Efficient Differentially Private Methods for a Transmission Disequilibrium Test in Genome Wide Association Studies

Akito Yamamoto[†] and Tetsuo Shibuya

Division of Medical Data Informatics, Human Genome Center,
The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan
[†]*E-mail: a-ymmt@ims.u-tokyo.ac.jp*

To achieve the provision of personalized medicine, it is very important to investigate the relationship between diseases and human genomes. For this purpose, large-scale genetic studies such as genome-wide association studies are often conducted, but there is a risk of identifying individuals if the statistics are released as they are. In this study, we propose new efficient differentially private methods for a transmission disequilibrium test, which is a family-based association test. Existing methods are computationally intensive and take a long time even for a small cohort. Moreover, for approximation methods, sensitivity of the obtained values is not guaranteed. We present an exact algorithm with a time complexity of $\mathcal{O}(nm)$ for a dataset containing n families and m single nucleotide polymorphisms (SNPs). We also propose an approximation algorithm that is faster than the exact one and prove that the obtained scores' sensitivity is 1. From our experimental results, we demonstrate that our exact algorithm is $10,000$ times faster than existing methods for a small cohort with $5,000$ SNPs. The results also indicate that the proposed method is the first in the world that can be applied to a large cohort, such as those with 10^6 SNPs. In addition, we examine a suitable dataset to apply our approximation algorithm. Supplementary materials are available at https://github.com/ay0408/DP-trio-TDT.

Keywords: Differential Privacy; GWAS; TDT.

1. Introduction

In recent years, the amount of human genome data has increased dramatically with the advances in genome technologies. Based on these data, it is important to analyze the relationship between genomes and diseases for genome research and personalized medicine. Genome wide association studies (GWAS) represent a type of statistical analysis to investigate genetic factors of diseases such as cancer. A typical GWAS statistically analyzes the links between millions of single nucleotide polymorphism (SNP) locations and diseases, and the analysis methods include case-control studies with contingency tables[2,14] and family-based transmission disequilibrium tests (TDTs).[1,17]

The use of statistics obtained from these tests is essential for the development of medicine, but it also poses privacy issues. If these statistics are made public as they are, genomic information of an individual might be leaked, and several studies[7,13] have been conducted on the identification of an individual using genomic information. In addition, the risks of being identified only from statistics using genomic information have been shown,[23] and some attack methods against GWAS have been proposed.[9,16] These studies resulted in the NIH ceasing to release aggregate GWAS data,[26] and now it is difficult to use these data freely.

In this situation, it is important to develop methods to release statistics based on genomic data, including GWAS data, while preventing the identification of individuals and maintaining the statistics' utility. For this purpose, we focused on the concept of differential privacy.[4] Differential privacy is a framework to protect the privacy of individuals in a database when releasing useful information such as genomic statistics. By adding perturbation to the original information, it creates a situation wherein it is almost impossible to distinguish whether a database contains a particular individual, regardless of what information the adversary has.

Using this concept, Fienberg et al.[6] proposed methods to release minor allele frequencies, chi-squared statistics, and p-values based on 3×2 contingency tables in GWAS. Since then, several studies[10,24,25] have proposed practical differential privacy mechanisms for case-control studies in GWAS. Other analyses in GWAS include family-based correlation analysis, and Wang et al.[22] proposed differentially private mechanisms for TDTs in the case of trio families, that is, one affected child per family.[19] However, their exact algorithm requires solving the shortest path problems with a large number of nodes, which are computationally intensive, and takes a long time to run even for a relatively small dataset. Furthermore, sensitivity of the score function obtained by their approximation algorithm is not guaranteed. Since the sensitivity affects the level of privacy that can be achieved in the concept of differential privacy, their algorithm is not strictly privacy-preserving.

In this study, we propose efficient methods to release the top K significant SNPs based on the TDT statistics in GWAS with the concept of differential privacy. We focus on the exponential mechanism, which has been shown to provide highly accurate results in various methods for releasing statistics based on contingency tables,[10,18,25] and we adopt the shortest Hamming distance (SHD) score as the score function. We present exact and approximation algorithms for calculating the SHD score, and the computational complexity is $\mathcal{O}(nm)$ and $\mathcal{O}(m)$ for a dataset with n families and m SNPs, respectively. For the approximation algorithm, we also prove that the sensitivity of the resulting SHD score is 1. This makes it possible to apply the exponential mechanism under differential privacy. Subsequently, we evaluate the run time and accuracy of our methods through experiments and show that our exact method is $10,000$ times faster than Wang et al.'s method[22] for a small cohort with $5,000$ SNPs, and is the first globally to be applied to a large cohort with 10^6 SNPs. We also show that our approximation algorithm is much faster than the exact one, taking only about 4 seconds to complete the calculation even for the large cohort.

