In Section S1, we review the related studies. In Section S2, we describe in detail the key statistics in GWAS and the existing differentially private methods for releasing them, including their problems. In addition, we describe differential privacy and the randomized response technique along with the related studies. In Sections S3 and S4, we supplement the details of our proposed methods for case-control studies. In Section S5, we propose local differentially private methods for releasing transmission disequilibrium test (TDT) statistics and EIGENSTRAT. In Section S6, we present detailed proofs of our theorems. In Section S7, we provide supplemental results and discussion on the experiments.

S1 Related Work

S1.1 Privacy-Preserving Methods for Genomic Data

To protect genomic data, various methods using homomorphic encryption, multiparty computation, and trusted execution environments [8, 9, 11, 29] have been developed, but they must additionally protect the calculation results. To release and utilize the analysis results while protecting privacy, applying the concept of differential privacy is widely considered. Regarding statistics in GWAS, Fienberg *et al.* [19] first proposed differentially private methods for releasing χ^2-statistics and P-values from the χ^2-test using a contingency table. Subsequently, several privacy-preserving methods were proposed for the Cochran–Armitage trend test [45], transmission disequilibrium tests [37, 43], and statistical analysis to correct for population stratification [31]. These existing studies assume that the data collector is trustworthy, and there is still no method for releasing genome statistics under local differential privacy. Furthermore, the existing methods for statistical tests using a contingency table have restrictions on the number of cases and controls, and the methods for correcting for population stratification can only protect phenotype information and do not satisfy differential privacy for genotype information.

S1.2 Local Differential Privacy for Statistical Tests

The existing study on statistical tests under local differential privacy is for the χ^2-tests in hypothesis testing [20]. They proposed three methods using the Laplace mechanism [13], randomized response [40], and bit flipping [5] similar to RAPPOR [16] to satisfy ϵ-local differential privacy. Among these, the method with the randomized response achieved the highest accuracy. Referring to the results, we employ the randomized response technique to develop local differentially private methods for genome statistics.

The randomized response is often utilized for frequency and mean estimation, and several studies have proposed mechanisms to minimize estimator errors [22, 25]. These existing studies include procedures to recover the original data from perturbed data, and there are two reconstruction methods: using a matrix inversion [24, 39] and using an EM algorithm [7, 18]. In this study, we present both of these two methods for privacy-preserving statistical genomic analysis and conduct an error analysis for the case using an inverse matrix. Furthermore, unlike the hypothesis testing addressed in the existing study [20], genome analysis often requires consideration of two-attribute data, for example, genotype and phenotype information. For this case, we propose novel randomized response techniques that provide stronger privacy guarantees than the existing method for multiple-attribute data [39].

S2 Preliminaries

S2.1 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 [4] in case-control studies, and family-based association studies [17, 34, 35]. 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 for correcting for this have been proposed [42]. In the following, we discuss the statistics used in these analyses, respectively.

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	b	$a+b$
	1	c	d	$c+d$
Total		$a+c$	$b+d$	$2N$

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

		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

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 [19, 45], 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 [45]. 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 [4] is commonly used to determine if there is a trend among binomial proportions in studies where the underlying genetic model is unknown [21]. 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

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

As in the case of the χ^2-tests, analysis of *sensitivity* under central differential privacy was conducted [45], but still, there are assumptions regarding the number of cases and controls, and the output accuracy is worse than the previous case.

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) [33] 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 [27, 32]. 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	M_1	a	b	$a+b$
Allele	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 [37]. 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.

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 [12], EIGENSTRAT [30], LAPSTRUCT [47], EMMAX [26], and others [10, 36]. Among these methods, EIGENSTRAT can provide stable correction in all cases [42], 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 [31, 41], 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 for satisfying local differential privacy for both kinds of information.

S2.2 Differential Privacy

Differential privacy [14] is a framework that enables data analysis while protecting a particular individual in the dataset from adversaries and was developed in the field of cryptography. In recent years, it has been well applied to statistical data of human genomes [2, 19, 37, 45, 46] and has also been widely used in various fields [1, 3]. The concept of differential privacy is based on the idea that the results obtained from two *neighboring* datasets that differ in just one element are nearly indistinguishable.

The privacy level can be evaluated by the parameter $\epsilon > 0$. As this value decreases, the privacy guarantee becomes stronger. In general, the value of ϵ is set in the range from 0.01 to 10 [23]. ϵ-differential privacy is defined as follows:

Definition S1. (*ϵ-Differential Privacy [14]*)
A randomized mechanism M is ϵ-differentially private if, for all neighboring datasets D and D' and any $S \subset \mathrm{range}(M)$, $\Pr[M(D) \in S] \leq e^{\epsilon} \cdot \Pr[M(D') \in S]$.

This definition of differential privacy is for the case where data collectors do not violate the privacy of individuals in the dataset and the data collectors are trusted. If the data collectors are likely to steal personal information, more attention should be paid to the privacy of each individual's data. The concept of considering such localized data is local differential privacy [28]. In contrast to local differential privacy, the concept described above is often referred to as central differential privacy.

Local Differential Privacy In the concept of local differential privacy, we can send data even to untrusted collectors while preserving privacy by adding noise and perturbations to sensitive input values of individuals, rather than to the output data. The definition of ϵ-local differential privacy is as follows:

Definition S2. (*ϵ-Local Differential Privacy [28]*)
A randomized mechanism M is ϵ-local differentially private if, for any input v_i and v_2 and any $y \in \mathrm{range}(M)$, $\Pr[M(v_1) = y] \leq e^{\epsilon} \cdot \Pr[M(v_2) = y]$.

