Supplements

In Section S1, we show the differentialy private algorithms using the normal Laplace and exponential mechanisms. In Section S2, we give detailed proofs for the theoretical guarantees of our methods. In Section S3, we describe the generation procedures of the simulation data used in our experiments.

S1. Laplace and Exponential Mechanisms

Algorithm S1 ϵ-differentially private algorithm for releasing the top K significant SNPs using the Laplace mechanism. [4]

Input: Genomic statistics data $g \in \mathbb{R}^m$, the *sensitivity* of the statistic Δ_g, number of data to release K, and privacy budget ϵ.

Output: Top K significant SNPs.

1: Add Laplace noise with mean 0 and scale $\frac{2K\Delta_g}{\epsilon}$ to each g_i, and get a noisy vector $\hat{g}$.
2: Choose the top K SNPs based on the elements of $\hat{g}$.

Algorithm S2 ϵ-differentially private algorithm for releasing the top K significant SNPs using the exponential mechanism. [4]

Input: The score of all m SNPs $q \in \mathbb{R}^m$, the *sensitivity* of the score Δ_q, number of data to release K, and privacy budget ϵ.

Output: Top K significant SNPs.

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

S2. Proofs

Theorem 1. *Algorithm 2 satisfies ϵ-differential privacy.*

Proof. Let W be the set of K data obtained by Algorithm 2. In addition, let $g, g' \in \mathbb{R}^m$ be neighboring statistics datasets. Because we assume that ~~g_k cannot be recovered from $\hat{g}_{k'}$ $(k' \neq k)$~~ the sorts in neighboring datasets are indistinguishable based on the probabilities of outputting the same K elements, it is enough to show that the steps 5 – 16 in Algorithm 2 satisfies ϵ-differential privacy. We let $\mathcal{A}$ be that procedure, and we will show $\Pr[\mathcal{A}(g) = W] \leq \mathrm{e}^\epsilon \cdot \Pr[\mathcal{A}(g') = W]$.

Here, we let $\tilde{c} : \mathbb{R} \to \mathbb{R}$ be the function to return noisy statistics from the original input statistics, and f_j be the frequency vector to obtain $\hat{g}_j$ in Algorithm 2. We let the procedure of adding noise to f_j and getting a noisy statistic (Steps 6 – 14) be a function $\tilde{d} : \mathbb{R}^m \to \mathbb{R}$

Also, we suppose that the real statistics of the released SNPs are $S_1, S_2, \cdots, S_K$ and the noisy statistics of them are $v_1, v_2, \cdots, v_K$, and we define $\mathcal{S} = \{S_1, S_2, \cdots, S_K\}$.

In addition, we let the $S_1, S_2, \cdots, S_K$'s indices in $\hat{g}$ be $I_1, I_2, \cdots, I_K$. Here, we assume that the sorting data in Step 1 cannot be identified from the output data (i.e., the sorts in g and g' are indistinguishable), so we consider these indices are equal in g and g'.

Under the above conditions, the following equation holds:

$$
\begin{aligned}
&\Pr[\mathcal{A}(g) = W] \\
= &\int_{v_1}\int_{v_2}\cdots\int_{v_K} \\
&\Pr[\tilde{c}(S_1) = v_1] \cdot \Pr[\tilde{c}(S_2) = v_2] \cdots \Pr[\tilde{c}(S_K) = v_K] \\
&\cdot \prod_{X \in g - \mathcal{S}} \Pr[\tilde{c}(X) < \min\{v_1, \cdots, v_K\}] \, \mathrm{d}v_1 \mathrm{d}v_2 \cdots \mathrm{d}v_K \\
= &\int_{v_1}\int_{v_2}\cdots\int_{v_K} \\
&\Pr[\tilde{d}(f_{I_1}) = v_1] \cdot \Pr[\tilde{d}(f_{I_2}) = v_2] \cdots \Pr[\tilde{d}(f_{I_K}) = v_K] \\
&\cdot \prod_{X \in g - \mathcal{S}} \Pr[\tilde{c}(X) < \min\{v_1, \cdots, v_K\}] \, \mathrm{d}v_1 \mathrm{d}v_2 \cdots \mathrm{d}v_K
\end{aligned}
$$