In Section 2, we describe basic assumptions and preliminary definitions. In Section 3, we present ϵ-differentially private methods for releasing the top K significant SNPs. In Section 4, we evaluate their utility based on simulation and real data. We summarize our study with directions for future work in Section 5. In the supplementary material, we provide detailed proofs for our algorithms, describe the datasets used in our experiments, and also show a method for releasing TDT statistics in combination with the Laplace mechanism.[3]

2. Preliminaries

A typical GWAS examines whether there is an association between marker loci, such as SNPs, and diseases. Test methods used in such studies include affected family-based control studies.[5,20,21] These methods can test whether there is a correlation between a marker locus and a disease that has a genetic linkage. In addition, TDT can also test for linkage when there is a correlation.

2.1. *TDT*

TDT[19] is a test for linkage disequilibrium, which examines the relationship between a disease and two or more alleles, depending on how many children are in a family. In TDT for n trio families, we consider $2n$ parents and n affected children. We focus on the case of testing for two alleles, such as SNPs. When the two alleles are M_1 and M_2, the $2n$ parents can be classified according to the type of allele transmitted to their child as shown in Table 1.

Table 1. Number of parents for TDT in one SNP.

		Non-Transmitted Allele		Total
		M_1	M_2	
Transmitted Allele	M_1	a	b	$a+b$
	M_2	c	d	$c+d$
Total		$a+c$	$b+d$	$2n$

Under the null hypothesis of no linkage or no correlation between a marker locus and a disease, the TDT statistics are expressed as follows:

$$\chi^2_{td} := \chi^2_{td}(b,c) = \frac{(b-c)^2}{b+c}.$$

These statistics approximately follow a chi-squared distribution with one degree of freedom. Since $b = c$ under the null hypothesis, when $b = c = 0$, we define $\chi^2_{td} = 0/0 = 0$. The possible combinations of (b,c) in one family are shown in Table 2, and b and c in n families can be calculated by the following equations: $b = n_1 + n_3 + 2\,n_4$ and $c = n_2 + n_3 + 2\,n_5$.

Table 2. Number of families for each (b,c).

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

2.2. *Differential Privacy*

Differential privacy[4] is a concept developed in the field of cryptography as a framework that allows statistical analysis of databases while preserving personal data in the database from adversaries. The idea of differential privacy is based on the fact that it should be almost impossible to distinguish between two neighboring datasets, that is, they differ in just one record. In this study, we define two neighboring datasets by exchanging the genomic data of exactly one family. The privacy level in differential privacy is evaluated by the parameter $\epsilon > 0$. The following is the definition of ϵ-differential privacy.

Definition 1. (ϵ*-Differential Privacy*)
A randomized mechanism M is ϵ-differentially private if, for all datasets D and D', which differ in only one family and any $S \subset$range(M),

$$\Pr[M(D) \in S] \leq e^{\epsilon} \cdot \Pr[M(D') \in S].$$

The closer ϵ is to zero, the more privacy is preserved, and the larger ϵ is, the less privacy is guaranteed. On the other hand, the higher is the privacy level, the lower is the data's utility, so the value of ϵ needs to be set with consideration of the trade-off between privacy and utility. In general, the value of ϵ is set in the range from 0.01 to 10,[8] but a smaller value should be chosen when more privacy is considered, such as the case with genomic data.

One main mechanism that satisfies the definition of ϵ-differential privacy is the exponential mechanism.[15] The exponential mechanism uses a score function, which indicates the desirability of the original output. Based on the *sensitivity* of the score function, elements with higher scores are made to have higher probability of being released. *Sensitivity* is defined as follows.

Definition 2. (*Sensitivity for the Exponential Mechanism*)
Let $\mathcal{D}^M$ be the collection of all datasets with M SNPs; then, the sensitivity of a score function $u : \mathcal{D}^M \times \{1, 2, \ldots, M\} \to \mathbb{R}$ is

$$\Delta u = \max_{r} \max_{D,D'} |u(D, r) - u(D', r)|,$$

where $r \in \{1, 2, \ldots, M\}$ and $D, D' \in \mathcal{D}^M$ differ in a single family.