S2.3 Randomized Response

Randomized response protects privacy by perturbing each individual's information, and it was first introduced by Warner [40] to encourage survey participants to answer sensitive questions honestly. This mechanism was shown to be differentially private [15] and has been used widely for crowdsourcing [16] and hypothesis testing [20]. In the following, we describe the randomized response approach in the case where all the participants in a dataset are divided into $m\,(\geq 2)$ mutually exclusive and exhaustive classes.

The randomized response with m classes follows an $m \times m$ distortion matrix:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix},$$

where $p_{uv} = \Pr[x' = u | x = v]$ $(u, v \in \{1, 2, \cdots, m\})$ denotes the probability that the randomized output is u when the real class of the participant is v. Here, the sum of probabilities in each column is 1. When the following inequality holds:

$$e^{\epsilon} \geq \max_{u=1,2,\cdots,m} \frac{\max_{v=1,2,\cdots,m} p_{uv}}{\min_{v=1,2,\cdots,m} p_{uv}},$$

the randomized response satisfies ϵ-differential privacy. To optimize the utility of statistical inferences [25] or maximize the sum of the diagonal elements [39], we should use the following distortion matrix $\mathbf{P}$, s.t.

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^{\epsilon}}{e^{\epsilon}+m-1} & (u = v) \\ \frac{1}{e^{\epsilon}+m-1} & (u \neq v) \end{cases}.$$

Even for the case where there are s (≥ 2) attributes to which the randomized response is applied, Wang *et al.* [39] proposed a method for generating a distortion matrix $\mathbf{P}$ was proposed. They let $\mathbf{P_1}$, $\mathbf{P_2}$, ..., $\mathbf{P_s}$ be distribution matrices to perturb the information about the s attributes, respectively. Then, the distortion matrix $\mathbf{P}$ for the entire data can be given by the following equation:

$$\mathbf{P} = \mathbf{P_1} \otimes \mathbf{P_2} \otimes \cdots \otimes \mathbf{P_s},$$

where $\otimes$ denotes the Kronecker product.

Here, when ϵ_i-differential privacy is achieved by $\mathbf{P_i}$, the randomized response using $\mathbf{P}$ satisfies $\sum_1^s \epsilon_i$-differential privacy. However, this method of generating a distortion matrix is not optimal in terms of privacy level, and we can generate matrices with stronger privacy guarantees. Therefore, in this study, we also propose a more powerful distortion matrix for releasing information with multiple attributes while satisfying differential privacy.

Other mechanisms under local differential privacy include RAPPOR [16] and its variants [24, 38], which encode an input value into a bit vector or bloom filter and apply perturbations to the bit array. These techniques are often used for crowdsourcing statistics because they are also applicable to cases where the inputs are not attribute values and to large-scale datasets.

In this study, we focus on the randomized response technique because the data used to calculate genome statistics can be regarded as attribute data and directly perturbed without encoding to other forms. In fact, in an existing study on private hypothesis testing under local differential privacy [20], the methods using the randomized response were more accurate than RAPPOR-related algorithms.

S3 Detailed Discussion for Case 1 : ϵ for the entire table

S3.1 2×2 Contingency Table

Here, we provide the EM algorithm for the case of a 2×2 contingency table. The details of distortion matrix $\mathbf{P}$ and the method using the inverse matrix are provided in the main document. The following procedure is similar to that in the previous study [18].

i. **Initialization:**
Let N be the number of individuals in the dataset. Create $x \in \mathbb{R}^{2N \times \{a,b,c,d\}}$ s.t.

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

Set $\theta_a^0 = \theta_b^0 = \theta_c^0 = \theta_d^0 = \frac{1}{4}$.

ii. **E-Step:**
For any allele h $(0 \leq h < 2N)$ and $i \in \{a, b, c, d\}$,

$$\begin{aligned} \theta_{h,i}^k = \Pr[z_{h,i} = 1 | x_{h,i}] &= \frac{\Pr[x_{h,i} | z_{h,i} = 1] \cdot \Pr[z_{h,i} = 1]}{\sum_{j \in \{a,b,c,d\}} \Pr[x_{h,i} | z_{h,j} = 1] \cdot \Pr[z_{h,j} = 1]} \\ &= \frac{\Pr[x_{h,i} | z_{h,i} = 1] \cdot \theta_i^{k-1}}{\sum_{j \in \{a,b,c,d\}} \Pr[x_{h,i} | z_{h,j} = 1] \cdot \theta_j^{k-1}}. \end{aligned}$$

iii. **M-Step:**

$$\theta_i^k = \frac{1}{2N} \sum_{h=0}^{2N-1} \theta_{h,i}^k$$

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 $\tilde{a} = 2N \cdot \theta_a^k$, $\tilde{b} = 2N \cdot \theta_b^k$, $\tilde{c} = 2N \cdot \theta_c^k$, and $\tilde{d} = 2N \cdot \theta_d^k$.

