor families as illustrated in Figure 2. Compared to neighbor individuals in the case–control studies, neighbor families have more moving directions making the privacy-protection algorithms much harder to develop. Details to obtain in the valid moving directions are listed in the Appendix.

2.1 Differentially private mechanisms based on the χ^2 − test statistics and P-values

With well-defined neighbor families, we perform sensitivity analysis (proofs in Appendix) on the TDT test statistics and P-values, which is the key element in applying Laplace and exponential mechanisms under the framework of DP.

Theorem 1: For one SNP, the sensitivity of the TDT statistic T in (1) for N trio-families is $8(N-1)/N$ given a fixed N with $N \geq 2$.

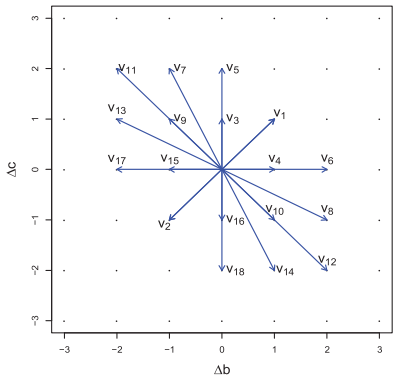

Fig. 2. Moving directions $\vec{v} = (\Delta b, \Delta c)$ among all neighbor families

Adding Laplace noise from distribution $Laplace(8(\mathrm{N}-1)/(\varepsilon N))$ to T achieves ε-differential privacy. We also consider releasing differentially private p-values, without perturbing the test statistics first.

Theorem 2: In one SNP, the sensitivity of p-value of χ^2-statistics with one degree of freedom for TDT is $F_{\chi_1^2}(4)(\approx 0.954)$ when $N \geq 4$, where $F_{\chi_1^2}(\cdot)$ is the cumulative density function for a χ^2 distribution with one degree of freedom.

From Theorem 2, we can see that the sensitivity of P-value (which is nearly 1) is quite large which may result in the perturbed outputs useless. Since in GWAS what we really concern are the SNPs in small P-values, following (Uhler, 2013), we consider a projected p-value of TDT test statistic t, defined by

$$P\mathrm{value}_{\mathrm{proj}}(t) = \min\{p\mathrm{value}(t), p^*\}, \tag{4}$$

where p^* is a predefined threshold.

Theorem 3: In one SNP, the sensitivity of the projected P-value defined in (4) is $|1 - F_{\chi_1^2}((t^*-4)^2/t^*) - p^*|$ for a predefined threshold p^*, and t^* is the quantile such that $F_{\chi_1^2}(t^*) = 1 - p^*$ when $N \geq 2$.

From our analysis, the most sensitive case for the test statistic corresponds to $(\mathrm{b}, \mathrm{c}) = (0, 2\mathrm{N})$ *or* $(2N,\ 0)$where b and c has the largest difference, while the case for the p-value corresponds to $(\mathrm{b}, \mathrm{c}) = (2, 2)$ where b and c has the smallest difference, which agrees with the finding in Uhler (2013) in allele-frequency test. To releasing top K significant SNPs based on test statistics and$P$-values, we adapt the procedures in Bhaskar (2010) and summarize the Algorithms 1–2.

Algorithm 1: ε-Differentially Private Algorithm for Releasing the K Most Significant SNPs Using the Laplace Mechanism.

Input: The number K of significant SNPs to release, the data function f (χ^2 statistics or P-values) for all M SNPs, the sensitivity of the data function S(f), privacy budget ε.
Output: Top K most significant SNPs after adding noise.

1. Let $\mathrm{f_i}$ be the data function for i-th SNP.
2. Add independent Laplace noise with parameter $\lambda = 2\mathrm{K}\ \mathrm{S(f)}/\varepsilon$ to $\mathrm{f_i}$ for M SNPs.
3. Pick the top K significant SNPs based on the noisy $\mathrm{f_i}$'s.

Algorithm 2: The ε-Differentially Private Algorithm for Releasing the K Most Significant SNPs Using the Exponential Mechanism.

Input: The number K of significant SNPs to release, the score function q (χ^2 statistics or Hamming distance score) for all M SNPs, the sensitivity of the score function S(q), privacy budget ε.

Output: K noisy significant SNPs.

1. Let $\mathcal{S} = \varnothing$ and q_i =the score of i-th SNP.
2. For each $i \in \{1, \ldots, M\}$, set the weight $w_i = \exp\left(\frac{\varepsilon q_i}{2KS(q)}\right)$ and $p_i = \frac{w_i}{\sum_{i=1}^{M} w_i}$ the probability for sampling i-th SNP.
3. Sample k elements independently from $\{1, \ldots, M\}$ with probabilities $\{p_1, \ldots, p_M\}$ without replacement. Add SNP $_k$ to $\mathcal{S}$ and set $q_k = -\infty$.
4. Repeat step 2, 3 until the size of $\mathcal{S}$ reaches K.