Following the above definition, we choose mechanism $\mathcal{M}_u^{\epsilon}$, which has the distribution:

$$\mathcal{M}_u^{\epsilon} = \frac{\exp\left(\frac{\epsilon u(D,r)}{2\Delta u}\right)}{\sum_{s\in\{1,\ldots,M\}} \exp\left(\frac{\epsilon u(D,s)}{2\Delta u}\right)}.$$

Then, releasing $\mathcal{M}_u^{\epsilon}$ satisfies the definition of ϵ-differential privacy.

In this study, we use the SHD score as the score function, which performed better than the chi-squared statistics and p-values in existing studies.[22,25] In the allelic test, various mechanisms using the SHD score have been proposed,[10,18,25] and it has been shown that this score's sensitivity is 1.[10] The SHD score indicates from how many neighboring datasets the statistics should be traced, from significant to non-significant and vice versa, and the definition of the SHD score is as follows.

Definition 3. (*The SHD score*)
Given a predefined threshold $c^* > 0$, the SHD score for i-th data D_i $(i = 1, 2, \ldots, M)$ is

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

where T_i and T_i' are the test statistics obtained from D_i and D_i', respectively, and $D_i, D_i' \in \mathcal{D}^M$ differ in a single family. For $i \notin \{1, \ldots, M\}$, $d_{\text{SH}}(D_i, i) = -\infty$.

3. Methods

In this study, we aim to release the K most significant SNPs. We first show an efficient exact algorithm to obtain the SHD score. Then, we propose an algorithm for calculating the approximation SHD score whose sensitivity is 1.

3.1. *Exact Algorithm*

Some differentially private releasing methods for trio families have been proposed by Wang et al..[22] However, the exponential mechanism in their methods, which gave relatively accurate results, has high time complexity and takes too much running time. In fact, it took 4.2 hours for a dataset with 187 families and 906 SNPs.[22] The reason for this is that it requires constructing a graph with $\mathcal{O}(n^5)$ nodes for a dataset with n families and solving the shortest path problems for m SNPs. To deal with this concern, we propose an efficient and rigorous exponential mechanism with the complexity of $\mathcal{O}(nm)$ for the same dataset, which does not need to consider graphs.

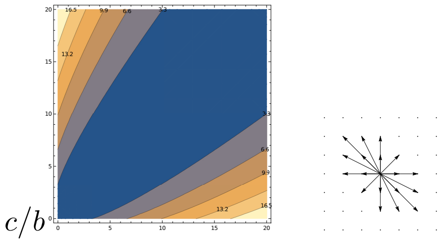

Fig. 1. Contour plots of the transmission disequilibrium test statistic for trio families and the possible moves of (b, c).

In this study, we adopt the SHD score as the score function in the exponential mechanism. Here, the TDT statistic for trio families is $\frac{(b-c)^2}{b+c}$, and the contours of this function can be shown as in Fig. 1. There can be 18 moving directions of (b, c) between two neighboring datasets, as shown in the figure. These 18 moves are due to the variations in the combinations of (b, c) shown in Table 2, and our algorithm calculates the SHD score by changing the values of n_k $(k = 1, 2, \ldots, 6)$. The detailed procedure is shown in Algorithm 1, and we prove that this algorithm gives the exact SHD score in Theorem 1.

Theorem 1. *Algorithm 1 outputs the exact SHD score.*

Proof. We consider two cases: $T < c^*$ and $T \geq c^*$. For each case, we look at how to change the values of n_1 to n_6 in Table 2 to maximize the change in TDT statistics. For a detailed proof, see Supplementary Theorem S1 in Supplementary Section S2.1. □

In Algorithm 1, the number of families to be changed is at most $2n$, so the computational complexity of this algorithm is $\mathcal{O}(n)$. If a dataset has m SNPs, we only need to apply Algorithm 1 m times, and thus the total complexity is $\mathcal{O}(nm)$. The ϵ-differentially private mechanism using the SHD scores for releasing the top K significant SNPs is represented in Algorithm 2. Here, as K increases, the accuracy of the output is expected to decrease because the weights of significant SNPs become smaller.

