In Section S1, we describe the key statistics in GWAS and the existing differentially private methods for releasing them. In Section S2, we present detailed proofs of our theorems. In Sections S3, S4 and S5, we supplement the details of our proposed methods. In Section S6, we provide supplemental results and discussion in our experiments.

S1 Statistics in GWAS

In GWAS, we often examine whether there is an association between marker loci, such as SNPs, and diseases. Test methods used in such analyses include χ^2-tests and Cochran-Armitage trend test [1] in case-control studies, and family-based association studies [5, 14, 15]. In addition, there is a risk in genomic analysis that population stratification in a dataset can result in failure to correctly measure the effect of SNPs on diseases, and several statistical methods to correct for this have been proposed [19]. In the following, we discuss the statistics used in these analyses, respectively.

S1.1 Statistics in Case-Control Studies

A typical genomic analysis often performs the χ^2-test with a 2×2 contingency table and that with a 3×2 contingency table. χ^2-statistics and P-values in the χ^2-test based on a 2×2 contingency table are mainly used to compare allele frequencies between the case and the control, whereas those based on a 3×2 contingency table are often used to compare genotype frequencies. When the number of individuals is N, examples of the contingency tables are shown in Tables S1 and S2, respectively.

Table S1. Allele counts distribution for case-control studies.

		Disease Status		Total
		0	1	
Allele	0	a	$m-a$	m
	1	$n-a$	$2N-m-n+a$	$2N-m$
Total		n	$2N-n$	$2N$

Table S2. Genotype counts distribution for case-control studies.

		Disease Status		Total
		0	1	
Genotype	0	a	$p-a$	p
	1	b	$q-b$	q
	2	$r-a-b$	$N-p-q-r+a+b$	$N-p-q$
Total		r	$N-r$	N

For a $K \times 2$ table t with counts $t_{i,j}$ and row sums s_i, the χ^2-statistic is

$$\chi^2(t) = \sum_{i=0}^{K-1} \frac{(t_{i,0} - t_{i,1})^2}{s_i}.$$

Under the null χ^2-distribution, the P-value corresponding to a value x of the χ^2-statistic with a 2×2 table is

$$P = \frac{1}{\sqrt{2\pi}} \int_x^\infty x^{-\frac{1}{2}} \cdot \mathrm{e}^{-\frac{x}{2}} \, \mathrm{d}x,$$

and that with a 3×2 table is

$$P = \mathrm{e}^{-\frac{x}{2}}.$$

When we aim to publish these statistics under central differential privacy and based on the Laplace mechanism, we should analyze the *sensitivity* of each statistic for each contingency table. The *sensitivities* of the χ^2-statistics and P-values for the 2×2 and 3×2 contingency tables have been analyzed in several existing studies [6, 20], but they assume that the number of the case is equal to that of the control. Therefore, further analysis in more general cases, including when using an $M \times N$ contingency table, is required. Also, when publishing the P-values, it was pointed out that considering the values of $\log(P)$ might provide more accurate results [20]. It might be possible to reduce the amount of adding Laplace noise by transforming the statistics to reduce its *sensitivity*, which is worth considering in the future.

Cochran-Armitage Trend Test The Cochran-Armitage trend test [1] is commonly used to determine if there is a trend among binomial proportions in studies where the underlying genetic model is unknown [7]. Here we consider the test for a 3×2 contingency table, and we assume that the genotype counts for the case and those for the control follow independent multinomial distributions with parameters (p_0, p_1, p_2), (p'_0, p'_1, p'_2), respectively, where the parameters are the genotype probabilities in the case and the control. When we take the assumption of no trend to be the null hypothesis, $H_0 : p_i = p'_i$ for $i = 0, 1, 2$. In order to test whether the major and minor alleles are codominant, the weights used in the test are set as $(t_0, t_1, t_2) = (0, 1, 2)$.

In the Cochran-Armitage trend test, we consider a 3×2 contingency table shown in Table S2. The χ^2-statistic in the Cochran-Armitage trend test for the data in the table is given by

$$\chi^2_{CA} = \frac{N \cdot \{(2p+q) \cdot r - N \cdot (2a+b)\}^2}{r \cdot (N-r) \cdot \{N \cdot (4p+q) - (2p+q)^2\}} \quad .$$

As in the case of the χ^2-tests, analysis of *sensitivity* under central differential privacy was conducted [20], but still, there are assumptions regarding the number of cases and controls, and the output accuracy is worse than the previous case.

S1.2 Statistics in Family-Based Association Studies

In family-based genomic analysis, we often examine linkage and correlation between marker loci and diseases. A transmission disequilibrium test (TDT) [13] is

the most common method for family-based studies. The simplest case for TDT is that of trio families where one family has one affected child, and various extended versions of TDT have been proposed [9, 12]. Here, we describe the case of trio families.

We assume that the dataset has n trio families, i.e., $2n$ parents and n affected children and 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 S3.

Table S3. 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:

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

These statistics approximately follow a χ^2-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 S4, 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 S4. 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

Regarding the publication of the TDT statistics, the analysis of *sensitivity* for the Laplace mechanisms was conducted [17]. Therefore, under central differential privacy, it is possible to release while truly preserving privacy. However, they did not consider the case with untrusted data collectors, so this study proposes an analysis procedure using the concept of local differential privacy.

S1.3 Statistics to Correct for Population Stratification

Population stratification is a common and important issue in large-scale genomic analyses. As an example, suppose that the sample data contains a mixture of populations with different genetic backgrounds, such as race. In case-control studies, differences in markers among race might be regarded as disease-related,

resulting in false positive results. To correct for population stratification, various statistical methods have been proposed, including genomic control [4], EIGENSTRAT [10], LAPSTRUCT [21], EMMAX [8], and others [16, 3]. Among these methods, EIGENSTRAT can provide stable correction in all cases [19], so we will discuss EIGENSTRAT in detail here.