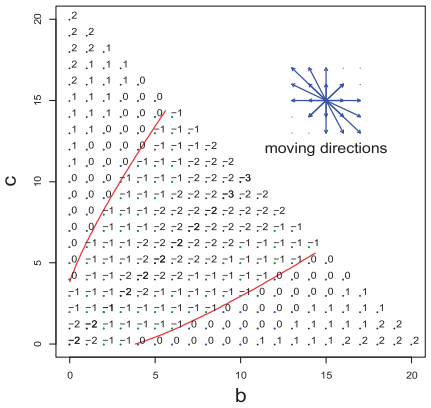

Fig. 3. The SHD scores for the points in D^1 from 10 families in one SNP. The red curves are from $T = c^*$, where $c^* = 95\% - \text{quantile of } \chi^2_1$. The points above the upper curve and below the lower curve are significant and the points between two curves and in boundaries are non-significant. The labels aside the points are their SHD scores from Algorithm 4

2.2 Differential privacy based on the shortest hamming distance score

Previous works (Johnson and Shmatikov, 2013; Simmons and Berger, 2016; Yu, 2010, 2014) show the DP mechanism based on the SHD score perform well in the case–control studies. We would like to develop and analyze this method to protect family's privacy in conducting TDT. The formal definition of SHD score in one SNP is in Definition 6. Johnson and Shmatikov (2013) showed that the sensitivity of the SHD score is 1. The SHD basically counts how many steps by moving the (b, c) pair among neighbor families such that the associated test statistic from being significant to being non-significant or vice versa, where the significance is defined by the test statistic is greater than or equal some predefined threshold $c^* > 0$. Since we are more concerned on the significant SNPs, we assign positive steps for the SNPs from being significant to being non-significant and negative steps for the SNPs from being non-significant to being significant. In this way, more significant SNPs obtain higher scores while non-significant SNPs get lower scores. In the exponential mechanism, the SNPs with higher scores have larger probabilities to be sampled.

Definition 6: (The SHD score (Johnson and Shmatikov, 2013)) Given a predefined threshold $c^* > 0$, the SHD score for i-th data $D_i := (b_i,\ c_i), i = 1,\ \ldots,\ M$, is

$$d_{SH}(D_i,i) = \begin{cases} 0, \text{ if } T_i \geq c^* \text{ and } \exists\ D_i', T_i' < c^* \\ 1 + \min_{D \sim D'} d_{SH}(D_i', i), \text{if } T_i \geq c^* \text{ and } \nexists\ D_i', T_i' < c^* \\ -1 + \max_{D \sim D'} d_{SH}(D_i', i), \text{ if } T_i < c^* \end{cases}$$

where T_i (T_i') is the test statistic associated with D_i (D_i') and $D_i \sim D_i'$ in $\mathcal{D}^M$. For $i \notin \{1, \ldots M\}$, setting $d_{SH}(D_i, i) = -\infty$.

To get the SHD score in case–control studies, early work (Johnson and Shmatikov, 2013; Yu, 2014) showed that it is a computational intensive task. Based on the minor allele frequencies using the iDASH healthcare privacy protection challenge data, Yu and Ji (2014) speeded up the computation by assuming that the control group is known. Later on, Simmons and Berger (2016) further improved the algorithm through a convex analysis and relaxed the constraint about knowing the control cohort.

The key to get the SHD score for TDT is to analyze how to move (b, c) pair among neighbor families affects the change of the test statistic $T = \frac{(b-c)^2}{b+c}$. Different from the allelic test statistic, the contour of T in TDT is not a straight line but a parabola, therefore, the change of T is no longer a constant. Hence, we have to take into account the starting point of (b, c) and the ending point along every move. Besides this, in TDT, there are 18 possible moving directions as shown in Figure 2, which makes the computation more challenging in our case. We develop two algorithms to calculate SHD score in private TDT: one is the exact algorithm (Algorithm 3) and the other is the approximation algorithm (Algorithm 4). In the exact algorithm, we take $(n_1, \ldots, n_6)$ the family numbers in each genotype (defined in Table 1) as the input. We construct a graph of nodes from all possible $(n_1, \ldots, n_6)$'s with fixed total family number and connected nodes from neighbor families. The brute force searching shortest distance in the graph gives the SHD score. We show that our exact algorithm gives exact SHD score (in Definition 6) in Theorem 4 with the proof in Appendix. The brute force searching (based on dynamic programming) is slow in the large cohort studies. To tackle its computation issue, we develop an approximate version of SHD for TDT, which may not have a guaranteed sensitivity but can be computed efficiently. In the approximation algorithm, we relax the constraint on the fixed total family number and only search the shortest distance path in the domain of (b, c). If initial T from the dataset is significant (non-significant), we move (b, c) in each step along the one of 18 possible moving direction in the largest (smallest) non-normalized directional gradient until the significance status of T is just altered. Figure 3 depicts the SHD scores from the approximation algorithm for 10 families in one SNP.

