Supplementary Materials

In Section S1, we present preliminary definitions for this study. In Section S2, we give detailed proofs for the theoretical guarantees of our methods. In Section S3, we provide descriptions of datasets used in the experiments. In Section S4, we present an algorithm for releasing the top K TDT statistics by adopting the Laplace mechanism.

S1. Preliminaries

Definition S1. *(ϵ-Differential Privacy)*
A randomized mechanism M is ϵ-differentially private if, for all datasets D and D' which differ in only one family and any $S \subset$range(M),

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

Definition S2. *(Sensitivity for the Exponential Mechanism)*
Let $\mathcal{D}^M$ be the collection of all datasets with M marker loci, the sensitivity of a score function $u : \mathcal{D}^M \times \{1, 2, \ldots, M\} \to \mathbb{R}$ is

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

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

Definition S3. *(The SHD score)*
Given a predefined threshold $c^* > 0$, the SHD score for i-th data D_i $(i = 1, 2, \ldots, M)$ is

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

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

The above definitions relate to the concepts used to release the top K significant marker loci privately. Furthermore, the Laplace mechanism[1] might be applied to release those statistics based on the concept of differential privacy.[2] In the Laplace mechanism, the *sensitivity* of a statistical function is considered. The definition of the *sensitivity* is as follows.

Definition S4. *(Sensitivity for the Laplace Mechanism)*
Let $\mathcal{D}^M$ be the collection of all datasets with M marker loci, the sensitivity of a function $f : \mathcal{D}^M \to \mathbb{R}^d$ is

$$\Delta f = \max_{D, D'} \|f(D) - f(D')\|_1,$$

where $D, D' \in \mathcal{D}^M$ differ in a single family.

For a statistic $f(D)$ obtained from the original dataset D, releasing $f(D) + b$ satisfies ϵ-differential privacy when b is random noise derived from a Laplace distribution with mean 0 and scale $\frac{\Delta f}{\epsilon}$.[1]

S2. Proofs

Table S1. 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

S2.1. *Exact Algorithm*

Theorem S1. *Algorithm 1 outputs the exact SHD score.*

Proof. We consider two cases: (I) $T < c^*$ and (II) $T \geq c^*$.

(I) $T < c^*$

In order to increase the value of the statistic $(b-c)^2/b+c$, we can consider increasing the difference between the values of b and c.

Firstly, we consider making b larger than c. Here, we discuss how to change the families included in each of the categories shown in Table S1. We start by looking at the case of changing one family in the category $(1,0)$. In this case, there are five possible changes as follows: (i) $(1,0) \to (0,1)$, (ii) $(1,0) \to (1,1)$, (iii) $(1,0) \to (2,0)$, (iv) $(1,0) \to (0,2)$, and (v) $(1,0) \to (0,0)$. For each of these cases, the statistics after the change are given below:

$$\text{(i) } \frac{(b-c-2)^2}{b+c},\ \text{(ii) } \frac{(b-c-1)^2}{b+c+1},\ \text{(iii) } \frac{(b-c+1)^2}{b+c+1},$$
$$\text{(iv) } \frac{(b-c-3)^2}{b+c+1},\ \text{(v) } \frac{(b-c-1)^2}{b+c-1}.$$

If $b > c$, the largest change is in case (iii), so we change the family into the category $(2,0)$. When a family is in the categories $(0,1)$, $(1,1)$, $(0,2)$, and $(0,0)$, we can change them into the category $(2,0)$ as well. The families in the category $(2,0)$ are not changed, because changing them decrease the statistics. Then, since

$$\frac{(b-c+4)^2}{b+c} > \frac{(b-c+3)^2}{b+c+1} > \frac{(b-c+2)^2}{b+c} > \frac{(b-c+2)^2}{b+c+2} > \frac{(b-c+1)^2}{b+c+1},$$

we can check n_5, n_2, n_3, n_6, and n_1 in that order and increase n_4, which is the number of families with $(2,0)$.

When making b smaller than c, the proof is very similar to the above.

(II) $T \geq c^*$

When $b > c$, we can think as in the case (I) and change the families so that the statistic becomes smaller. In this case, we consider increasing the number of families included in the category $(0,2)$. Since

$$\frac{(b-c-4)^2}{b+c} < \frac{(b-c-3)^2}{b+c+1} < \frac{(b-c-2)^2}{b+c+2} < \frac{(b-c-2)^2}{b+c} < \frac{(b-c-1)^2}{b+c+1},$$

we check n_4, n_1, n_6, n_3, and n_2 in that order.