Suppose that we have genotype data for M SNPs of N individuals. We let X be the $M \times N$ matrix and X_{sh} be the genotype at SNP s of individual h. In addition, we let Y be the N-dimensional vector representing the phenotype information. First, we construct an $N \times N$ empirical covariance matrix Ψ whose elements satisfy the following equation:

$$\Psi_{ij} = \frac{1}{M} \sum_{s=0}^{M-1} \frac{(X_{si} - 2\hat{p_s})(X_{sj} - 2\hat{p_s})}{2\hat{p_s}(1-\hat{p_s})} \ ,$$

where $\hat{p_s}$ is the allele frequency for the type 1 allele at SNP s. Subsequently, we calculate the top K eigenvalues of Ψ and the corresponding eigenvectors v_1, v_2, $\cdots$, v_K. Using these vectors, we conduct a multiple regression analysis on the following equations:

$$Y_h = \beta + \beta_1 \cdot v_{1h} + \beta_2 \cdot v_{2h} + \cdots + \beta_K \cdot v_{Kh} \qquad (h = 0, 1, \cdots, N-1)$$

and obtain $\hat{Y}$. Then, we define $Y^* = Y - \hat{Y}$. Similarly, we obtain $\hat{X}_s$ and define $X_s^* = X_s - \hat{X}_s$. Consequently, we can obtain the χ^2-statistic for SNP s calculated by

$$\chi^2_{EG} = (N - K - 1) \cdot \frac{(X_s^* \cdot Y^*)^2}{|X_s^*|^2 \ |Y^*|^2} \ .$$

Several methods for conducting EIGENSTRAT while preserving privacy have been proposed [11, 18], along with a method for EMMAX. The procedure of the existing methods for EIGENSTRAT is as follows:

1. Obtain X_s^* and Y^* as in the original procedure.
2. Let

$$U_s = X_s^* \cdot Y^* + Lap\left(0, \frac{2 \max_h |X_{sh}|}{\epsilon}\right) \quad \text{and} \quad Y_{dp} = |Y^*| + Lap\left(0, \frac{2}{\epsilon}\right) \ ,$$

where $Lap(0, \lambda)$ is random noise derived from a Laplace distribution with mean 0 and scale λ.

3. Estimate ϵ-phenotypically differentially private statistic as

$$(N - K - 1) \cdot \frac{U_s^2}{|X_s^*|^2 \ |Y_{dp}|^2} \ .$$

This method can satisfy differential privacy for phenotype information, but not for genotype information because the added noise to $X_s^\star \cdot Y^\star$ depends on datasets and we should add perturbations to $|X_s^*|^2$ as well. Therefore, in this study, we propose a method to satisfy local differential privacy for both kinds of information.

S2 Proofs

Theorem 1. *The expected values of $\tilde{a}$, $\tilde{b}$, $\tilde{c}$, and $\tilde{d}$ are a, b, c, and d, and the variables of them are $\frac{2}{e^\epsilon-1}\,a+\frac{2(e^\epsilon+2)}{(e^\epsilon-1)^2}\,N$, $\frac{2}{e^\epsilon-1}\,b+\frac{2(e^\epsilon+2)}{(e^\epsilon-1)^2}\,N$, $\frac{2}{e^\epsilon-1}\,c+\frac{2(e^\epsilon+2)}{(e^\epsilon-1)^2}\,N$, and $\frac{2}{e^\epsilon-1}\,d+\frac{2(e^\epsilon+2)}{(e^\epsilon-1)^2}\,N$, respectively.*

Proof. We first consider the expected value and variance of a'. From the distortion matrix $\mathbf{P}$,

$$\begin{aligned}
\mathrm{E}(a') &= a\cdot\frac{e^\epsilon}{e^\epsilon+3}+(b+c+d)\cdot\frac{1}{e^\epsilon+3}\\
&= \frac{e^\epsilon}{e^\epsilon+3}\cdot a+\frac{1}{e^\epsilon+3}\cdot(2N-a)=\frac{e^\epsilon-1}{e^\epsilon+3}\cdot a+\frac{2}{e^\epsilon+3}\cdot N\ ,\\
\mathrm{Var}(a') &= a\cdot\frac{e^\epsilon}{e^\epsilon+3}\cdot\frac{3}{e^\epsilon+3}+(2N-a)\cdot\frac{1}{e^\epsilon+3}\cdot\frac{e^\epsilon+2}{e^\epsilon+3}\\
&= \frac{2(e^\epsilon-1)}{(e^\epsilon+3)^2}\cdot a+\frac{2(e^\epsilon+2)}{(e^\epsilon+3)^2}\cdot N\ .
\end{aligned}$$

Here, $\tilde{a}$ can be calculated using a' as follows:

$$\tilde{a}=a'\cdot\frac{e^\epsilon+2}{e^\epsilon-1}-(2N-a')\cdot\frac{1}{e^\epsilon-1}=\frac{e^\epsilon+3}{e^\epsilon-1}\cdot a'-\frac{2}{e^\epsilon-1}\cdot N\ .$$

Thus, the expected value and variance of $\tilde{a}$ can be obtained from the following equations:

$$\begin{aligned}
\mathrm{E}(\tilde{a}) &= \frac{e^\epsilon+3}{e^\epsilon-1}\cdot\mathrm{E}(a')-\frac{2}{e^\epsilon-1}\cdot N=a\ ,\\
\mathrm{Var}(\tilde{a}) &= \left(\frac{e^\epsilon+3}{e^\epsilon-1}\right)^2\cdot\mathrm{Var}(a')=\frac{2}{e^\epsilon-1}\cdot a+\frac{2(e^\epsilon+2)}{(e^\epsilon-1)^2}\cdot N\ .
\end{aligned}$$

As for $\tilde{b}$, $\tilde{c}$, and $\tilde{d}$, we can show in the same way. □

Theorem 2. *The expected values of $\tilde{p}$, $\tilde{q}$, $\tilde{r}$, $\tilde{s}$, $\tilde{t}$, and $\tilde{u}$ are p, q, r, s, t, and u, and the variables of them are $\frac{4}{e^\epsilon-1}\,p+\frac{e^\epsilon+4}{(e^\epsilon-1)^2}\,N$, $\frac{4}{e^\epsilon-1}\,q+\frac{e^\epsilon+4}{(e^\epsilon-1)^2}\,N$, $\frac{4}{e^\epsilon-1}\,r+\frac{e^\epsilon+4}{(e^\epsilon-1)^2}\,N$, $\frac{4}{e^\epsilon-1}\,s+\frac{e^\epsilon+4}{(e^\epsilon-1)^2}\,N$, $\frac{4}{e^\epsilon-1}\,t+\frac{e^\epsilon+4}{(e^\epsilon-1)^2}\,N$, and $\frac{4}{e^\epsilon-1}\,u+\frac{e^\epsilon+4}{(e^\epsilon-1)^2}\,N$, respectively.*