Note that $z_{h,j}$ is unobserved data and $\Pr[x_{h,i} | z_{h,j} = 1]$ satisfies the following equations:

$$\Pr[x_{h,i} | z_{h,j} = 1] = \begin{cases} \begin{cases} \frac{e^\epsilon}{e^\epsilon + 3} & (x_{h,i} = 1) \\ \frac{3}{e^\epsilon + 3} & (x_{h,i} = 0) \end{cases} & (i = j) \\ \begin{cases} \frac{1}{e^\epsilon + 3} & (x_{h,i} = 1) \\ \frac{e^\epsilon + 2}{e^\epsilon + 3} & (x_{h,i} = 0) \end{cases} & (i \neq j) \end{cases} .$$

S3.2 3×2 Contingency Table

We consider the case of using the following 3×2 contingency table. This table can be used to conduct the χ^2-test, and the Cochran–Armitage trend test to determine if there is a trend among binomial proportions.

		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

In this case, each individual's genotype can be classified into one of the six attributes from p to u; therefore, we use a 6×6 distortion matrix for randomized response. As in the case of a 2×2 contingency table, the distortion matrix $\mathbf{P}$ can maximize the sum of diagonal elements when the elements are as follows:

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^\epsilon}{e^\epsilon+5} & (u = v) \\ \frac{1}{e^\epsilon+5} & (u \neq v) \end{cases} .$$

The privacy guarantees provided to row and column information by this distortion matrix are $\log\left(\frac{e^\epsilon+1}{2}\right)$ and $\log\left(\frac{e^\epsilon+2}{3}\right)$, respectively.

As in the previous case, we can obtain private p', q', r', s', t', and u' from the randomized response and reconstruct the original table based on these values using either the inverse matrix of $\mathbf{P}$ with the following elements:

$$\mathbf{P}^{-1}_{uv} = \begin{cases} \frac{e^\epsilon+4}{e^\epsilon-1} & (u = v) \\ \frac{-1}{e^\epsilon-1} & (u \neq v) \end{cases} ,$$

or EM algorithm.

When using the inverse matrix, the expected values and variances of $(\tilde{p}, \tilde{q}, \tilde{r}, \tilde{s}, \tilde{t}, \tilde{u})^\mathsf{T} = \mathbf{P}^{-1} \cdot (p', q', r', s', t', u')^\mathsf{T}$ are shown in Theorem 2. The proof is provided in Section S6.

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.*

When using the EM algorithm, the procedure is similar to that for the case of a 2×2 contingency table and is described as follows:

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 $\theta^0_p = \theta^0_q = \theta^0_r = \theta^0_s = \theta^0_t = \theta^0_u = \frac{1}{6}$.

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

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

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, $\Pr[x_{h,i} | z_{h,j} = 1]$ satisfies the following equations:

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

Using the reconstructed values, we can conduct the χ^2-test and the Cochran–Armitage trend test under local differential privacy.

S4 Detailed Discussion for Case 2: ϵ_1 for row and ϵ_2 for column

S4.1 2×2 Contingency Table

We consider the following table:

		Disease Status		Total
		0	1	
Allele	0	a	b	$a+b$
	1	c	d	$c+d$
Total		$a+c$	$b+d$	$2N$

Here, we discuss the first column of the 4×4 distortion matrix $\mathbf{P}$: $(p_{00}, p_{10}, p_{20}, p_{30})^{\mathsf{T}}$, because $\mathbf{P}$ is expected to be a symmetric matrix. Among these four values, the largest is p_{00} and the smallest is p_{30}. Our goal is to maximize the privacy guarantee for the entire table; therefore, we should minimize the value of p_{00}/p_{30}.

First, we let

$$A = \frac{p_{00}}{p_{30}},\ B = \frac{p_{10}}{p_{30}},\ C = \frac{p_{20}}{p_{30}},\ \text{and}\ D = \frac{p_{30}}{p_{30}} = 1.$$

Because the privacy levels of the row and column information are ϵ_1 and ϵ_2, respectively, the values of A, B, C, and D satisfy the following conditions:

$$\begin{cases} A \geq B, C \geq D = 1 \\ \frac{A+B}{C+D} = e^{\epsilon_1} \\ \frac{A+C}{B+D} = e^{\epsilon_2} \end{cases} \iff \begin{cases} A \geq B, C \geq 1 \\ (e^{\epsilon_2}+1)\,B - (e^{\epsilon_1}+1)\,C = e^{\epsilon_1} - e^{\epsilon_2} \quad -\star \\ C = e^{\epsilon_2}\cdot B + e^{\epsilon_2} - A \end{cases}.$$

Since

$$\frac{e^{\epsilon_2}+1}{e^{\epsilon_1}+1} - e^{\epsilon_2} = \frac{1-e^{\epsilon_1+\epsilon_2}}{e^{\epsilon_1}+1} \geq 0\ ,$$

A is minimized when B or C is 1. When $C = 1$, we can obtain $B = (2e^{\epsilon_1} - e^{\epsilon_2} + 1)/(e^{\epsilon_2}+1)$ from $\star$. Thus, we should consider the following two cases:

$$\text{(I)}\ \ \frac{2e^{\epsilon_1} - e^{\epsilon_2} + 1}{e^{\epsilon_2}+1} \geq 1 \iff \epsilon_1 \geq \epsilon_2 \ \ \text{and}\ \ \text{(II)}\ \ \epsilon_1 < \epsilon_2.$$