Theorem 4: Algorithm 3 outputs exactly the SHD score (in Definition 6).

3 Results

We are often interested in a few most significant SNPs. In most real datasets, there is a gap between the extremely significant SNPs and other SNPs in terms of P-value. This makes it possible to apply a DP mechanism to maintain both utility and privacy. Under the DP framework, we would like to apply the ε-differential privacy mechanisms we analyzed in the previous section to protect the privacy of releasing top K most significant SNPs in TDT. We quantify the

Algorithm 3: Exact Algorithm to calculate the SHD score in one SNP.

Input: Information regarding one SNP and the threshold c^* for T.
Output: The SHD score in one SNP.

1. Compute $(n_1, \ldots, n_6)$ defined in Table 1.
2. Construct a graph of nodes as $\{(n_1', \ldots, n_6') : \sum n_i = \sum n_i'\}$ where two nodes are connected if $\sum |n_i - n_i'| = 2$.
3. Compute $T = (n_1 + 2n_4 - n_2 - 2n_5)^2/(n_1 + n_2 + 2n_3 + 2n_4 + 2n_5)$ for all the nodes.
4. Find nodes which have a) $T > c^*$ and b) a neighbor with $T \leq c^*$. Put them in a set S.
5. If $(n_1, \ldots, n_6)$ has $T > c^*$, compute its shortest distance to S and take it as its score; otherwise compute the distance and then use 1–distance as the score. Dijkstra algorithm solves this.

Algorithm 4: Approximation Algorithm to calculate the SHD score in one SNP.

Input: Information regarding one SNP and the threshold c^* for T_i.
Output: The SHD score in one SNP.

1. Get (b_i, c_i) for T_i.
2. If $T_i < c^*$, move (b_i, c_i) to the direction among the valid moving directions in the domain that maximizes the non-normalized directional gradient until T_i is just above or equals c^* then $d_{SH}((b_i, c_i),\ i) = -(\#$ of steps); if $T_i \geq c^*$, move (b_i, c_i) to the direction among the valid moving directions in the domain that minimizing the non-normalized directional gradient until T_i is below c^* then $d_{SH}((b_i, c_i),\ i) = (\#$ of steps$) - 1$.

utility by the accuracy defined by the proportion of the reported top K significant SNPs that are correctly selected, i.e. $|\mathrm{A} \cap \mathrm{B}|/|\mathrm{A}|$ where A is the set of true top K significant SNPs and B is the set reported from a DP mechanism.

In a DP mechanism, there is always a tradeoff among the family size N in TDT, the privacy budget ε and the accuracy. Without the concern of privacy, the family number determines the power of the TDT; larger family number gives more statistically powerful of TDT thus returns more accurate significant SNPs. Privacy budget ε controls the level of privacy protection (from Definition 1); smaller ε provides stronger privacy protection. An ε-DP mechanism perturbs the TDT results by adding noise (e.g. in the Laplace mechanism or resampling the results in the exponential mechanism). Under a certain ε, applying DP mechanisms on a dataset of a large family size can lead to a more accurate result, otherwise on a dataset of a very small family size could make the result useless. The number of most significant SNPs to release also plays an important role in privacy and utility. Suppose the total number of SNPs we are concerned is $M = 10\,000$ and 10 SNPs are extremely significant. To release top $K \leq 10$, a well-designed DP mechanism can give an almost perfect result under a given privacy budget, but to choose a much larger K, perturbation could reduce the gap between extremely SNPs and others, and therefore making these DP mechanisms un-useful. We can see this from Algorithms 1–2. A large K leads larger variances (in the Laplace noise) in the Laplace mechanism and smaller weights on the significant SNPs in the exponential mechanism. Hence, given a dataset, under a fixed family size, to maintain both accuracy and privacy under the framework of DP, how powerful of the TDT test can achieve and how much privacy of the dataset can support (i.e. how many top significant SNPs to release) are essentially determined by the data. We would like to demonstrate these points and investigate the performance of our propose DP mechanisms in TDT under various settings in the simulation studies and give a guideline for real practices.