Proof. Similar to the proof of Theorem 1, we first consider the expected value and variance of p'. From the distortion matrix,

$$\begin{aligned}
\mathrm{E}(p') &= p\cdot\frac{e^\epsilon}{e^\epsilon+5}+(N-p)\cdot\frac{1}{e^\epsilon+5}=\frac{e^\epsilon-1}{e^\epsilon+5}\cdot p+\frac{1}{e^\epsilon+5}\cdot N\ ,\\
\mathrm{Var}(p') &= p\cdot\frac{e^\epsilon}{e^\epsilon+5}\cdot\frac{5}{e^\epsilon+5}+(N-p)\cdot\frac{1}{e^\epsilon+5}\cdot\frac{e^\epsilon+4}{e^\epsilon+5}\\
&= \frac{4(e^\epsilon-1)}{(e^\epsilon+5)^2}\cdot p+\frac{e^\epsilon+4}{(e^\epsilon+5)^2}\cdot N\ .
\end{aligned}$$

Here, $\tilde{p}$ can be calculated using p' as follows:

$$\tilde{p} = \frac{e^\epsilon + 5}{e^\epsilon - 1} \cdot p' - \frac{1}{e^\epsilon - 1} \cdot N \ .$$

Thus, the expected value and variance of $\tilde{p}$ are as follows:

$$\begin{aligned} \mathrm{E}(\tilde{p}) &= \frac{e^\epsilon + 5}{e^\epsilon - 1} \cdot \mathrm{E}(p') - \frac{1}{e^\epsilon - 1} \cdot N = p \ , \\ \mathrm{Var}(\tilde{p}) &= \left(\frac{e^\epsilon + 5}{e^\epsilon - 1} \right)^2 \cdot \mathrm{Var}(p') = \frac{4}{e^\epsilon - 1} \cdot p + \frac{e^\epsilon + 4}{(e^\epsilon - 1)^2} \cdot N \ . \end{aligned}$$

As for $\tilde{q}$, $\tilde{r}$, $\tilde{s}$, $\tilde{t}$, and $\tilde{u}$, we can show in the same way. □

Theorem 3. *The sensitivity of the covariance matrix Ψ for EIGENSTRAT under central differential privacy is greater than N, where the number of individuals is $N\,(> 2)$.*

Proof. We discuss the value of

$$\frac{(X_{si} - 2\hat{p_s})(X_{sj} - 2\hat{p_s})}{2\hat{p_s}(1 - \hat{p_s})} \quad (i \neq j). \tag{1}$$

When $(X_{si}, X_{sj}) = (0, 0)$, the range of $\hat{p_s}$ is from 0 to $\frac{2N-4}{2N} = 1 - \frac{2}{N}$. Since

$$(1) = \frac{(-2\hat{p_s})(-2\hat{p_s})}{2\hat{p_s}(1 - \hat{p_s})} = -2 + \frac{2}{1 - \hat{p_s}} \ ,$$

we can show $0 \leq (1) \leq -2 + N$.

When $(X_{si}, X_{sj}) = (0, 1)$ and $(1, 0)$, the range of $\hat{p_s}$ is from $\frac{1}{2N}$ to $\frac{2N-3}{2N} = 1 - \frac{3}{2N}$. Since

$$(1) = \frac{(-2\hat{p_s})(1 - 2\hat{p_s})}{2\hat{p_s}(1 - \hat{p_s})} = -2 + \frac{1}{1 - \hat{p_s}} \ ,$$

we can show $-1 + \frac{1}{2N-1} \leq (1) \leq -2 + \frac{2N}{3}$.

When $(X_{si}, X_{sj}) = (0, 2)$ and $(2, 0)$,

$$(1) = \frac{(-2\hat{p_s})(2 - 2\hat{p_s})}{2\hat{p_s}(1 - \hat{p_s})} = -2 \ .$$

When $(X_{si}, X_{sj}) = (1, 1)$, the range of $\hat{p_s}$ is from $\frac{1}{N}$ to $\frac{2N-2}{2N} = 1 - \frac{1}{N}$. Since

$$(1) = \frac{(1 - 2\hat{p_s})(1 - 2\hat{p_s})}{2\hat{p_s}(1 - \hat{p_s})} = -2 + \frac{1}{2\hat{p_s}(1 - \hat{p_s})} \ ,$$

we can show $0 \leq (1) \leq -2 + \frac{N^2}{2(N-1)}$.

When $(X_{si}, X_{sj}) = (1,2)$ and $(2,1)$, the range of $\hat{p_s}$ is from $\frac{3}{2N}$ to $\frac{2N-1}{2N} = 1 - \frac{1}{2N}$. Since

$$(1) = \frac{(1-2\hat{p_s})(2-2\hat{p_s})}{2\hat{p_s}(1-\hat{p_s})} = -2 + \frac{1}{\hat{p_s}} \ ,$$

we can show $-1 + \frac{1}{2N-1} \leq (1) \leq -2 + \frac{2N}{3}$.

When $(X_{si}, X_{sj}) = (2,2)$, the range of $\hat{p_s}$ is from $\frac{2}{N}$ to 1. Since

$$(1) = \frac{(2-2\hat{p_s})(2-2\hat{p_s})}{2\hat{p_s}(1-\hat{p_s})} = -2 + \frac{2}{\hat{p_s}} \ ,$$

we can show $0 \leq (1) \leq -2 + N$.

Thus, the maximum change in (1) between neighboring datasets is N, when $(X_{si}, X_{sj}, \hat{p_s})$ is varied from $\left(0, 0, 1 - \frac{2}{N}\right)$ to $\left(0, 2, 1 - \frac{1}{N}\right)$ or from $\left(2, 2, \frac{2}{N}\right)$ to $\left(0, 2, \frac{1}{N}\right)$, and vice versa.