In Case (I), A is minimized when

$$(A, B, C, D) = \left(\frac{2e^{\epsilon_1+\epsilon_2} + e^{\epsilon_2} - 1}{e^{\epsilon_2}+1}, \frac{2e^{\epsilon_1} - e^{\epsilon_2} + 1}{e^{\epsilon_2}+1}, 1, 1\right).$$

Then, the first column of the distortion matrix $\mathbf{P}$ is as follows:

$$\frac{1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} \begin{pmatrix} 2e^{\epsilon_1+\epsilon_2} + e^{\epsilon_2} - 1 \\ 2e^{\epsilon_1} - e^{\epsilon_2} + 1 \\ e^{\epsilon_2}+1 \\ e^{\epsilon_2}+1 \end{pmatrix}.$$

The second, third, and fourth columns can be obtained in the same manner, and eventually, the elements of $\mathbf{P}$ satisfy the following equations:

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

After the randomized response using this matrix $\mathbf{P}$, we can reconstruct the original table as in the previous section. When using an inverse matrix, the elements of $\mathbf{P}^{-1}$ are as follows:

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

When using an EM algorithm, the procedure is the same as that in Section S3, and $\Pr[x_{h,i}|z_{h,j}=1]$ satisfies the following equations:

$$\Pr[x_{h,i}|z_{h,j}=1] = \begin{cases} \begin{cases} \frac{2e^{\epsilon_1+\epsilon_2}+e^{\epsilon_2}-1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=1) \\ \frac{2e^{\epsilon_1}+e^{\epsilon_2}+3}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=0) \end{cases} & (i=j) \\ \begin{cases} \frac{2e^{\epsilon_1}-e^{\epsilon_2}+1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=1) \\ \frac{2e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_2}+1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=0) \end{cases} & ((i,j) \text{ or } (j,i)=(a,b),(c,d)) \\ \begin{cases} \frac{1}{2(e^{\epsilon_1}+1)} & (x_{h,i}=1) \\ \frac{2e^{\epsilon_1}+1}{2(e^{\epsilon_1}+1)} & (x_{h,i}=0) \end{cases} & (\text{otherwise}) \end{cases}.$$

In Case (II), A is minimized when

$$(A,B,C,D) = \left(\frac{2e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}-1}{e^{\epsilon_1}+1}, 1, \frac{-e^{\epsilon_1}+2e^{\epsilon_2}+1}{e^{\epsilon_1}+1}, 1\right).$$

Then, as in Case (I), the elements of $\mathbf{P}$ and $\mathbf{P}^{-1}$ are as follows:

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

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

When using the EM algorithm, $\Pr[x_{h,i}|z_{h,j}=1]$ satisfies the following equations:

$$\Pr[x_{h,i}|z_{h,j}=1] = \begin{cases} \begin{cases} \frac{2e^{\epsilon_1+\epsilon_2}+e^{\epsilon_1}-1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=1) \\ \frac{e^{\epsilon_1}+2e^{\epsilon_2}+3}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=0) \end{cases} & (i=j) \\ \begin{cases} \frac{-e^{\epsilon_1}+2e^{\epsilon_2}+1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=1) \\ \frac{2e^{\epsilon_1+\epsilon_2}+3e^{\epsilon_1}+1}{2(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)} & (x_{h,i}=0) \end{cases} & ((i,j) \text{ or } (j,i)=(a,c),(b,d)) \\ \begin{cases} \frac{1}{2(e^{\epsilon_2}+1)} & (x_{h,i}=1) \\ \frac{2e^{\epsilon_2}+1}{2(e^{\epsilon_2}+1)} & (x_{h,i}=0) \end{cases} & (\text{otherwise}) \end{cases}.$$

S4.2 3×2 Contingency Table

As in the case of 2×2 table, 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,

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

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:

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

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 $\Pr[x_{h,i}|z_{h,j}=1]$ in the EM algorithm for reconstruction satisfy the following equations:

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

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

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 $\Pr[x_{h,i}|z_{h,j}=1]$ satisfy the following equations:

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

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

In Case (III), P is minimized when

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

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

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

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

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

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 $\Pr[x_{h,i}|z_{h,j}=1]$ satisfy the following equations:

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

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

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

For this case, we illustrate the privacy levels for the entire table when using the existing and our methods in Fig. S1.

Even in the case for a 3×2 contingency table, Fig. S1 indicates that the larger and closer the values of ϵ_1 and ϵ_2 are, the higher the privacy guarantee for the entire data can be achieved by our method.

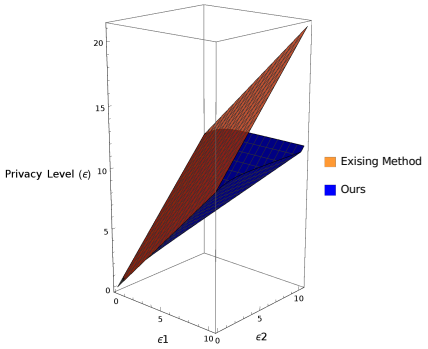

Fig. S1. Comparison of the privacy level for a 3×2 contingency table between the existing method (using the Kronecker-product) and ours. The x and y-axis represent the privacy budget given to row and column information, respectively. The z-axis represents the privacy level for the entire table.