3.1 Simulation studies

In the simulation studies, we compare various DP mechanisms for TDT in a small cohort and a large cohort. We investigate how the parameters (N, M, K, ε) affect the accuracy in the simulated two cohorts, where N is the family number, M is the total SNPs number in the study, K is number of selected top significant SNPs to release, and ε is the privacy budget.

To generate the simulated cohorts, suppose there are true 10 significant SNPs associated with some disease of interest. In the small cohort, we set family number $N = 150$ and the SNPs number $M = 5000$ and in the large cohort, we set $N = 5000$, $M = 10^6$. To generate the test statistic for i-th SNP, we first generate $S_i := b_i + c_i$, the sum of the (b_i, c_i) pair, uniformly sampled from discrete integers from 0 to $2N$. Given S_i, we generate b_i from Binomial distribution with size S_i and probability 0.5 then set $c_i = S_i - b_i$. From (b_i, c_i)'s, the corresponding test statistics T_i's have approximately χ_1^2 distribution, which simulate the case that there is no linkage disequilibrium in those SNPs. To get the SNPs with linkage disequilibrium, we select the SNPs with top $K = 10$ S_i's and regenerate their (b_i, c_i) pairs, where b_i is from Binomial distribution with size S_i and probability 0.65, then $c_i = S_i - b_i$. In this way, these 10 SNPs are significantly greater than the others. To apply the exact algorithm in Algorithm 3, since its input is the family numbers in each genotype, given (b_i, c_i) for the i-th SNP, we generate a random combination of n's that satisfy the equations in section 1.1. We plot the top 100 largest test statistics from two cohorts in Figure 4. In our setting, the true significant SNPs are the top 10 largest test statistics, indicated by triangles in the figure. Comparing two cohorts in Figure 4, under the Bonferroni correction as $100(1 - 0.05/M)\%$-quantile of χ_1^2 distribution, even without adding any perturbation from a DP mechanism, from the small cohort, we can only select 8 true significant SNPs above the threshold, while in the large cohort, the true 10 significant SNPs are dramatically greater the non-significant SNPs and can be picked accurately. We can expect if we would like to ask for the top 10 SNPs from the small cohort, the released results under DP cannot have much accuracy, but from the large cohort, a well-designed DP mechanism can maintain both the privacy and accuracy.

We compare the performance of the DP mechanisms for TDT in terms of accuracy under a certain privacy budget ε and the number K of releasing most significant SNPs. Our proposed DP mechanisms for TDT include: the Laplace and exponential mechanisms based on test statistics, the exponential mechanisms based on p-values and projected p-values, and the exponential mechanism based on the shortest Hamming distance (SHD). We set the privacy budget ε from 0.1 to 3 and the releasing number $K = 1, 3, 5, 10$. We get the sensitivities for the mechanisms based on test statistics, P-values and projected P-values from Theorem 1-3, and the SHD score from Algorithm 3–4. To release top K most significant SNPs, we apply the Laplace mechanism from Algorithm 1 and the exponential

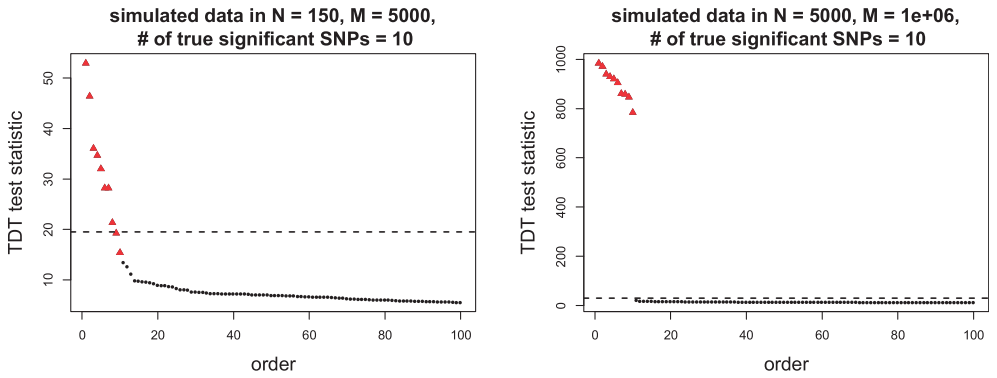

Fig. 4. Top 100 largest TDT test statistics in the simulated data. The left panel is from a small cohort (with family number N = 150 and SNPs number M = 5000) and the right panel from a large cohort (with N = 5000 and M = 10^6). The triangle points are the true top 10 significant SNPs. The horizontal dashed lines are the thresholds at 100(1–0.05/M)%-quantile of χ^2_1 distribution, where M = 5000, 10^6 in left and right panels respectively