Therefore, the value of Ψ_{ij} can change by

$$\frac{1}{M} \sum_{s=0}^{M-1} N = N,$$

and consequently, the *sensitivity* of Ψ is larger than N. □

S3 EM Algorithm for Case 1: ϵ for the entire table using a 3×2 contingency table

In the main document, we provide the reconstruction procedure when using a 2×2 contingency table. Here, we show the algorithm for the case of a 3×2 contingency table.

i. **Initialization:**
Let N be the number of individuals in the dataset. Create $x \in \mathbb{R}^{N \times \{p,q,r,s,t,u\}}$ s.t.

$$x_{h,i} = \begin{cases} 1 & (\text{Individual } h \text{ belongs to } i.) \\ 0 & (\text{otherwise}) \end{cases}.$$

Set $\theta_p^0 = \theta_q^0 = \theta_r^0 = \theta_s^0 = \theta_t^0 = \theta_u^0 = \frac{1}{6}$.

ii. **E-Step:**
For any individual h $(0 \leq h < N)$ and $i \in \{p, q, r, s, t, u\}$,

$$\begin{aligned} \theta_{h,i}^k = \Pr[z_{h,i} = 1 | x_{h,i}] &= \frac{\Pr[x_{h,i} | z_{h,i} = 1] \cdot \Pr[z_{h,i} = 1]}{\sum_{j \in \{p,q,r,s,t,u\}} \Pr[x_{h,i} | z_{h,j} = 1] \cdot \Pr[z_{h,j} = 1]} \\ &= \frac{\Pr[x_{h,i} | z_{h,i} = 1] \cdot \theta_i^{k-1}}{\sum_{j \in \{p,q,r,s,t,u\}} \Pr[x_{h,i} | z_{h,j} = 1] \cdot \theta_j^{k-1}}. \end{aligned}$$

iii. **M-Step:**

$$\theta_i^k = \frac{1}{N} \sum_{h=0}^{N-1} \theta_{h,i}^k$$

iv. Repeat steps ii and iii until $\sum_i \left| \theta_i^k - \theta_i^{k-1} \right| < \delta$ for some $\delta > 0$, then calculate $\tilde{p} = N \cdot \theta_p^k$, $\tilde{q} = N \cdot \theta_q^k$, $\tilde{r} = N \cdot \theta_r^k$, $\tilde{s} = N \cdot \theta_s^k$, $\tilde{t} = N \cdot \theta_t^k$, and $\tilde{u} = N \cdot \theta_u^k$.

Here, $z_{h,j}$ satisfies the following equations:

$$\Pr[x_{h,i} | z_{h,j}] = \begin{cases} \begin{cases} \frac{e^\epsilon}{e^\epsilon + 5} & (x_{h,i} = 1) \\ \frac{5}{e^\epsilon + 5} & (x_{h,i} = 0) \end{cases} & (i = j) \\ \begin{cases} \frac{1}{e^\epsilon + 5} & (x_{h,i} = 1) \\ \frac{e^\epsilon + 4}{e^\epsilon + 5} & (x_{h,i} = 0) \end{cases} & (i \neq j) \end{cases}.$$

S4 EM Algorithm for the case of TDT

We also present the EM algorithm in local differentially private methods for TDT.

i. **Initialization:**
Let n be the number of families in the dataset.
Create $x \in \mathbb{R}^{n\times\{(1,0),(0,1),(1,1),(2,0),(0,2),(0,0)\}}$ s.t.

$$x_{h,i} = \begin{cases} 1 & ((\text{b,c}) \text{ of family } h \text{ is } i.) \\ 0 & (\text{otherwise}) \end{cases} .$$

Set $\theta^0_{(1,0)} = \theta^0_{(0,1)} = \theta^0_{(1,1)} = \theta^0_{(2,0)} = \theta^0_{(0,2)} = \theta^0_{(0,0)} = \frac{1}{6}$.

ii. **E-Step:**
For any family h $(0 \le h < n)$ and $i \in \{(1,0),(0,1),(1,1),(2,0),(0,2),(0,0)\}$,

$$\begin{aligned} \theta^k_{h,i} = \Pr[z_{h,i} = 1 | x_{h,i}] &= \frac{\Pr[x_{h,i}|z_{h,i}=1]\cdot\Pr[z_{h,i}=1]}{\sum_j \Pr[x_{h,i}|z_{h,j}=1]\cdot\Pr[z_{h,j}=1]} \\ &= \frac{\Pr[x_{h,i}|z_{h,i}=1]\cdot\theta^{k-1}_i}{\sum_j \Pr[x_{h,i}|z_{h,j}=1]\cdot\theta^{k-1}_j}. \end{aligned}$$

iii. **M-Step:**

$$\theta^k_i = \frac{1}{n}\sum_{h=0}^{n-1}\theta^k_{h,i}$$

iv. Repeat steps ii and iii until $\sum_i \left|\theta^k_i - \theta^{k-1}_i\right| < \delta$ for some $\delta > 0$, then calculate the number of families with each (b,c) attribute value by $n\cdot\theta^k_i$.

Here, $z_{h,j}$ in the EM algorithm satisfies the following equations:

$$\Pr[x_{h,i}|z_{h,j}] = \begin{cases} \begin{cases} \frac{e^\epsilon}{e^\epsilon+5} & (x_{h,i}=1) \\ \frac{5}{e^\epsilon+5} & (x_{h,i}=0) \end{cases} & (i=j) \\ \begin{cases} \frac{1}{e^\epsilon+5} & (x_{h,i}=1) \\ \frac{e^\epsilon+4}{e^\epsilon+5} & (x_{h,i}=0) \end{cases} & (i\neq j) \end{cases} .$$

S5 Detailed Discussion for Case 2: ϵ_1 for row and ϵ_2 for column using a 3×2 contingency table

In this section, we discuss the 6×6 distortion matrix $\mathbf{P}$ to perturb the following table data.

		Disease Status		Total
		0	1	
Genotype	0	p	q	$p+q$
	1	r	s	$r+s$
	2	t	u	$t+u$
Total		$p+r+t$	$q+s+u$	N

As in the case for 2×2 contingency table, we focus on the first column of $\mathbf{P}$: $(p_{00}, p_{10}, p_{20}, p_{30}, p_{40}, p_{50})^{\mathsf{T}}$. Here, we let

$$P = \frac{p_{00}}{p_{50}},\ Q = \frac{p_{10}}{p_{50}},\ R = \frac{p_{20}}{p_{50}},\ S = \frac{p_{30}}{p_{50}},\ T = \frac{p_{40}}{p_{50}},\ \text{and } U = \frac{p_{50}}{p_{50}} = 1.$$