S5 Methods for TDT and EIGENSTRAT

S5.1 Local Differentially Private Methods for TDT

Next, we propose a local differentially private method for a transmission disequilibrium test (TDT), which is the most commonly used method in family-based studies. Note that, following the existing studies [37, 44], we aim to protect the presence information of one family, not that of one individual. In this study, we consider the case of n trio families using the following table:

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

The test statistics is calculated using the values of b and c, and there are six possible pairs (b, c) in one family: $(1, 0)$, $(0, 1)$, $(1, 1)$, $(2, 0)$, $(0, 2)$, and $(0, 0)$. Here, we consider perturbing the information of these six attributes using the randomized response technique. The distortion matrix $\mathbf{P}$ for this case is a 6×6 matrix whose elements are as follows:

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^\epsilon}{e^\epsilon+5} & (u = v) \\ \frac{1}{e^\epsilon+5} & (u \neq v) \end{cases} .$$

After conducting the randomized response based on $\mathbf{P}$, we can reconstruct the original (b, c) attribute using an inverse matrix or EM algorithm, as in the previous case. The EM algorithm for this case is as follows:

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 $(0 \leq h < n)$ 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 $\sum_i \left|\theta_i^k - \theta_i^{k-1}\right| < \delta$ for some $\delta > 0$, then calculate the number of families with each (b, c) attribute value by $n \cdot \theta_i^k$.

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

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

Using the reconstructed number of (b, c) pairs, we can obtain the TDT statistic and conduct a family-based analysis.

Unlike the existing methods [37, 44], our method satisfies the definition of local differential privacy and can be utilized for the cases with untrusted data collectors. In addition, while this study considers protecting the participation information of each family in the dataset, we could conduct the randomized response to each parent's information using a similar approach. Besides, we intend to enhance our technique for use in various other TDT settings [27, 32].

S5.2 Local Differentially Private Methods for EIGENSTRAT

Finally, we propose a method for conducting EIGENSTRAT to correct for population stratification using the randomized response. Our method differs from existing methods [31, 41] in that it can release the statistics while satisfying local differential privacy for both genotype and phenotype information of the targeted single nucleotide polymorphism (SNP) and individual. In the following, we consider obtaining a statistic for a given SNP s and assume that we cannot infer information about the other SNPs from the private statistic.

Let X_{sh} and Y_h be the genotype at SNP s and the phenotype information of individual h, respectively. When $X_{sh} \in \{0, 1, 2\}$ and $Y_h \in \{0, 1\}$ for any h, the number of possible pairs of (X_{sh}, Y_h) is $3 \times 2 = 6$. Here, by perturbing the information of the pair using the randomized response, we can satisfy local differential privacy for both X_{sh} and Y_h. The distortion matrix $\mathbf{P} \in \mathbb{R}^{6\times6}$ is the same as that in the case for TDT, and we let (X'_{sh}, Y'_h) be the perturbed values from (X_{sh}, Y_h). In the statistical tests discussed in the previous cases, we only need to know the number of elements in each cluster (e.g., a, b, c, and d in a 2×2 contingency table), but in EIGENSTRAT, we have to use each individual's (X_{sh}, Y_h) values. It is difficult to recover the individual's data from the perturbed values after the randomized response; therefore, in this case, we conduct the analysis based on (X'_{sh}, Y'_h).

Here, it should be noted that this method requires genotype information at SNPs other than SNP s. Although this study assumes that no information on other SNPs will be leaked from the released statistic, we plan to develop stronger privacy-preserving methods that strictly protect all genotype information in the

future. For other statistical methods for correcting for population stratification, such as LMM-based EMMAX, we can perform the genomic analysis under local differential privacy by considering a set of genotype and phenotype information as in this method. If we aim to assign the privacy budget to each of X_{sh} and Y_h, we should follow procedures similar to our methods for the case-control studies using a 3×2 contingency table.

S6 Proofs

Here, we provide the proofs of theorems on the characteristics of private values in contingency tables.

For the case of 2×2 contingency table, we consider perturbing the values of a, b, c, and d in the following table:

		Disease Status		Total
		0	1	
Allele	0	a	b	$a+b$
	1	c	d	$c+d$
Total		$a+c$	$b+d$	$2N$

Here, we let a', b', c', and d' be the perturbed values of a, b, c, and d, respectively, by the randomized response using the 4×4 distortion matrix $\mathbf{P}$, s.t.

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^\epsilon}{e^\epsilon+3} & (u = v) \\ \frac{1}{e^\epsilon+3} & (u \neq v) \end{cases} .$$

Then, $\tilde{a}$, $\tilde{b}$, $\tilde{c}$, and $\tilde{d}$ can be recovered from a', b', c', and d' using the inverse matrix of $\mathbf{P}$ by the following equation:

$$\left(\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}\right)^\mathsf{T} = \mathbf{P}^{-1} \cdot (a', b', c', d')^\mathsf{T} .$$

We show the expected values and variables of these values by Theorem 1.

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:

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

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

For the case of 3×2 contingency table, we consider perturbing the values of p, q, r, s, t, and u in the following table:

		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

Here, we let p', q', r', s', t', and u' be the perturbed values using the 6×6 distortion matrix $\mathbf{P}$, s.t.

$$\mathbf{P}_{uv} = \begin{cases} \frac{e^\epsilon}{e^\epsilon + 5} & (u = v) \\ \frac{1}{e^\epsilon + 5} & (u \neq v) \end{cases} \ .$$

Then, $\tilde{p}$, $\tilde{q}$, $\tilde{r}$, $\tilde{s}$, $\tilde{t}$, and $\tilde{u}$ can be recovered by the following equation:

$$\left(\tilde{p}, \tilde{q}, \tilde{r}, \tilde{s}, \tilde{t}, \tilde{u}\right)^\mathsf{T} = \mathbf{P}^{-1} \cdot \left(p', q', r', s', t', u'\right)^\mathsf{T} .$$

We show the expected values and variables of these values by Theorem 2.