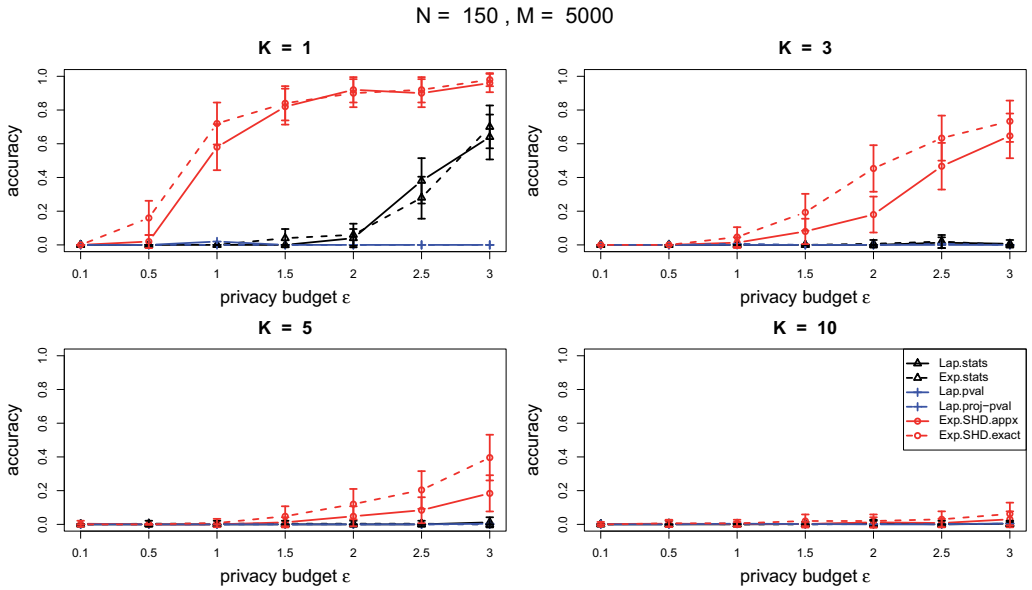

Fig. 5. Performance comparison of DP mechanisms for TDT in terms of accuracy under various privacy budgets and numbers of SNPs to release in a small cohort (with family size N = 150 and SNP number M = 5000). The solid and dashed curves with circle points are for the exponential mechanism based on SHD from the approximation algorithm and the exact algorithm. The solid and dashed curves with triangle points are for the Laplace and exponential mechanism based on the test statistics. The solid and dashed curves with square points are for the exponential mechanisms based *P*-value and projected *P*-value. The accuracy is reported in the average from repeatedly applying a DP mechanism 50 times. The small bars along the curves are the 95% confidence intervals

mechanism from Algorithm 2. In the small cohort, the exact algorithm takes about 4 hours and used up 38.5 GB memory on a 3.0 GHz 8-Core Intel Xeon E5 machine. (We are planning to develop a more efficient version for the exact algorithm.) Due to the computation complexity in the current version of the exact algorithm, we only apply the DP mechanism based on the SHD scores from the exact algorithm in the small cohort.

From the performance in the small cohort in Figure 5 and the large cohort in Figure 6, DP mechanisms based on *P*-values and projected *P*-values are almost useless due to their high sensitivities (see Theorem 2–3). The Laplace mechanism and the exponential mechanism based on the test statistics perform comparably. From both small and large cohort, the DP mechanism based on the SHD outperforms the others, and in the small cohort, the method from the exact algorithm performs slightly better than the one from the approximation algorithm. In the small cohort, since there is not a significant gap between the significant SNPs and non-significant ones from Figure 4, releasing top $K = 3, 5, 10$ SNPs, even under a large privacy budget, the performance of a DP mechanism is hard to have a great accuracy. When $K = 1$, the DP mechanism based on SHD achieves accuracy around 0.8 under $\varepsilon = 1.5$ while the DP mechanisms based on the test statistics requires $\varepsilon > 3$ for the same level of accuracy. In the large cohort, since there is a big gap between the significant SNPs and the others from Figure 4, we can ask for more top significant SNPs under a certain level of privacy protection and accuracy. The DP mechanism based on SHD achieves accuracy more than 0.8 under a small $\varepsilon = 0.5$ across all $K \leq 10$.