Given that $p_{20} = p_{40}$ and $p_{30} = p_{50}$, the values from P to U can satisfy the following conditions:

$$\begin{cases} P \geq Q, R, T \geq S = U = 1 \\ R = T \\ \frac{P+Q}{R+S} = \frac{P+Q}{T+U} = e^{\epsilon_1} \\ \frac{P+R+T}{Q+S+U} = e^{\epsilon_2} \end{cases} \iff \begin{cases} P \geq Q, R, T \geq S = U = 1 \\ R = T \\ (e^{\epsilon_2}+1)\,Q - (e^{\epsilon_1}+2)\,R = e^{\epsilon_1} - 2e^{\epsilon_2} \\ R = \frac{e^{\epsilon_2}}{2} \cdot Q + e^{\epsilon_2} - \frac{P}{2} \end{cases}.$$

If

$$\frac{e^{\epsilon_2}}{2} \geq \frac{e^{\epsilon_2}+1}{e^{\epsilon_1}+2} \iff e^{\epsilon_1+\epsilon_2} \geq 2,$$

P is minimized when Q or R is 1. When Q can be 1,

$$\frac{2\,(e^{\epsilon_1} - e^{\epsilon_2} + 1)}{e^{\epsilon_2}+1} \geq 1 \iff 2e^{\epsilon_1} - 3e^{\epsilon_2} + 1 \geq 0.$$

If $e^{\epsilon_1+\epsilon_2} < 2$, P is minimized when $P = Q$ or $P = R$. When P can be Q under $P \geq R$,

$$\begin{aligned} &\frac{2e^{\epsilon_1+\epsilon_2} + 2e^{\epsilon_1}}{-e^{\epsilon_1+\epsilon_2} + e^{\epsilon_1} + 4} \geq \frac{e^{\epsilon_1+\epsilon_2} - e^{\epsilon_1} + 4e^{\epsilon_2}}{-e^{\epsilon_1+\epsilon_2} + e^{\epsilon_1} + 4} \\ \iff\ & e^{\epsilon_1+\epsilon_2} + 3e^{\epsilon_1} - 4e^{\epsilon_2} \geq 0. \end{aligned}$$

Therefore, we should consider the following four cases:

$$\begin{aligned} &\text{(I)} && e^{\epsilon_1+\epsilon_2} \geq 2 \ \wedge\ 2e^{\epsilon_1} - 3e^{\epsilon_2} + 1 \geq 0, \\ &\text{(II)} && e^{\epsilon_1+\epsilon_2} \geq 2 \ \wedge\ 2e^{\epsilon_1} - 3e^{\epsilon_2} + 1 < 0, \\ &\text{(III)} && e^{\epsilon_1+\epsilon_2} < 2 \ \wedge\ e^{\epsilon_1+\epsilon_2} + 3e^{\epsilon_1} - 4e^{\epsilon_2} \geq 0, \\ \text{and}\ &\text{(IV)} && e^{\epsilon_1+\epsilon_2} < 2 \ \wedge\ e^{\epsilon_1+\epsilon_2} + 3e^{\epsilon_1} - 4e^{\epsilon_2} < 0. \end{aligned}$$

In Case (I), P is minimized when

$$(P,Q,R,S,T,U)=\left(\frac{2(e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}-1)}{e^{\epsilon_2}+1},\frac{2(e^{\epsilon_1}-e^{\epsilon_2}+1)}{e^{\epsilon_2}+1},1,1,1,1\right).$$

The second through sixth columns can be discussed in the same manner, and the elements of $\mathbf{P}$ are as follows:

$$\mathbf{P}_{uv}=\begin{cases}\frac{e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}-1}{(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (u=v)\\ \frac{e^{\epsilon_1}-e^{\epsilon_2}+1}{(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & ((u,v)\text{ or }(v,u)=(0,1),(2,3),(4,5))\\ \frac{1}{2(e^{\epsilon_1}+2)} & (\text{otherwise})\end{cases}.$$

The inverse matrix $\mathbf{P}^{-1}$ and $z_{h,j}$ in the EM algorithm for reconstruction satisfy the following equations:

$$\mathbf{P}^{-1}_{uv}=\begin{cases}\frac{e^{\epsilon_1+\epsilon_2}-1}{(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (u=v)\\ \frac{-e^{\epsilon_1}+e^{\epsilon_2}}{(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & ((u,v)\text{ or }(v,u)=(0,1),(2,3),(4,5))\\ \frac{-1}{2(e^{\epsilon_1}-1)} & (\text{otherwise})\end{cases},$$

$$\Pr[x_{h,i}|z_{h,j}]=\begin{cases}\begin{cases}\frac{e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}-1}{(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{e^{\epsilon_1}+e^{\epsilon_2}+3}{(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & (i=j)\\ \begin{cases}\frac{e^{\epsilon_1}-e^{\epsilon_2}+1}{(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_2}+1}{(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & \begin{array}{l}((i,j)\text{ or }(j,i)=\\ (p,q),(r,s),(t,u))\end{array}\\ \begin{cases}\frac{1}{2(e^{\epsilon_1}+2)} & (x_{h,i}=1)\\ \frac{2e^{\epsilon_1}+3}{2(e^{\epsilon_1}+2)} & (x_{h,i}=0)\end{cases} & (\text{otherwise})\end{cases}.$$

In Case (II), P is minimized when

$$(P,Q,R,S,T,U)=\left(\frac{3e^{\epsilon_1+\epsilon_2}+2e^{\epsilon_1}-2}{e^{\epsilon_1}+2},1,\frac{-e^{\epsilon_1}+3e^{\epsilon_2}+1}{e^{\epsilon_1}+2},1,\frac{-e^{\epsilon_1}+3e^{\epsilon_2}+1}{e^{\epsilon_1}+2},1\right),$$

and the elements of $\mathbf{P}$ are as follows:

$$\mathbf{P}_{uv}=\begin{cases}\frac{3e^{\epsilon_1+\epsilon_2}+2e^{\epsilon_1}-2}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (u=v)\\ \frac{-e^{\epsilon_1}+3e^{\epsilon_2}+1}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (|u-v|=2,4)\\ \frac{1}{3(e^{\epsilon_2}+1)} & (\text{otherwise})\end{cases}.$$