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,

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

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 manner. □

S7 Experiments and Discussion

In this section, we show the experimental results on accuracy when releasing each genomic statistics by our methods. For χ^2-tests using a contingency table, Cochran–Armitage trend test, and TDT, we compared ours to the existing methods using the Laplace mechanism [19, 37, 45]. For EIGENSTRAT, we employed the existing method satisfying phenotypical differential privacy [31, 41] instead of the Laplace mechanism, because the *sensitivity* has not yet been analyzed.

In the experiments, we randomly generated simulation data on 100 SNPs for statistical analyses using $1,000$ individuals or families. For the family-based TDT, we assigned one (b, c) value among $(1, 0)$, $(0, 1)$, $(1, 1)$, $(2, 0)$, $(0, 2)$, and $(0, 0)$ to each family for each SNP. Then, we calculated the number of families in each category and obtained the TDT statistics for $1,000$ families. For the others, we first generated genotype data at 100 SNPs of $1,000$ individuals and their phenotype data. The genotype value at each SNP and phenotype information was randomly set as 0, 1, or 2, and as 0 or 1, respectively. Based on the generated data, we constructed contingency tables and obtained the statistics for each statistical test. As an example of the distribution of statistics obtained from simulation data, we show the case of the χ^2-test using a 2×2 contingency table in Fig. S2.

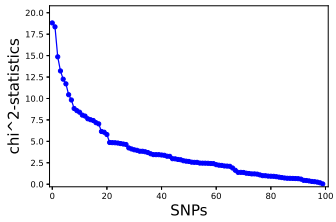

Fig. S2. A distribution of χ^2-statistics obtained from simulation data.

S7.1 χ^2-Tests

Here, we show the results in the case of the χ^2-tests using a contingency table. As a reconstruction procedure after the randomized response, we employed the method using an inverse matrix, because it requires a short execution time, the expected values and variances of the recovered elements of the table are theoretically guaranteed, and the data used in this experiment was relatively large.

To evaluate the accuracy of our methods, we measured the absolute value of the difference between the original and differentially private statistic at each SNP while varying the privacy level ϵ for the entire table.

2×2 Contingency Table In the case of using a 2×2 contingency table, we compared our method to the existing method [45] with the Laplace mechanism. Note that the existing method is for a restricted situation in which the number

of cases and controls are equal under central differential privacy. The results averaged over 20 runs are shown in Fig. S3. SNPs are ordered by the original χ^2-statistic in descending order (cf. Fig. S2).

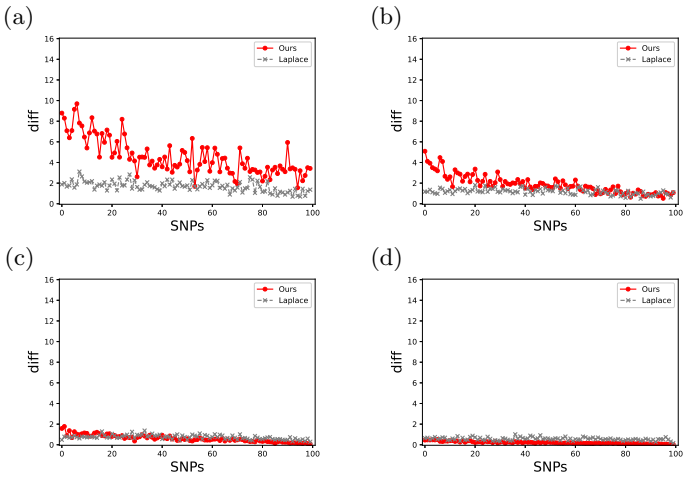

Fig. S3. Differences between original and differentially private χ^2-statistics on 100 SNPs using a 2×2 contingency table when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

Fig. S3 indicates that our method is as accurate as the existing Laplace mechanism even under local differential privacy. Given that local differential privacy can achieve equivalent privacy guarantees to central differential privacy with a larger ϵ [6] and our method requires no restriction on the number of cases and controls, it outperforms the existing method in terms of both privacy assurance and utility. One important characteristic of our method is that the differences become larger when the original χ^2-statistic is larger, whereas the Laplace mechanism does not depend on the statistic values. This might be because the variance of reconstructed elements increases as the original value increases (see Theorem 1). When the statistic is large, an imbalance is expected in the contingency table, that is, some large elements will exist. As a result, the reconstructed table tends to vary greatly, and the private statistic also differ much from the original statistic. In addition, a smaller ϵ causes a more severe increase in the variance, which is more prominent in Cases (a) and (b). For these cases, we plan to consider the use of RAPPOR [16] and its enhancement to improve accuracy.

3×2 Contingency Table In the case of using a 3×2 contingency table, we compared our method to the existing method [19]. As in the case of a 2×2 contingency table, 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. S4.