Algorithm 1 Exact algorithm to find the SHD score for TDT statistics.

Input: Information about a single SNP, that is, n_1, n_2, n_3, n_4, n_5, n_6, and the threshold c^* for the TDT statistics.

Output: The SHD score in one SNP.

```
$T = (n_1 - n_2 + 2n_4 - 2n_5)^2/(n_1 + n_2 + 2n_3 + 2n_4 + 2n_5)$

if $T < c^*$ then
    Increase the number of families with $(b, c) = (2, 0)$.
    $d_1 = 0$, $N_k = n_k (k = 1, \ldots, 6)$
    while $T < c^*$ do
        Check the value of $N_5$, $N_2$, $N_3$, $N_6$, and $N_1$ in that order, and if a value greater than 0 is found, decrease it by one and continue to the next step.
        $N_4 \leftarrow N_4 + 1$
        $T = (N_1 - N_2 + 2N_4 - 2N_5)^2/(N_1 + N_2 + 2N_3 + 2N_4 + 2N_5)$
        $d_1 \leftarrow d_1 - 1$
    end while

    Increase the number of families with $(b, c) = (0, 2)$.
    $d_2 = 0$, $N_k = n_k (k = 1, \ldots, 6)$
    As in the above case, check $N_4$, $N_1$, $N_3$, $N_6$, and $N_2$ in that order, and increase $N_5$, then decrease $d_2$ until $T \geq c^*$.

    The SHD score is $\max\{d_1, d_2\}$.

else if $T \geq c^*$ then
    if $n_1 + 2n_4 > n_2 + 2n_5$ then
        As in the case of $T < c^*$, check $n_4$, $n_1$, $n_6$, $n_3$, and $n_2$ in that order, and increase $n_5$ until $T < c^*$.
    else
        Check $n_5$, $n_2$, $n_6$, $n_3$, and $n_1$ in that order, and increase $n_4$ until $T < c^*$.
    end if
    The SHD score is (the number of steps) $-1$.
end if
```

3.2. *Approximation Algorithm*

We also propose an algorithm to find the approximation SHD score whose sensitivity is 1. The computational complexity of our algorithm is $\mathcal{O}(1)$ when finding the SHD score for a single SNP, which is much faster than the exact algorithm. We also prove that sensitivity of the obtained score is 1, which has not been shown in the existing approximation algorithm.[22]

In our approximation algorithm, we focus on only variables b and c in calculating the TDT statistics($= (b - c)^2/(b + c)$). First, we consider the case wherein the original data are

Algorithm 2 ϵ-differentially private algorithm for releasing the top K significant SNPs using the exponential mechanism with the SHD score.

Input: The SHD score of all m SNPs, number K of SNPs to release, and privacy budget ϵ.
Output: Top K significant SNPs.

```
1: Let $S = \emptyset$ and $q_i$ be the SHD score of the $i$-th SNP.
2: For each $i \in \{1, \dots, m\}$, set the weight $w_i = \exp\left(\frac{\epsilon q_i}{2K}\right)$ and the probability $p_i = \frac{w_i}{\sum_{i=1}^{m} w_i}$ for sampling the $i$-th SNP.
3: Sample $k$ from $\{1, \dots, m\}$ with probabilities $\{p_1, \dots, p_m\}$; add $k$-th SNP to $S$ and set $q_k = -\infty$.
4: Repeat steps 2 and 3 until the size of $S$ reaches $K$.
```

not significant. If $b + c < c^*$, we start by increasing $b + c$ to c^*. Then, we increase $|b - c|$ to c^*. Since the maximum changes in the sum and difference of b and c are 2 and 4, respectively, the approximation score can be calculated by $-\left\lceil \frac{c^*-(b+c)}{2} + \frac{c^*-\{|b-c|+c^*-(b+c)\}}{4} \right\rceil = -\left\lceil \frac{2c^*-(b+c)-|b-c|}{4} \right\rceil$. If $b + c \geq c^*$, we increase the difference between b and c to $\sqrt{(b+c) \cdot c^*}$. When the original data are significant, we reduce it to $\sqrt{(b+c) \cdot c^*}$. The above procedures are summarized in Algorithm 3. We show that the sensitivity of the score in this way is 1 by Theorem 2. For a proof of this theorem, please see Supplementary Section S2.2.