The elements of $\mathbf{P}^{-1}$ and $z_{h,j}$ satisfy the following equations:

$$\mathbf{P}^{-1}_{uv}=\begin{cases}\frac{3e^{\epsilon_1+\epsilon_2}-2e^{\epsilon_1}+3e^{\epsilon_2}-4}{3(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (u=v)\\ \frac{e^{\epsilon_1}-3e^{\epsilon_2}+2}{3(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (|u-v|=2,4)\\ \frac{-1}{3(e^{\epsilon_2}-1)} & (\text{otherwise})\end{cases},$$

$$\Pr[x_{h,i}|z_{h,j}]=\begin{cases}\begin{cases}\frac{3e^{\epsilon_1+\epsilon_2}+2e^{\epsilon_1}-2}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{e^{\epsilon_1}+6e^{\epsilon_2}+8}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & (i=j)\\ \begin{cases}\frac{-e^{\epsilon_1}+3e^{\epsilon_2}+1}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{3e^{\epsilon_1+\epsilon_2}+4e^{\epsilon_1}+3e^{\epsilon_2}+5}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & \begin{array}{l}((i,j) \text{ or } (j,i)=\\ (p,r),(p,t),(q,s),\\ (q,u),(r,t),(s,u))\end{array}\\ \begin{cases}\frac{1}{3(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{3e^{\epsilon_2}+2}{3(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & (\text{otherwise})\end{cases}.$$

In Case (III), P is minimized when

$$\begin{aligned}&(P,Q,R,S,T,U)\\ &=\left(\frac{2e^{\epsilon_1}(e^{\epsilon_2}+1)}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}+4},\frac{2e^{\epsilon_1}(e^{\epsilon_2}+1)}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}+4},\frac{e^{\epsilon_1+\epsilon_2}-e^{\epsilon_1}+4e^{\epsilon_2}}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}+4},1,\frac{e^{\epsilon_1+\epsilon_2}-e^{\epsilon_1}+4e^{\epsilon_2}}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}+4},1\right).\end{aligned}$$

Then, as in the previous cases, the elements of $\mathbf{P}$ and $\mathbf{P}^{-1}$ and $z_{h,j}$ for the EM algorithm satisfy the following equations:

$$\mathbf{P}_{uv}=\begin{cases}\frac{e^{\epsilon_1}}{2(e^{\epsilon_1}+2)} & (u=v\ \vee\ (u,v) \text{ or } (v,u)=(0,1),(2,3),(4,5))\\ \frac{e^{\epsilon_1+\epsilon_2}-e^{\epsilon_1}+4e^{\epsilon_2}}{4(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (|u-v|=2,4)\\ \frac{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}+4}{4(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (\text{otherwise})\end{cases},$$

$$\mathbf{P}^{-1}_{uv}=\begin{cases}\frac{-e^{\epsilon_1}+e^{\epsilon_2}}{(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (u=v)\\ \frac{e^{\epsilon_1+\epsilon_2}-1}{(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & ((u,v) \text{ or } (v,u)=(0,1),(2,3),(4,5))\\ \frac{e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}-2e^{\epsilon_2}}{2(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (|u-v|=2,4)\\ \frac{-e^{\epsilon_1+\epsilon_2}-e^{\epsilon_1}+2}{2(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (\text{otherwise})\end{cases},$$

$$\Pr[x_{h,i}|z_{h,j}]=\begin{cases}\begin{cases}\frac{e^{\epsilon_1}}{2(e^{\epsilon_1}+2)} & (x_{h,i}=1)\\ \frac{e^{\epsilon_1}+4}{2(e^{\epsilon_1}+2)} & (x_{h,i}=0)\end{cases} & \begin{array}{l}(i=j\ \vee\ (i,j) \text{ or } (j,i)\\ =(p,q),(r,s),(t,u))\end{array}\\ \begin{cases}\frac{e^{\epsilon_1+\epsilon_2}-e^{\epsilon_1}+4e^{\epsilon_2}}{4(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{3e^{\epsilon_1+\epsilon_2}+5e^{\epsilon_1}+4e^{\epsilon_2}+8}{4(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & \begin{array}{l}((i,j) \text{ or } (j,i)=\\ (p,r),(p,t),(q,s),\\ (q,u),(r,t),(s,u))\end{array}\\ \begin{cases}\frac{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}+4}{4(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{5e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_1}+8e^{\epsilon_2}+4}{4(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & (\text{otherwise})\end{cases}.$$

In Case (IV), P is minimized when

$$(P,Q,R,S,T,U)$$
$$=\left(\frac{e^{\epsilon_2}(e^{\epsilon_1}+2)}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}+3},\frac{2e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_1}-2e^{\epsilon_2}}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}+3},\frac{e^{\epsilon_2}(e^{\epsilon_1}+2)}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}+3},1,\frac{e^{\epsilon_2}(e^{\epsilon_1}+2)}{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}+3},1\right),$$

and the elements of $\mathbf{P}$ and $\mathbf{P}^{-1}$ and $z_{h,j}$ satisfy the following equations:

$$\mathbf{P}_{uv}=\begin{cases}\frac{e^{\epsilon_2}}{3(e^{\epsilon_2}+1)} & (u=v\ \vee\ |u-v|=2,4)\\ \frac{2e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_1}-2e^{\epsilon_2}}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & ((u,v)\text{ or }(v,u)=(0,1),(2,3),(4,5))\\ \frac{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}+3}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (\text{otherwise})\end{cases},$$

$$\mathbf{P}^{-1}_{uv}=\begin{cases}\frac{e^{\epsilon_2}}{3(e^{\epsilon_2}-1)} & (u=v\ \vee\ |u-v|=2,4)\\ \frac{2e^{\epsilon_1+\epsilon_2}-3e^{\epsilon_1}+4e^{\epsilon_2}-3}{3(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & ((u,v)\text{ or }(v,u)=(0,1),(2,3),(4,5))\\ \frac{-e^{\epsilon_1+\epsilon_2}-2e^{\epsilon_2}+3}{3(e^{\epsilon_1}-1)(e^{\epsilon_2}-1)} & (\text{otherwise})\end{cases},$$

$$\Pr[x_{h,i}|z_{h,j}]=\begin{cases}\begin{cases}\frac{e^{\epsilon_2}}{3(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{2e^{\epsilon_2}+3}{3(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & (i=j\ \vee\ (i,j)\text{ or }(j,i)=(p,r),(p,t),(q,s),(q,u),(r,t),(s,u))\\ \begin{cases}\frac{2e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_1}-2e^{\epsilon_2}}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{e^{\epsilon_1+\epsilon_2}+8e^{\epsilon_2}+6}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & ((i,j)\text{ or }(j,i)=(p,q),(r,s),(t,u))\\ \begin{cases}\frac{-e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}+3}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=1)\\ \frac{4e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_1}+5e^{\epsilon_2}+3}{3(e^{\epsilon_1}+2)(e^{\epsilon_2}+1)} & (x_{h,i}=0)\end{cases} & (\text{otherwise})\end{cases}.$$

S6 Supplemental Results and Discussion

S6.1 χ^2-Tests

3×2 Contingency Table In the case of using a 3×2 contingency table, we compared our method to the existing method [6]. As in the case of a 2×2 contingency table (see the main document), the existing method utilizes the Laplace mechanism under central differential privacy and has the limitation on the number of cases and controls. The results are shown in Fig. S1.