3.2 Real data

We apply our proposed DP mechanisms for TDT in the study of a rare disease—Kawasaki Disease (KD) (Shimizu, 2011). In this datast, we have 187 KD families. After filteation, we get 906 SNPs of ponential interest. We would like to report top *K* most significant SNPs under DP. From the comparison of applying the DP

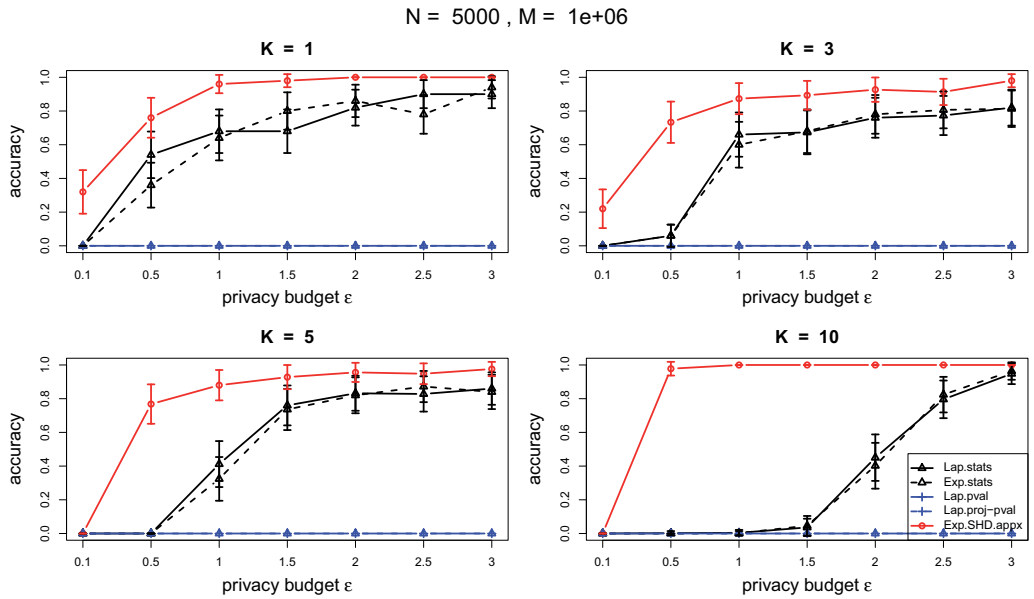

Fig. 6. Performance comparison of DP mechanisms for TDT in terms of accuracy under various privacy budgets and numbers of NSP to release in a large cohort with family size N = 5000 and SNP number M = 10^6. (The curve labels are the same as in Figure 5.)

mechanisms to TDT to the small and large cohorts in the simulation studies, we can expect the small size of this real dataset makes the problem of protecting familiy-based information challenging. We just want to show a real world application here.

Since the DP mechanism on the Shortest Hamming Distance (SHD) outperforms the other proposed mechanisms from simulation studies, we only apply this method to the real dataset. To get SHD scores, we use both exact and approixmation algorithms (Algorithm 3–4). The exact algorithm took about 4.2 hours to compute the SHD for this real dataset. Due to the small size of the data, it could be hard to ensure both high coverage accurage and privacy protection as shown in the simulations. We relax the cirterion a little bit. Instead of concerning whether the top K truly significant SNPs have been selected, we are interested in the average rank difference between the reported top K significant SNPs and their true ranks. We measure the utility by rank error, defined by $\frac{1}{K}\sum_{k=1}^{K}|R_k - k|$ where R_k is the rank of reported k-th significant SNP under DP. We set the privacy budget $\varepsilon = 3, 5$ and the number of significnat SNPs to release $K = 3, 5$. We set the threshold as 95%-quantile of chi-squared distribution with one degree of freedom for the SHD score. We report the average utilitites in Table 3 from 200 repeated procedures. From Table 3, the utilities from these two methods in calculating the SHD scores are similar. Comparing the utilities under $K = 3, 5$, to maintain a certain of privacy protection and utility, our small datset is more suitable for releasing very few most significant SNPs, this pattern was also observed from the simulation studies. From the rank error in Table 3, when $K = 1$ and $\varepsilon = 3$, the reported most significant SNP is among the truly top 10 most signifiant SNPs with high probability and when $K = 1$ and $\varepsilon = 5$, the reported SNP is either the true top one SNP or the top two for most of the time.