Theorem 2. *Sensitivity of the SHD score obtained by Algorithm 3 is 1.*

Algorithm 3 Approximation algorithm to find the SHD Score for TDT statistics.

Input: Information about a single SNP, that is, n_1, n_2, n_3, n_4, n_5, n_6, and the threshold c^* for the TDT statistics.
Output: The SHD score in one SNP.

```
$b = n_1 + n_3 + 2n_4$, $c = n_2 + n_3 + 2n_5$
$T = (b - c)^2/(b + c)$
if $T < c^*$ then
    if $b + c < c^*$ then
        The SHD score is $-\left\lceil \frac{2c^*-(b+c)-|b-c|}{4} \right\rceil$.
    else if $b + c \geq c^*$ then
        The SHD score is $-\left\lceil \frac{\sqrt{(b+c)\cdot c^*}-|b-c|}{4} \right\rceil$.
    end if
else if $T \geq c^*$ then
    The SHD score is $\left\lceil \frac{|b-c|-\sqrt{(b+c)\cdot c^*}}{4} \right\rceil - 1$.
end if
```

Even when using this approximation score, Algorithm 2 can be used to release the top K significant SNPs privately.

4. Experiments

We first measured the run time of our algorithms in a small cohort and a large cohort using two types of simulation data: one wherein families were only in n_1, n_2, and n_6 categories, and one wherein families were distributed across n_1 to n_6. Then, we calculated the accuracy rate of the top K significant SNPs in each case and examined the effect of the value of K and ϵ. The threshold c^* for calculating the SHD score was same as the threshold for statistical tests.

4.1. *Simulation Data*

For both cases in (I) small cohort and (II) large cohort, we generated simulation data for two situations: (i) all families were in the n_1 or n_2 or n_6 categories, and (ii) families were distributed in n_1 to n_6 categories. The distributions of the statistics in the datasets generated by the following procedures are shown in Supplementary Section S2.

(I) Small Cohort

We set the family number $N = 150$ and SNP number $M = 5,000$ as in the experiments by Wang et al.[22] Here, we consider generating a dataset for the i-th SNP.

(i) First, we let S_i be a random natural number in the range of 0 to $2N$. Then, we generate n_1 from binomial distribution with size S_i and probability 0.5. Finally, we set $n_2 = S_i - n_1$ and $n_6 = 2N - n_1 - n_2$. In addition, for the 10 SNPs, the probability in the binomial distribution to generate n_1 is set to 0.75 to create some significant datasets.

(ii) We set n_1 to n_6 by the following equations:

$$n_1 = Binomial\left(2N,\ \frac{1}{6}\right), \qquad n_2 = Binomial\left(2N - n_1,\ \frac{1}{5}\right),$$

$$n_3 = Binomial\left(2N - n_1 - n_2,\ \frac{1}{4}\right), \qquad n_4 = Binomial\left(2N - n_1 - n_2 - n_3,\ \frac{1}{3}\right),$$

$$n_5 = Binomial\left(2N - n_1 - n_2 - n_3 - n_4,\ \frac{1}{2}\right), n_6 = 2N - n_1 - n_2 - n_3 - n_4 - n_5.$$

For the generation of 10 significant datasets, the probabilities in the binomial distribution are set to $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{3}$, in that order.

(II) Large Cohort

We set $N = 5,000$ and $M = 10^6$ as in the experiments by Wang et al.[22] The way to generate non-significant datasets is the same as in (I). When generating 10 significant datasets,

(i) the probability in the binomial distribution to calculate n_1 is set to 0.55, and

(ii) the probabilities are set to $\frac{11}{60}$, $\frac{2}{11}$, $\frac{1}{4}$, $\frac{11}{30}$, and $\frac{5}{11}$, in that order.

4.2. *Results*

4.2.1. *Run Time*

We measured the run time of calculating the SHD score based on the generated data described above. We conducted five runs for each case, and the averages are shown in Table 3.

Table 3. Run Time of our algorithms for a (I) small cohort and (II) large cohort, when the (i) distribution of families is unbalanced and (ii) families are distributed across all categories.