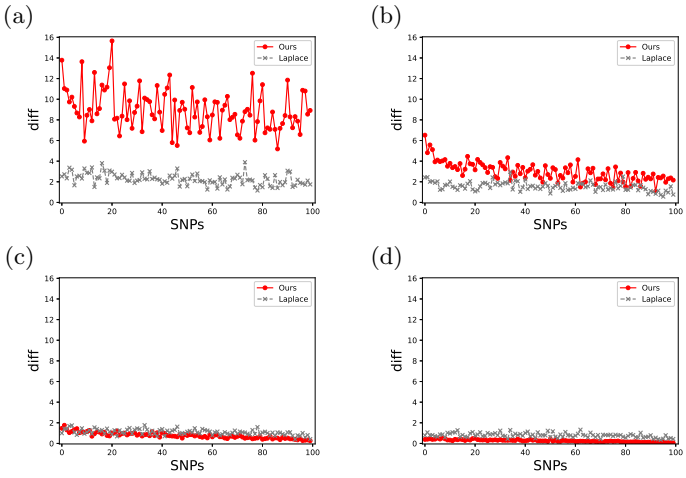

Fig. S1. Differences between original and differentially private χ^2-statistics on 100 SNPs using a 3×2 contingency table when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

Compared to the existing method, our method can achieve higher accuracy when ϵ is larger. However, the trend of the difference being larger when the original statistics increases and the value of ϵ decreases can also be seen in Fig. S1. In addition, the differences are greater than those in the case of a 2×2 contingency table, especially when ϵ is small. This might be because that the elements in a 3×2 contingency table are smaller than those in a 2×2 table for genomic statistical tests. Here, note that the variance of the recovered value of each element does not make much difference whether using either contingency table (see Theorems 1 and 2) and we generated the table data at random. If the amount of change in an element is the same, the χ^2-statistic is expected to vary larger when the original element is smaller, and consequently, the differences from the original statistics would become larger for a 3×2 contingency table. In addition, when the variance is larger, i.e., when ϵ is small, the change in the statistic would be more apparent as shown in Case (a) of Fig. 1 (see the main document) and Fig. S1.

S6.2 Cochran-Armitage Trend Test

Next, we show the results on the Cochran-Armitage trend test using a 3×2 contingency table. As in the case of the χ^2-tests, we measured difference between the original and differential private statistics. The existing method [20] as a comparison employs the Laplace mechanism. The results are shown in Fig. S2.

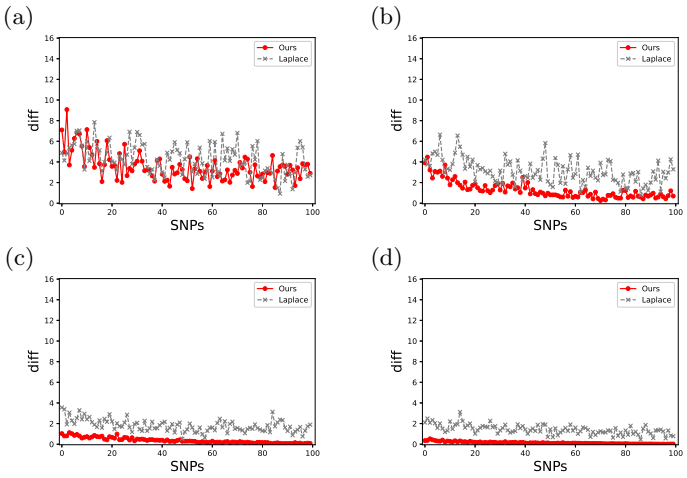

Fig. S2. Differences between original and differentially private χ^2-statistics for the Cochran-Armitage trend test on 100 SNPs when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

When using the Laplace mechanism, the accuracy becomes worse than the results in Fig. S1 because the *sensitivity* of the statistic for the Cochran-Armitage trend test is greater than that for the χ^2-test. On the other hand, our method can provide higher accuracy than the previous case. One possible reason is that even if each element of the contingency table is dispersed to some extent, the linear trend of the entire table will be less varied than the relevance of the row and column information. However, it is the same for both tests that the variance of recovered elements becomes larger when the original value increases and ϵ decreases, and Fig. S2 has the similar trend to the figures in the previous case.

S6.3 TDT

Then, we show the results on the TDT for family-based studies in Fig. S3. Unlike the previous cases, the existing method using the Laplace mechanism [17] does not have restrictions on (b, c) values to calculate the TDT statistics. Therefore, the main difference between the existing method and ours is that whether it is under central or local differential privacy.

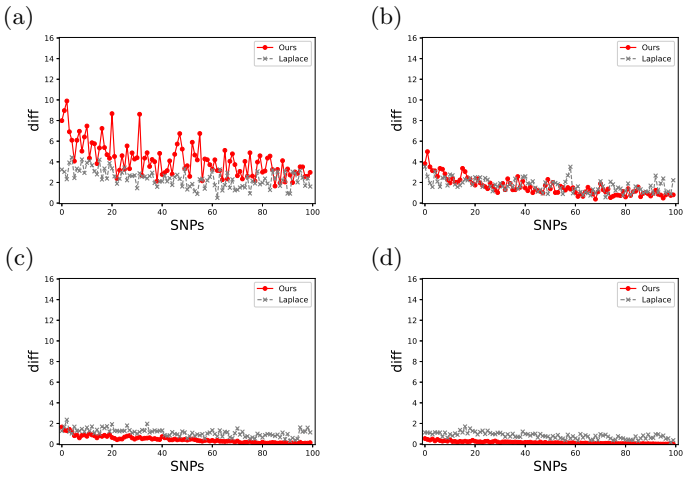

Fig. S3. Differences between original and differentially private TDT statistics on 100 SNPs when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

Fig. S3 shows that our method outperforms the existing method when ϵ is large, but as in the previous cases, the existing method is superior when $\epsilon = 1$ and the original statistics are large. It has been pointed out that local differential privacy might achieve stronger privacy assurance than central differential privacy [2], and the appropriate values of ϵ for genomic statistical analysis under local differential privacy is open to further discussion.