Imagine in practice, a researcher asks a query on this dataset and she would like to know a few most significant SNPs under DP TDT. She sets K, the number of most significant SNPs she would like to explore, and takes a privacy budget ε that bounds the error that she can tolerate. Since this is a small cohort, if she sets K too large and ε too small, we will suggest her recruiting more patients into the study

Table 3. Utility of the DP mechanism based on the SHD score calculated from exact and approximation algorithms for TDT in Kawasaki disease dataset

Method: Exp. Mech. SHD		Exact Alg.	Approximation Alg.
K	ε	Rank error	Rank error
$K = 1$	$\varepsilon = 3$	6.2	6.0
	$\varepsilon = 5$	0.8	0.8
$K = 3$	$\varepsilon = 3$	190.8	187.5
	$\varepsilon = 5$	65.9	73.3

to guarantee both privacy and utility of our reported results under DP. Or we can suggest her reducing the number of K in order to get a reasonable answer. Under a proper K and ε, for example, $K = 1$ and $\varepsilon = 3$, the user will get the top most significant SNP 'rs12569527' under DP with a high probability. Athough outcomes of such a test-run may not be directly used for scientific discovery, our point is that the user will get a sense how the dataset is likely to be useful through her exploratory analysis, and can therefore decide on whether it is necessary to gain access to those data through the Institutional Review Board (IRB) application. This is the key idea of our proposed framework, which is illustrated in Figure 1.

4 Discussion and conclusion

In this paper, we make the assumption that an attacker intends to violate family information. We apply the Laplace and the exponential mechanisms, two major ε-differentially private mechanisms, to achieve differentially private TDT in GWAS. However, there are still some limitations in the proposed method.

The proposed methods and the individual protected methods in Uhler (2013), Yu (2014) and Yu and Ji (2014) only focus on defining two neighbor datasets by *exchanging* one family/individual. Other ways of generating neighbor datasets can be used. To get the

SHD score, we developed both exact and approximation algorithms. The approximation algorithm is efficient, but the sensitivity of the SHD score is not guaranteed to be 1 any more. We are planning to develop a more efficient version of the exact algorithm.

The choice of privacy budget ε, as we discussed in the section of 'Results', depends on the trade-offs among family size, privacy and utility. Since there is no explicit formula to quantify the relationship among those three quantities, the setting of ε depends on the context and on the application. How to pick a proper ε is still an open question.

The proposed method protects privacy for trio-families. We can further develop privacy-preserving methods to the extended TDT on the families with more than one child and testing more alleles in a marker locus (Spielman, 1993) in the same manner. However, more discussion on the changes of two neighbor datasets should be considered. Furthermore, TDT is based on a χ^2 test statistic and it has good asymptotic properties when the counts b and c are large. However, when the counts are small, TDT may lose some detection ability. In this case, other tests such as the Binomial exact test and continuity correction may be considered. Recently there are modified TDTs (Ewens and Spielman, 2004; Wittkowski and Liu, 2002) that can be more powerful under some circumstances. Hence, we can choose different TDT tests and then apply our privacy-preserving mechanisms to obtain higher utility. Additionally, in practice, real data always involve missing data and there are several imputation techniques to get more information. How to properly use imputation techniques in a privacy-preserving manner still needs further investigation.

Our analysis does not consider the correlation structure of SNPs but note that a DP mechanism does not depend on the correlation among SNPs to guarantee DP. However, to better use the dataset, if the dataset involves highly correlated SNPs, one can apply the DP TDT to the tag SNPs, the representatives in each correlated block (Gabriel *et al.*, 2002); otherwise it might not be interesting that the reported top K most significant SNPs come from the same correlated block. Similar to (Johnson and Shmatikov, 2013), we can extend our DP mechanisms to TDT on the correlated SNPs.

Overall, from our simulation experiments and application in the Kawasaki Disease dataset, by applying TDT to trio-families in GWAS and defining the neighbor datasets by exchanging family genotypes, we found that the exponential mechanism based on the shortest Hamming distance score preserves both higher utility and privacy than the other proposed methods to protect family-based information.

Contribution statement

Analysis in the methods: M.W.; Algorithms: M.W., Z.J., S.W, X.J.; Data Processing and software: M.W., J.K., H.Y.; Editing and revising the manuscript: M.W., L.OM., X.J.

Acknowledgements

We would like to thank the A.E. Bonnie Berger and the two anonymous reviewers for their editing work and helpful suggestion. We thank Jane C. Burns MD for contributing the Kawasaki Disease trio genotype data for this project. MW would like to thank the training from UCSD.

Funding

This work was supported by the Patient-Centered Outcomes Research Institute (PCORI) under contract ME-1310-07058, the National Institute of Health (NIH) under award numbers R00HG008175, R01HG008802, R01HG007078, R01GM114612, R01GM118574, R01GM118609, U01EB023685, U54HL108460 and National Library of Medicine (NLM) R00LM011392, R21LM012060.