(I)	exact	appx.
(i)	0.875 s	0.020 s
(ii)	0.972 s	0.019 s

(II)	exact	appx.
(i)	1081.773 s	4.047 s
(ii)	1338.327 s	4.163 s

The existing method by Wang et al.[22] took four hours even for a small cohort, but our algorithm is 10,000 times faster than that. For a large cohort, our exact algorithm can compute within about 20 minutes, indicating that it is practical. To the best of our knowledge, this is the first high-speed algorithm for a large cohort.

4.2.2. *Accuracy*

We varied the values of K and ϵ and calculated the accuracy for top K significant SNPs' output by the exponential mechanism. For the four cases described in Subsection 4.1, the accuracies of the exact and approximation algorithms are plotted in Figs. 2, 3, 4, and 5.

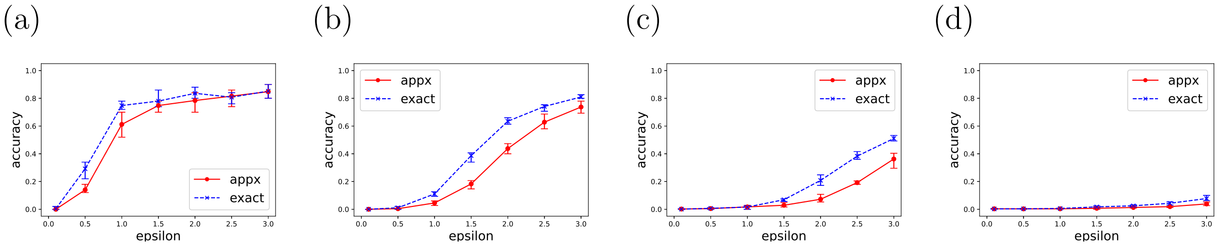

Fig. 2. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (I)(i).

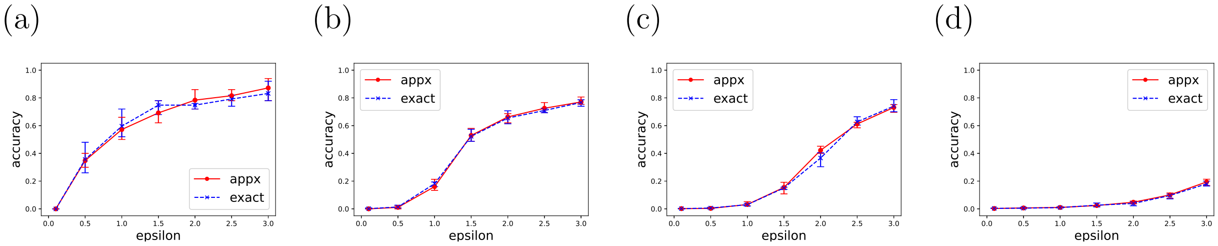

Fig. 3. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (I)(ii).

First, we discuss the case of a small cohort. Figs. 2 and 3 indicate that when $K = 1$, high accuracy could be obtained even if we set ϵ as small as 1.0 to 1.5. On the other hand, when $K = 10$, no high accuracy was obtained. For practical use, it might be better to set K as 3 or 5, and the value of ϵ as 1.5 to 2.5. Moreover, interestingly, in case (ii), wherein families are distributed across all categories, the approximation algorithm achieved almost the same accuracy as the exact algorithm. One possible reason for this is that there is a reasonable number of families included in the n_4 and n_5 categories.

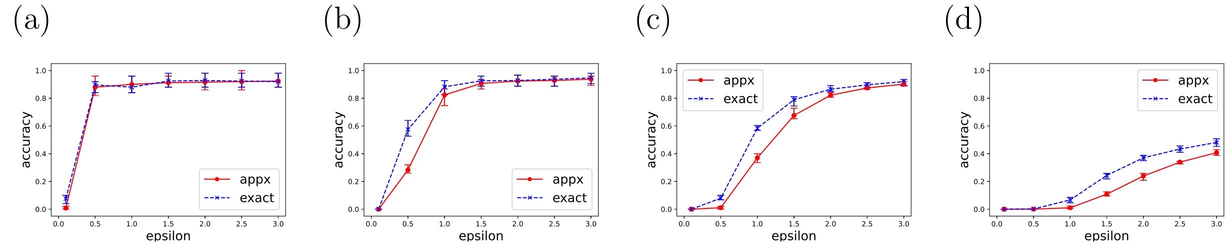

Fig. 4. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (II)(i).