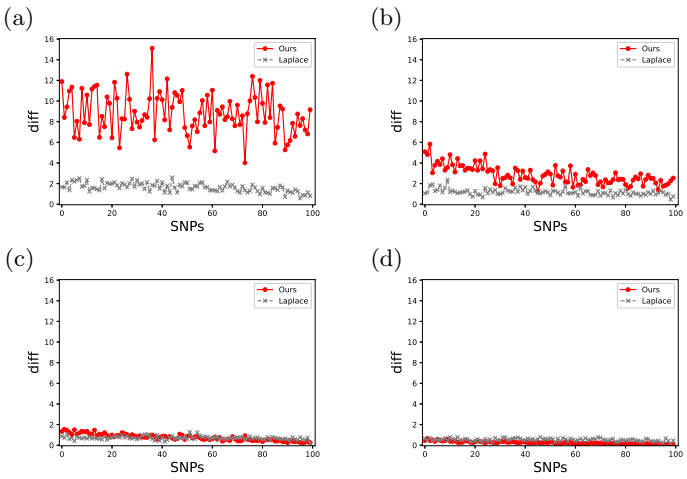

Fig. S4. Differences between original and differentially private χ^2-statistics on 100 SNPs using a 3×2 contingency table when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

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. S4. 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. S3 and Fig. S4.

S7.2 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 [45] as a comparison employs the Laplace mechanism. The results are shown in Fig. S5.

When using the Laplace mechanism, the accuracy becomes worse than the results in Fig. S4 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

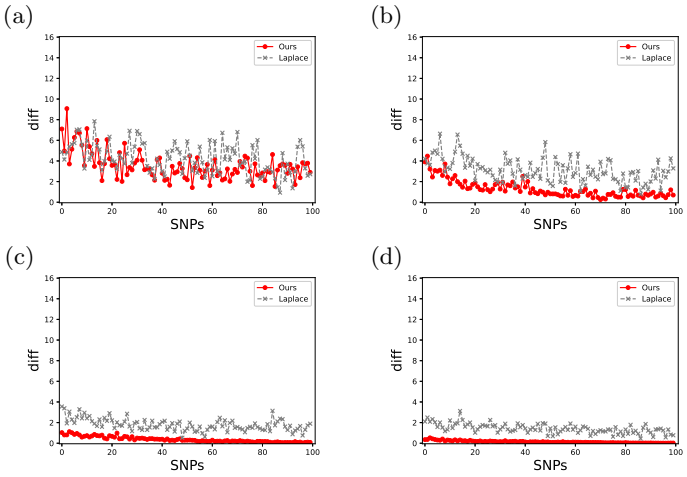

Fig. S5. Differences between original and differentially private χ^2-statistics for the Cochran–Armitage trend test on 100 SNPs when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

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. S5 has the similar trend to the figures in the previous case.

S7.3 TDT

Then, we show the results on the TDT for family-based studies in Fig. S6. Unlike the previous cases, the existing method using the Laplace mechanism [37] 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.

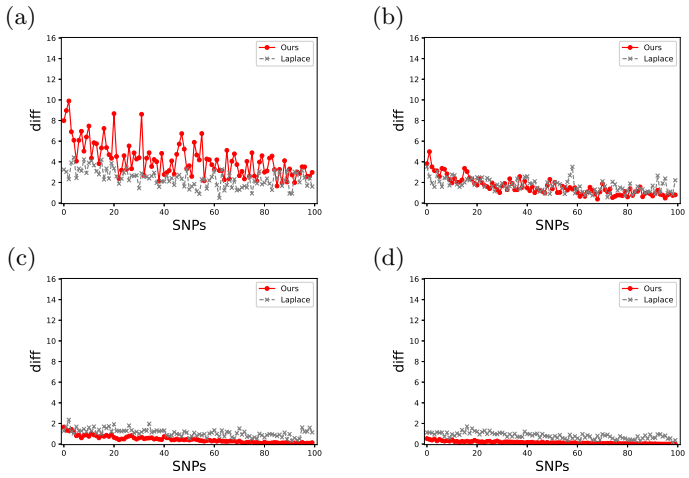

Fig. S6. Differences between original and differentially private TDT statistics on 100 SNPs when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

Fig. S6 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 [6], and the appropriate values of ϵ for genomic statistical analysis under local differential privacy is open to further discussion.

S7.4 EIGENSTRAT

Finally, we show the results on the EIGENSTRAT for correcting for population stratification in Fig. S7.

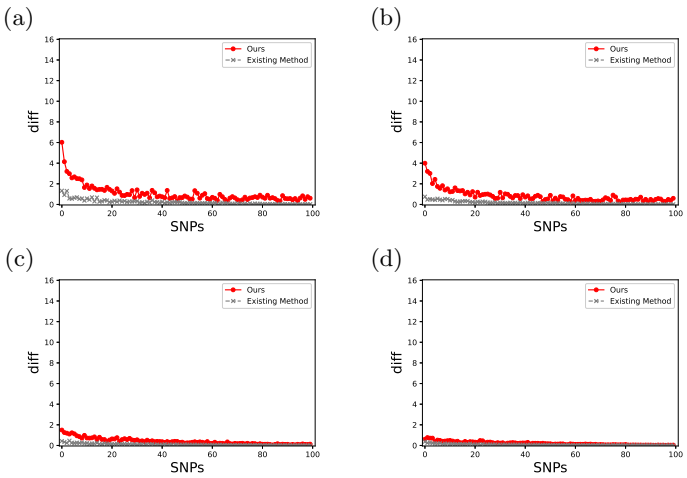

Fig. S7. Differences between original and differentially private EIGENSTRAT statistics on 100 SNPs when (a) $\epsilon = 2$, (b) $\epsilon = 3$, (c) $\epsilon = 5$, and (d) $\epsilon = 7$.

