In this supplementary material, we overview the transmission disequilibrium test (TDT) focused on in our experiments and give a detailed description of the used data.

S1 Preliminaries

S1.1 TDT

TDT (Spielman *et al.*, 1993) is a test for linkage disequilibrium, which examines the relationship between a disease and two or more alleles. In TDT for n trio families, we consider $2n$ parents and n affected children. In this study, we focus on the case of testing for 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 S1.

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	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:

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

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 S2, and b and c in n families can be calculated by the following equations: $b = n_1 + n_3 + 2\,n_4$ and $c = n_2 + n_3 + 2\,n_5$.

Table S2. 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 Experiments

S2.1 Simulation Data

S2.1.1 Small Cohort

We set the family number $N = 150$ and SNP number $M = 5,000$ as in the experiments by Wang *et al.*, 2017. Here, we explain how to generate a family dataset for the i-th SNP.

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.

S2.1.2 Large Cohort

We set $N = 5,000$ and $M = 10^6$ as in the experiments by Wang *et al.*, 2017. 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.

S2.2 Real Data

We generated family datasets based on the TDT data on nonsyndromic metopic craniosynostosis by Justice *et al.*, 2020. 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:

$$\begin{aligned} b &= n_1 + n_3 + 2\,n_4 \\ c &= n_2 + n_3 + 2\,n_5. \end{aligned}$$

The distribution of TDT statistics in the datasets generated by the above procedure is shown in Fig S1.

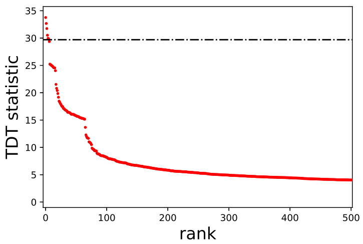

Fig. S1. The top 500 TDT statistics based on real data. The dotted line is the threshold set by Justice et al., 2020.

References

Justice, C. M. *et al.* (2020). A genome-wide association study implicates the BMP7 locus as a risk factor for nonsyndromic metopic craniosynostosis. *Hum. Genet.*, **139**(8), 1077–1090.

Spielman, R. S. *et al.* (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**(3), 506–516.

Wang, M. *et al.* (2017). Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies. *Bioinformatics*, **33**(23), 3716–3725.