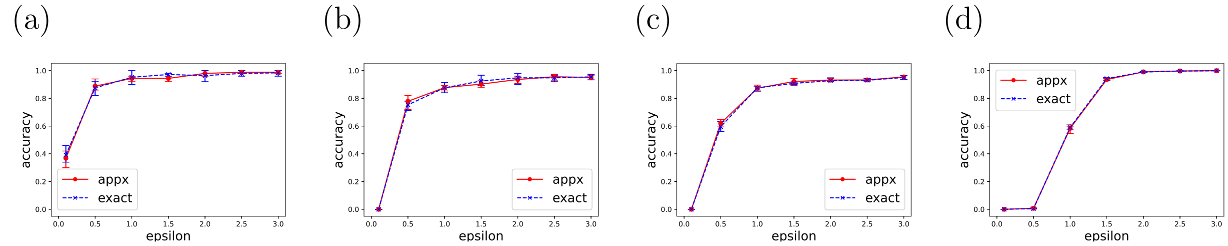

Fig. 5. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (II)(ii).

For the case of a large cohort, the figures' outline is roughly the same as that for a small cohort, but with high accuracy. In fact, it is expected that the statistics on significant SNPs will be larger for a large cohort than for a small cohort, and therefore, the top SNPs will be selected with higher probability by the exponential mechanism.

4.3. *Real Data*

We also evaluated our algorithms based on real data. The statistical data we used are TDT statistics from Justice et al..[11] This data are for 215 families and $649,669$ SNPs. We generated family datasets based on the statistics and applied our algorithms to them, and we confirmed that our algorithms indeed work for a large cohort. The detailed generating procedure are described in Supplementary Section S3.2. For these datasets, we varied the values of K and ϵ and calculated the rate at which significant SNPs were output. For each case, 30 repeated procedures were carried out, and the average results are presented in Table 4. The average run time was 45 seconds for the exact algorithm and 2.2 seconds for the approximation one.

Table 4. Probability of extracting significant SNPs.

K		1				2			3		5
ϵ		2	3	5	7	3	5	7	5	7	7
Significance	Exact	73.3	90.0	83.3	90.0	63.3	75.0	85.0	61.1	74.4	43.3
Rate (%)	Appx.	73.3	80.0	90.0	96.7	60.0	75.0	86.7	56.7	72.2	39.3

These results show that when $K = 1$, a significant SNP can be released with high probability even if the value of ϵ is as small as 2. Even when $K = 5$, if the value of ϵ is 7, significant SNPs can be detected with a probability of 40–50%. In the case of $K = 2$ or 3, it was implied that our algorithms would be reasonably useful with ϵ values of 3 or 5.

5. Conclusion

In this study, we presented efficient privacy-preserving methods for releasing significant SNPs based on TDT statistics in GWAS. Our exact algorithm is about $10,000$ times faster than the previous method[22] for small cohorts. Our experimental results indicated that our algorithms are the first in the world to be practical even for large cohorts, such as those with 10^6 SNPs. We have also shown that sensitivity of the SHD score obtained by our approximation algorithm is 1. Our simulation studies have suggested that the approximation algorithm can be as accurate as the exact algorithm when there is no imbalance in the combination of genotypes in a family dataset. If we want to release the top K TDT statistics privately, we could consider adopting the Laplace mechanism.[3] The detailed algorithm is shown in Supplementary Section S4.

Limitations of this study include the restricted situations in which the approximation algorithm can achieve high accuracy and the probability of extracting non-significant SNPs.

For future research, we need to consider multi-allelic TDT[12] or the case wherein one family has two or more affected children, not only the case of trio families. Also, it might be desirable to investigate score functions other than the SHD score for the exponential mechanism.

Acknowledgement

This work was supported by JSPS KAKENHI Grant 20H05967, 20K21827, 21H05052. and JST CREST Grant JPMJCR1402JST.