When $b < c$, same as for the case of $b > c$. □

Algorithm S1 Exact algorithm to find the SHD Score for TDT statistics.

Input: Information about a single marker locus, i.e., n_1, n_2, n_3, n_4, n_5, n_6, and the threshold c^* for the TDT statistics.

Output: The SHD score in one marker locus.

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

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

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

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

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

S2.2. *Approximation Algorithm*

Theorem S2. *The sensitivity of the SHD score obtained by Algorithm 2 is 1.*

Proof.

(I) $(b - c)^2/(b + c) < c^*$

(i) $b + c < c^*$

When $b \geq c$, $(b+c)+|b-c| = 2b$. Since the maximum change in b is 2, that in the SHD score is $\lceil \frac{4}{4} \rceil = 1$. It is similar for $b < c$.

(ii) $b + c \geq c^*$

Let $b + c = s$, $|b - c| = d$, and we calculate the maximum change in $\sqrt{kc^*} - s$.

When the change in s is 2, d changes by at most 2. Therefore, we can consider the following inequality:

$$\begin{aligned}\{\sqrt{(s+2)c^*} - (d-2)\} - \{\sqrt{sc^*} - d\} &= \frac{2c^*}{\sqrt{(s+2)c^*} + \sqrt{sc^*}} + 2 \\ &\leq \frac{\sqrt{c^*}}{\sqrt{s}} + 2 \leq 3. \qquad [\because s \geq c^*]\end{aligned}$$

When the change in s is 1, since d changes by at most 3,

$$\begin{aligned}\{\sqrt{(s+1)c^*} - (d-3)\} - \{\sqrt{sc^*} - d\} &= \frac{c^*}{\sqrt{(s+1)c^*} + \sqrt{sc^*}} + 3 \\ &\leq \frac{\sqrt{c^*}}{2\sqrt{s}} + 3 \leq \frac{7}{2}. \qquad [\because s \geq c^*]\end{aligned}$$

When s does not change, the maximum change in d is 4.

Thus, the SHD score changes by at most $\lceil \frac{4}{4} \rceil = 1$.

(II) $(b-c)^2/(b+c) \geq c^*$

Same as the case (I)(ii).

Consequently, the sensitivity of the SHD score from Algorithm 2 is 1. □

Algorithm S2 Approximation algorithm to find the SHD Score for TDT statistics.

Input: Information about a single marker locus, i.e., n_1, n_2, n_3, n_4, n_5, n_6, and the threshold c^* for the TDT statistics.

Output: The SHD score in one marker locus.

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

S3. Experiments

S3.1. *Simulation Data*

For both cases in (I) small cohort and (II) large cohort, we generated simulation data for two situations: (i) all families were in the n_1 or n_2 or n_6 categories, and (ii) families were distributed in n_1 to n_6 categories. We show the distribution of the statistics for each simulation data in Fig S1. As we can see from this figure, about 10 of the top marker loci are significant.

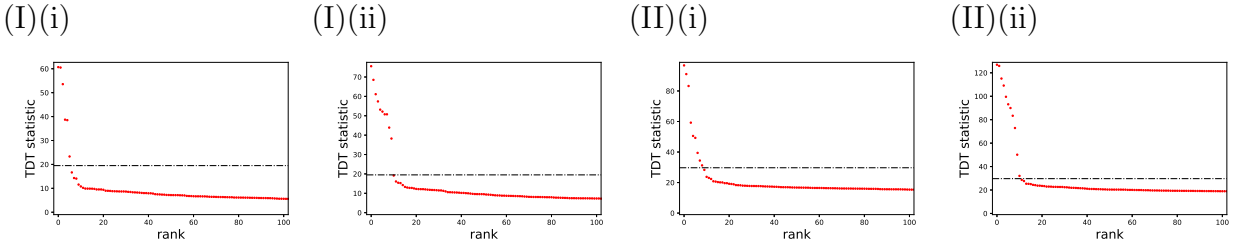

Fig. S1. The top 100 TDT statistics in the simulation data for each case. The dotted lines are thresholds at $100(1-0.05/M)\%$-quantile of χ^2-distribution with one degree of freedom, based on the Bonferroni correction.