S6.4 EIGENSTRAT

Finally, we show the results on the EIGENSTRAT for correcting for population stratification in Fig. S4.

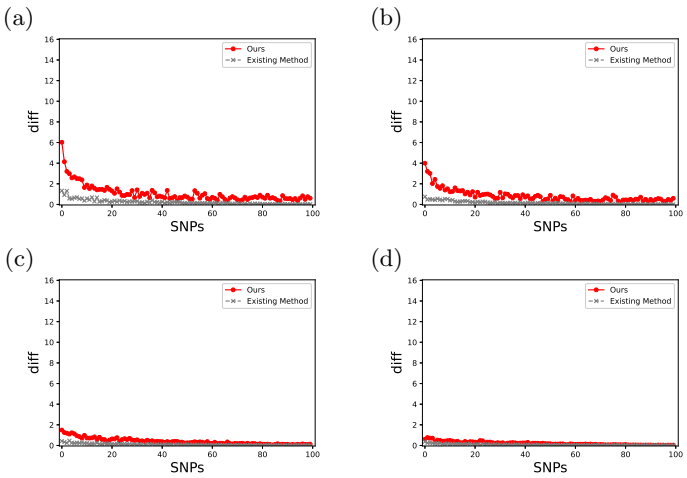

Fig. S4. Differences between original and differentially private EIGENSTRAT statistics on 100 SNPs when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

Because the existing method [11, 18] can only protect phenotype information, the differences are smaller than those in the previous cases. On the other hand, our method also protects the targeted genotype information, which increases the differences, but maintains high accuracy. This might be because all the SNP information in the data is used to compute the statistic, so we should develop methods with stronger privacy guarantees in the future.

References

1. Armitage, P.: Tests for linear trends in proportions and frequencies. Biometrics **11**(3), 375–386 (1955). https://doi.org/10.2307/3001775
2. Bernau, D., Robl, J., Grassal, P.W., Schneider, S., Kerschbaum, F.: Comparing local and central differential privacy using membership inference attacks. In: Data and Applications Security and Privacy XXXV: 35th Annual IFIP WG 11.3 Conference, DBSec 2021, Calgary, Canada, July 19–20, 2021, Proceedings. pp. 22–42 (2021). https://doi.org/10.1007/978-3-030-81242-3_2
3. Bouaziz, M., Mullaert, J., Bigio, B., Seeleuthner, Y., Casanova, J.L., Alcais, A., Abel, L., Cobat, A.: Controlling for human population stratification in rare variant association studies. Sci. Rep. **11**, 19015 (2021). https://doi.org/10.1038/s41598-021-98370-5
4. Devlin, B., Roeder, K.: Genomic control for association studies. Biometrics **55**(4), 997–1004 (1999). https://doi.org/10.1111/j.0006-341x.1999.00997.x
5. Falk, C.T., Rubinstein, P.: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann. Hum. Genet. **51**(3), 227–233 (1987). https://doi.org/10.1111/j.1469-1809.1987.tb00875.x
6. Fienberg, S.E., Slavkovic, A., Uhler, C.: Privacy preserving GWAS data sharing. In: IEEE 11th International Conference on Data Mining Workshops. pp. 628–635 (2011). https://doi.org/10.1109/ICDMW.2011.140
7. Ghodsi, M., Amiri, S., Hassani, H., Ghodsi, Z.: An enhanced version of Cochran-Armitage trend test for genome-wide association studies. Meta Gene **9**, 225–229 (2016). https://doi.org/10.1016/j.mgene.2016.07.001
8. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., yee Kong, S., Freimer, N.B., Sabatti, C., Eskin, E.: Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42**, 348–354 (2010). https://doi.org/10.1038/ng.548
9. Kaplan, N.L., Martin, E.R., Weir, B.S.: Power studies for the transmission/disequilibrium tests with multiple alleles. Am. J. Hum. Genet. **60**(3), 691–702 (1997)
10. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**, 904–909 (2006). https://doi.org/10.1038/ng1847
11. Simmons, S., Sahinalp, C., Berger, B.: Enabling privacy-preserving GWASs in heterogeneous human populations. Cell Syst. **3**(1), 54–61 (2016). https://doi.org/10.1016/j.cels.2016.04.013
12. Spielman, R.S., Ewens, W.J.: A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. **62**(2), 450–458 (1998). https://doi.org/10.1086/301714
13. Spielman, R.S., McGinnis, R.E., Ewens, W.J.: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. **52**(3), 506–516 (1993)
14. Terwilliger, J.D., Ott, J.: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum. Hered. **42**(6), 337–346 (1992). https://doi.org/10.1159/000154096
15. Thomson, G.: Mapping disease genes: family-based association studies. Am. J. Hum. Genet. **57**(2), 487–498 (1995)
16. Thornton, T., McPeek, M.S.: ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am. J. Hum. Genet. **86**(2), 172–184 (2010). https://doi.org/10.1016/j.ajhg.2010.01.001

17. Wang, M., Ji, Z., Wang, S., Kim, J., Yang, H., Jiang, X., Ohno-Machado, L.: Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies. Bioinformatics **33**(23), 3716–3725 (2017). https://doi.org/10.1093/bioinformatics/btx470
18. Wei, J., Lin, Y., Yao, X., Zhang, J., Liu, X.: Differential privacy-based genetic matching in personalized medicine. IEEE Transactions on Emerging Topics in Computing **9**(3), 1109–1125 (2021). https://doi.org/10.1109/TETC.2020.2970094
19. Wu, C., DeWan, A., Hoh, J., Wang, Z.: A comparison of association methods correcting for population stratification in case-control studies. Ann. Hum. Genet **75**(3), 418–27 (2011). https://doi.org/10.1111/j.1469-1809.2010.00639.x
20. Yamamoto, A., Shibuya, T.: More practical differentially private publication of key statistics in GWAS. Bioinformatics Advances **1**(1) (2021). https://doi.org/10.1093/bioadv/vbab004
21. Zhang, J., Niyogi, P., McPeek, M.S.: Laplacian eigenfunctions learn population structure. PLoS One **4**(12), e7928 (2009). https://doi.org/10.1371/journal.pone.0007928