References

1. B. Benyamin, P. M. Visscher, and A. F. McRae. Family-based genome-wide association studies. *Pharmacogenomics.*, 10(2):181–190, 2009.
2. T. Dickhaus, K. Straßburger, D. Schunk, C. Morcillo-Suarez, T. Illig, and A. Navarro. How to analyze many contingency tables simultaneously in genetic association studies. *Stat. Appl. Genet. Mol Biol.*, 11(4):1544–6115.1776, 2012.
3. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *S. Halevi and T. Rabin, (eds) Theory of Cryptography*, 3876:265–284, 2006.
4. Cynthia Dwork. Differential privacy. *M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, (eds) Automata, Languages and Programming*, 4052, 2006.
5. C. T. Falk and P. Rubinstein. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.*, 51(3):227–233, 1987.
6. S. E. Fienberg, A. Slavkovic, and C. Uhler. Privacy preserving GWAS data sharing. In *IEEE 11th International Conference on Data Mining Workshops*, pages 628–635, Vancouver, Canada, December 2011.
7. N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *Plos Genet.*, 4(8):e1000167, 2008.
8. J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410, Vienna, Austria, July 2014.
9. K. B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover, G. D. Thomas, S. J. Chanock, and N. Chatterjee. A new statistic and its power to infer membership in a genome-wide association study using genotype

frequencies. *Nat. Genet.*, 41(11):1253–1257, 2009.

10. A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *KDD'13*, pages 1079–1087, Chicago, Illinois, USA, August 2013.
11. C. M. Justice, A. Cuellar, K. Bala, J. A. Sabourin, M. L. Cunningham, K. Crawford, J. M. Phipps, Y. Zhou, D. Cilliers, J. C. Byren, D. Johnson, S. A. Wall, J. E. V. Morton, P. Noons, E. Sweeney, A. Weber, K. E. M. Rees, L. C. Wilson, E. Simeonov, R. Kaneva, N. Yaneva, K. Georgiev, A. Bussarsky, C. Senders, M. Zwienenberg, J. Boggan, T. Roscioli, G. Tamburrini, M. Barba, K. Conway, V. C. Sheffield, L. Brody, J. L. Mills, D. Kay, R. J. Sicko, P. H. Langlois, R. K. Tittle, L. D. Botto, M. M. Jenkins, J. M. LaSalle, W. Lattanzi, A. O. M. Wilkie, A. F. Wilson, P. A. Romitti, and S. A. Boyadjiev. A genome-wide association study implicates the BMP7 locus as a risk factor for nonsyndromic metopic craniosynostosis. *Hum. Genet.*, 139(8):1077–1090, 2020.
12. N. L. Kaplan, E. R. Martin, and B. S. Weir. Power studies for the transmission/disequilibrium tests with multiple alleles. *Am. J. Hum. Genet.*, 60(3):691–702, 1997.
13. Z. Lin, A. B. Owen, and R. B. Altman. Genetics. genomic research and human subject privacy. *Science*, 305(5681):183, 2004.
14. A. G. Matthews, C. Haynes, C. Liu, and J. Ott. Collapsing SNP genotypes in case-control genome-wide association studies increases the type I error rate and power. *Stat. Appl. Genet. Mol. Biol.*, 7(1):1544–6115.1325, 2008.
15. F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103, Providence, RI, USA, October 2007.
16. S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin. Genomic privacy and limits of individual detection in a pool. *Nat. Genet.*, 41:965–967, 2009.
17. P. Sebastiani, N. Timofeev, D. A. Dworkis, T. T. Perls, and M. H. Steinberg. Genome-wide association studies and the genetic dissection of complex traits. *Am. J. Hematol.*, 84(8):504–515, 2009.
18. S. Simmons and B. Berger. Realizing privacy preserving genome-wide association studies. *Bioinformatics*, 32(9):1293–1300, 2016.
19. R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, 52(3):506–516, 1993.
20. J. D. Terwilliger and J. Ott. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum. Hered.*, 42(6):337–346, 1992.
21. G. Thomson. Mapping disease genes: family-based association studies. *Am. J. Hum. Genet.*, 57(2):487–498, 1995.
22. M. Wang, Z. Ji, S. Wang, J. Kim, H. Yang, X. Jiang, and L. Ohno-Machado. Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies. *Bioinformatics*, 33(23):3716–3725, 2017.
23. R. Wang, Y. Fuga Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: Information leaks in genome wide association study. In *CCS'09*, pages 534–544, Chicago, Illinois, USA, November 2009.
24. A. Yamamoto and T. Shibuya. More practical differentially private publication of key statistics in GWAS. in press.
25. F. Yu and Z. Ji. Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. *BMC Med.Inform. Decis. Mak.*, 14(S3), 2014.
26. E. A. Zerhouni and E. G. Nabel. Protecting aggregate genomic data. *Science*, 322(5898):44, 2008.