In the similar way to the discussion of [1]'s Theorem 4, we can find the lower bound of $\Pr[\mathcal{A}(g') = W]$ when g is fixed by considering a "fictional" data g' such that

$$
\tilde{c}(g'_i) = \begin{cases} \tilde{c}(g_i) - 2\Delta_g & (i \in W) \\ \tilde{c}(g_i). & (i \notin W) \end{cases}
$$

Here, when the original statistic changes d, regardless of the index of the statistic, F_k^{real} and F_k^{imag} of the frequency vector F change by $d \cdot \cos\left(-\frac{2jk}{m}\pi\right)$ and $d \cdot \sin\left(-\frac{2jk}{m}\pi\right)$, respectively. Therefore, the following inequality holds:

$$
\log \frac{\Pr[\tilde{d}(f_{I_i}) = v_i]}{\Pr[\tilde{d}(f'_{I_i}) = v_i]} \leq \frac{2\Delta_g}{\lambda},
$$

where f'_j is the frequency vector in the calculation of $\mathcal{A}(g')$, and λ is the scale of the Laplace distribution in Step 6. Consequently, we can show

$$
\log \frac{\Pr[\mathcal{A}(g) = W]}{\Pr[\mathcal{A}(g') = W]} \leq \frac{2K\Delta_g}{\lambda}. \tag{1}
$$

Thus, if we set $\lambda = \frac{2K\Delta_g}{\epsilon}$, (1) is equal to ϵ, and Algorithm 2 satisfies ϵ-differential privacy. □

Theorem 2. *Let Δ_{X_k} be the sensitivity of X_k in a vector $X \in \mathbb{R}^m$ and define the operations in steps 2 through 5 of*

Algorithm 3 as a function $f : \mathbb{R}^m \longrightarrow \mathbb{C}^m$. *The sensitivity of* X'_k *in the vector* $X' = IDFT(f(X))$ *is* $\frac{2s-1}{m} \cdot \Delta_{X_k}$.

Proof. Let $F = f(X)$. When X_k changes by d, the amount of change in F_j $(j = 0, 1, \cdots, s-1, m-s+1, m-s+2, \cdots, m-1)$ is

$$\begin{aligned} & d \cdot \left(\cos\left(-\frac{2jk}{m}\pi \right) + i \cdot \sin\left(-\frac{2jk}{m}\pi \right) \right) \\ = \; & d \cdot \left(\cos\left(\frac{2jk}{m}\pi \right) - i \cdot \sin\left(\frac{2jk}{m}\pi \right) \right) \end{aligned}$$

Here, since

$$\begin{cases} X'_k = \frac{1}{m} \sum_{j=0}^{m-1} \left(\cos\left(\frac{2jk}{m}\pi \right) + i \cdot \sin\left(\frac{2jk}{m}\pi \right) \right) \cdot F_j, \\ F_j = 0 \quad (j = s, s+1, \cdots, m-s), \end{cases}$$

the amount of change in $|X'_k|$ is

$$\begin{aligned} & \left| \frac{d}{m} \cdot \left\{ \sum_{j<s, m-s<j} \left(\cos\left(\frac{2jk}{m}\pi \right) + i \cdot \sin\left(\frac{2jk}{m}\pi \right) \right) \right. \right. \\ & \qquad \left. \left. \cdot \left(\cos\left(\frac{2jk}{m}\pi \right) - i \cdot \sin\left(\frac{2jk}{m}\pi \right) \right) \right\} \right| \\ = \; & \frac{d}{m} \cdot (2s-1). \end{aligned}$$

Therefore, the *sensitivity* of X'_k is $\frac{2s-1}{m} \cdot \Delta_{X_k}$. □

S3. Simulation Data

A. Good Value of s

TABLE S1
2×2 CONTINGENCY TABLE FOR A TRANSMISSION DISEQUILIBRIUM TEST IN ONE SNP FOR TRIO FAMILIES.

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

In this experiment, we set the number of families in the dataset to $n = 2,000$, and generate the values of b and c in Table 1 for each SNP by the following equation:

$$S = Random(0, 2n), \quad b = Binomial(S, 0.5), \quad c = S - b,$$

where $Random(0, 2n)$ is a random integer between 0 and $2n$, and $Binomial(S, 0.5)$ is the number of successes after S trials with the success probability of 0.5. In addition, for 10 significant SNPs, we set the probability in the binomial distribution to generate b to 0.55. This is because the statistics for significant SNPs are often much larger than the others in large-scale genetic analyses such as GWAS.