S3.2. *Real Data*

We generated family datasets based on the TDT data on nonsyndromic metopic craniosynostosis by Justice et al. 2020.[3] The data contains 215 families and 649669 SNPs, and 6 SNPs were tested to be significant. Here, we explain how to generate simulation data based on this data. 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:

$$b = n_1 + n_3 + 2\,n_4$$
$$c = n_2 + n_3 + 2\,n_5.$$

The distribution of TDT statistics in the datasets generated bby the above procedure is shown in Fig S2.

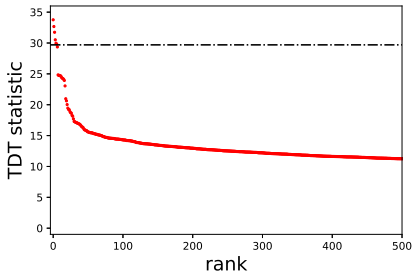

Fig. S2. The top 500 TDT statistics based on real data. The dotted lines are thresholds at $100(1-5\times10^{-8})\%$-quantile of χ^2-distribution with one degree of freedom.

S4. Algorithm for Releasing Top K TDT Statistics

Similar to methods by Bhaskar et al.,[4] we consider combining the Laplace mechanism with our method for releasing the top K significant marker loci. Then, the TDT statistics for the top K marker loci can be released while achieving ϵ-differential privacy, and the procedure is summarized in Algorithm S3.

Algorithm S3 ϵ-differentially private algorithm for releasing TDT statistics of the top K significant marker loci using the exponential mechanism with the SHD score and the Laplace mechanism.

Input: The SHD score of all m marker loci, the number K of marker loci to release, the privacy budget ϵ, and the sensitivity of the TDT statistics s.

Output: TDT statistics of the top K significant marker loci.

1: Let $S = \emptyset$ and q_i be the SHD score of i-th marker locus.
2: For each $i \in \{1, \ldots, m\}$, set the weight $w_i = \exp\left(\frac{\epsilon q_i}{4K}\right)$ and the probability $p_i = \frac{w_i}{\sum_{i=1}^{m} w_i}$ for sampling i-th marker locus.
3: Sample k from $\{1, \ldots, m\}$ with probabilities $\{p_1, \ldots, p_m\}$, add k-th marker locus to S and set $q_k = -\infty$.
4: Repeat step 2 and 3 until the size of S reaches K.
5: Add a Laplace noise with mean 0 and scale $\frac{2Ks}{\epsilon}$ for the top K marker loci.

References

1. C. Dwork, F. McSherry, K. Nissim and A. Smith, Calibrating noise to sensitivity in private data analysis, *S. Halevi and T. Rabin, (eds) Theory of Cryptography* **3876**, 265 (2006).
2. C. Dwork, Differential privacy, *Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, (eds) Automata, Languages and Programming* **4052** (2006).
3. C. M. Justice, A. Cuellar, K. Bala, J. A. Sabourin, M. L. Cunningham, K. Crawford, J. M. Phipps, Y. Zhou, D. Cilliers, J. C. Byren, D. Johnson, S. A. Wall, J. E. V. Morton, P. Noons, E. Sweeney, A. Weber, K. E. M. Rees, L. C. Wilson, E. Simeonov, R. Kaneva, N. Yaneva, K. Georgiev, A. Bussarsky, C. Senders, M. Zwienenberg, J. Boggan, T. Roscioli, G. Tamburrini, M. Barba, K. Conway, V. C. Sheffield, L. Brody, J. L. Mills, D. Kay, R. J. Sicko, P. H. Langlois, R. K. Tittle, L. D. Botto, M. M. Jenkins, J. M. LaSalle, W. Lattanzi, A. O. M. Wilkie, A. F. Wilson, P. A. Romitti and S. A. Boyadjiev, A genome-wide association study implicates the BMP7 locus as a risk factor for nonsyndromic metopic craniosynostosis, *Hum. Genet.* **139**, 1077 (2020).
4. R. Bhaskar, S. Laxman, A. Smith and A. Thakurta, Discovering frequent patterns in sensitive data, in *KDD'10*, (Washington, DC, USA, 2010).