Because the existing method [31, 41] 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. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. p. 308–318 (2016)
2. Almadhoun, N., Ayday, E., Ulusoy, O.: Differential privacy under dependent tuples—the case of genomic privacy. Bioinformatics **36**(6), 1696–1703 (2019)
3. Alnemari, A., Raj, R.K., Romanowski, C.J., Mishra, S.: Interactive range queries for healthcare data under differential privacy. In: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). pp. 228–237 (2021)
4. Armitage, P.: Tests for linear trends in proportions and frequencies. Biometrics **11**(3), 375–386 (1955)
5. Bassily, R., Smith, A.: Local, private, efficient protocols for succinct histograms. In: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing. pp. 127–135 (2015)
6. 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)
7. Blair, G., Imai, K., Zhou, Y.Y.: Design and analysis of the randomized response technique. Journal of the American Statistical Association **110**(511), 1304–1319 (2015)
8. Blatt, M., Gusev, A., Polyakov, Y., Goldwasser, S.: Secure large-scale genome-wide association studies using homomorphic encryption. PNAS **117**(21), 11608–11613 (2020)
9. Bonte, C., Makri, E., Ardeshirdavani, A., Simm, J., Moreau, Y., Vercauteren, F.: Towards practical privacy-preserving genome-wide association study. BMC Bioinform. **19**, 537 (2018)
10. 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)
11. Cho, H., Wu, D.J., Berger, B.: Secure genome-wide association analysis using multiparty computation. Nat. Biotechnol. **36**, 547–551 (2018)
12. Devlin, B., Roeder, K.: Genomic control for association studies. Biometrics **55**(4), 997–1004 (1999)
13. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. S. Halevi and T. Rabin, (eds) Theory of Cryptography **3876**, 265–284 (2006)
14. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) Automata, Languages and Programming. pp. 1–12 (2006)
15. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
16. Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. pp. 1054–1067 (2014)
17. 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)

18. Fanti, G., Pihur, V., Úlfar Erlingsson: Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. Proceedings on Privacy Enhancing Technologies (PoPETS) **issue 3, 2016** (2016)
19. 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)
20. Gaboardi, M., Rogers, R.: Local private hypothesis testing: Chi-square tests. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. vol. 80, pp. 1626–1635 (2018)
21. 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)
22. Holohan, N., Leith, D.J., Mason, O.: Optimal differentially private mechanisms for randomised response. IEEE Transactions on Information Forensics and Security **12**(11), 2726–2735 (2017). https://doi.org/10.1109/TIFS.2017.2718487
23. Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B.C., Roth, A.: Differential privacy: An economic method for choosing epsilon. In: 2014 IEEE 27th Computer Security Foundations Symposium. pp. 398–410 (2014)
24. Kairouz, P., Bonawitz, K., Ramage, D.: Discrete distribution estimation under local privacy. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. pp. 2436–2444 (2016)
25. Kairouz, P., Oh, S., Viswanath, P.: Extremal mechanisms for local differential privacy. J. Mach. Learn. Res. **17**(1), 492–542 (2016)
26. 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)
27. 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)
28. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? In: 2008 49th Annual IEEE Symposium on Foundations of Computer Science. pp. 531–540 (2008)
29. Kockan, C., Zhu, K., Dokmai, N., Karpov, N., Kulekci, M.O., Woodruff, D.P., Sahinalp, S.C.: Sketching algorithms for genomic data analysis and querying in a secure enclave. Nat. Methods **17**, 295–301 (2020)
30. 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)
31. Simmons, S., Sahinalp, C., Berger, B.: Enabling privacy-preserving GWASs in heterogeneous human populations. Cell Syst. **3**(1), 54–61 (2016)
32. 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)
33. 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)
34. Terwilliger, J.D., Ott, J.: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum. Hered. **42**(6), 337–346 (1992)
35. Thomson, G.: Mapping disease genes: family-based association studies. Am. J. Hum. Genet. **57**(2), 487–498 (1995)

36. 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)
37. 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)
38. Wang, T., Blocki, J., Li, N., Jha, S.: Locally differentially private protocols for frequency estimation. In: Proceedings of the 26th USENIX Conference on Security Symposium. p. 729–745 (2017)
39. Wang, Y., Wu, X., Hu, D.: Using randomized response for differential privacy preserving data collection. In: Palpanas, T., Stefanidis, K. (eds.) Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016. vol. 1558 (2016)
40. Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**(309), 63–66 (1965)
41. 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)
42. 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)
43. Yamamoto, A., Shibuya, T.: Differentially private linkage analysis with TDT — the case of two affected children per family. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 765–770 (2021)
44. Yamamoto, A., Shibuya, T.: Efficient differentially private methods for a transmission disequilibrium test in genome wide association studies. In: Pacific Symposium on Biocomputing 2022. pp. 85–96 (2021)
45. Yamamoto, A., Shibuya, T.: More practical differentially private publication of key statistics in GWAS. Bioinformatics Advances **1**(1) (2021)
46. Yu, F., Ji, Z.: 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)
47. Zhang, J., Niyogi, P., McPeek, M.S.: Laplacian eigenfunctions learn population structure. PLoS One **4**(12), e7928 (2009)