TABLE S2
NUMBER OF FAMILIES FOR EACH (b, c).

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

B. Small Cohort

We set the family number $N = 150$ and SNP number $M = 5,000$ as in the experiments by [3]. Here, we explain how to generate a family dataset for the i-th SNP. Note that the possible combinations of (b, c) in one family are shown in Table S2, and b and c in n families can be calculated by the following equations: $b = n_1 + n_3 + 2\,n_4$ and $c = n_2 + n_3 + 2\,n_5$.

First, we let S_i be a random natural number in the range of 0 to $2N$. Then, we generate n_1 from binomial distribution with size S_i and probability 0.5. Finally, we set $n_2 = S_i - n_1$ and $n_6 = 2N - n_1 - n_2$. In addition, for the 10 SNPs, the probability in the binomial distribution to generate n_1 is set to 0.75 to create some datasets for significant SNPs. The distribution of the statistics is shown in Fig. S1.

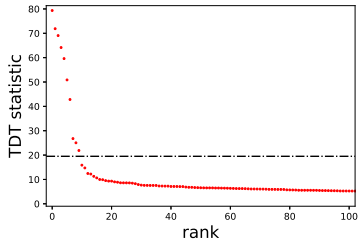

Fig. S1. The top 100 TDT statistics in simulation data for a small cohort. The dotted line is the threshold at $100(1-0.05/m)\%$-quantile of χ^2-distribution with one degree of freedom, based on the Bonferroni correction.

C. Large Cohort

We set $N = 5,000$ and $M = 10^6$ as in the experiments by [3]. The way to generate non-significant datasets is the same as in a small cohort. When generating 10 significant datasets, the probability in the binomial distribution to calculate n_1 is set to 0.56. The distribution of the statistics is shown in Fig. S2.

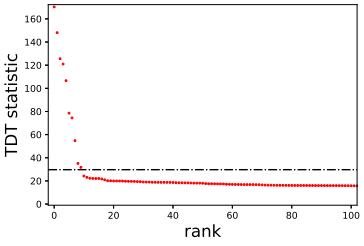

Fig. S2. The top 100 TDT statistics in simulation data for a large cohort. The dotted line is the threshold at $100(1-0.05/m)\%$-quantile of χ^2-distribution with one degree of freedom, based on the Bonferroni correction.

D. Real Data

We generated family datasets based on the TDT data on nonsyndromic metopic craniosynostosis by [2]. The data contains 215 families and $649,669$ SNPs, and 6 SNPs were tested to be significant. Based on this data, we generated a dataset containing $10,000$ statistics. The detailed procedure is explained below. First, we prepared the TDT statistics for all SNPs according to their Q-Q plot. Next, we find b and c such that they yield each statistic. Then, we determine the values from n_1 to n_6 using random numbers so that the following two equations are satisfied:$b = n_1 + n_3 + 2\,n_4, c = n_2 + n_3 + 2\,n_5$.

REFERENCES

[1] Bhaskar, R., Laxman, S., Smith, A., Thakurta, A.: Discovering frequent patterns in sensitive data. In: KDD'10. pp. 503–512. Washington, DC, USA (Jul 2010)

[2] Justice, C.M., Cuellar, A., Bala, K., Sabourin, J.A., Cunningham, M.L., Crawford, K., Phipps, J.M., Zhou, Y., Cilliers, D., Byren, J.C., Johnson, D., Wall, S.A., Morton, J.E.V., Noons, P., Sweeney, E., Weber, A., Rees, K.E.M., Wilson, L.C., Simeonov, E., Kaneva, R., Yaneva, N., Georgiev, K., Bussarsky, A., Senders, C., Zwienenberg, M., Boggan, J., Roscioli, T., Tamburrini, G., Barba, M., Conway, K., Sheffield, V.C., Brody, L., Mills, J.L., Kay, D., Sicko, R.J., Langlois, P.H., Tittle, R.K., Botto, L.D., Jenkins, M.M., LaSalle, J.M., Lattanzi, W., Wilkie, A.O.M., Wilson, A.F., Romitti, P.A., Boyadjiev, S.A.: A genome-wide association study implicates the BMP7 locus as a risk factor for nonsyndromic metopic craniosynostosis. Hum. Genet. **139**(8), 1077–1090 (2020)

[3] 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)

[4] 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)