Conflict of Interest: none declared.

References

(onlince source) 'HiSeq™ Sequencing Systems' from Specification Sheet in Illumina® Sequencing, in the link https://www.illumina.com/systems/sequencing-platforms/hiseq-x.html

Bhaskar,R. *et al.* (2010) Discovering frequent patterns in sensitive data. In: *Proceedings of the 16th ACM SIGKDD International Conference on KnowledgeDiscovery and Data Mining – KDD '10*, **2010**, p. 503.

Chen,F. *et al.* (2017) PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. *Bioinformatics*, btw758.

Church,G. (2005) The Personal Genome Project. *Mol. Syst. Biol.*, **1**, no. 1.

Clarke,L. *et al.* (2012) The 1000 Genomes Project: data management and community access. *Nat. Methods*, **9**, 459–462.

Clayton,D. (2010) On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics*, **11**, 661–673.

Collins,F. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.

Craig,D. *et al.* (2011) Assessing and managing risk when sharing aggregate genetic variant data. *Nat. Rev. Genet.*, **12**, 730–736.

Dwork,C. (2006) Differential privacy. *Int. Colloq. Autom. Lang. Program*, **4052**, 1–12.

Dwork,C. *et al.* (2006) Calibrating noise to sensitivity in private data analysis. *Theory Cryptogr*, **3876**, 265–284.

Ewens,W. and Spielman,R. (2004) The TDT is a statistically valid test: comments on Wittkowski and Liu. *Hum. Hered.*, **58**, 59–60.

Gabriel,S. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

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

Gutmann,A. W. *et al.* (2012) Privacy and progress in whole genome sequencing. In: *Presidential Committee for the Study of Bioethical* 2012.

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.

Jacobs,K. *et al.* (2009) A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat. Genet.*, **41**, 1253–1257.

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 - KDD '13*, p. 1079.

Levin-Decanini,T. *et al.* (2013) Parental broader autism subphenotypes in {ASD} affected families: relationship to gender, child's symptoms, {SSRI} treatment, and platelet serotonin. *Autism Res.*, **6**, 621–630.

Lin,Z. *et al.* (2004) Genomic research and human subject privacy. *Science*, **305**, 183.

Malin,B. and Sweeney,L. (2001) Inferring genotype from clinical phenotype through a knowledge based algorithm. In: *Proceedings of the Pacific Symposium on Biocomputing*, pp. 41–52.

Malin,B. and Sweeney,L. (2004) How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J. Biomed. Inform.*, **37**, 179–192.

McSherry,F. and Talwar,K. (2007) Mechanism Design via Differential Privacy. In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103.

Ott,J. *et al.* (2011) Family-based designs for genome-wide association studies. *Nat. Rev. Genet.*, **12**, 465–474.

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

Shimizu,C. *et al*. (2011) Transforming growth factor-β signaling pathway in patients with Kawasaki disease. *Circ. Cardiovasc. Genet*., **4**, 16–25.

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

Simmons,S. and Berger,B. (2016) Realizing privacy preserving genome-wide association studies. *Bioinformatics*, **32**, 1293–1300.

Spielman,R. *et al*. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet*., **52**, 506.

Sweeney,L. *et al*. (2013) Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment). *Computers and Society*, arXiv.

Uhler,C. *et al*. (2013) Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confidentiality*, **5**, 137–166.

Visscher,P. and Hill,W. (2009) The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet*., **5**, e1000628.

Wang,R. *et al*. (2009) Learning your identity and disease from research papers. In: *Proceedings of the 16th ACM conference on Computer and communications security – CCS '09, 2009*, pp. 534–44.

Wang,S. *et al*. (2014) Differentially private genome data dissemination through top-down specialization. *BMC Med. Inform. Decis. Mak*., **14**, S2.

Wang,S. *et al*. (2016) HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS. *Bioinformatics*, **32**, 211–218.

Wittkowski,K. and Liu,X. (2002) A statistically valid alternative to the TDT. *Hum. Hered*., **54**, 157–164.

Yang,Z. *et al*. (2014) Whole-exome sequencing for the identification of susceptibility genes of Kashin-Beck disease. *PLoS One*, **9**, e92298.

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

Yu,F. *et al*. (2010) Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In: Domingo-Ferrer (J.ed.) *Privacy in Statistical Databases*. **vol. 8744**, Springer International Publishing, Cham, pp. 170–184.

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

Zhang,Y. *et al*. (2015) FORESEE: Fully Outsourced secuRe gEnome Study basEd on homomorphic Encryption. *BMC Med Inf. Decis Mak*., **15**, S5.