JUSTC

Article

http://justc.ustc.edu.cn

Received: November 03, 2023; Accepted: December 11, 2023

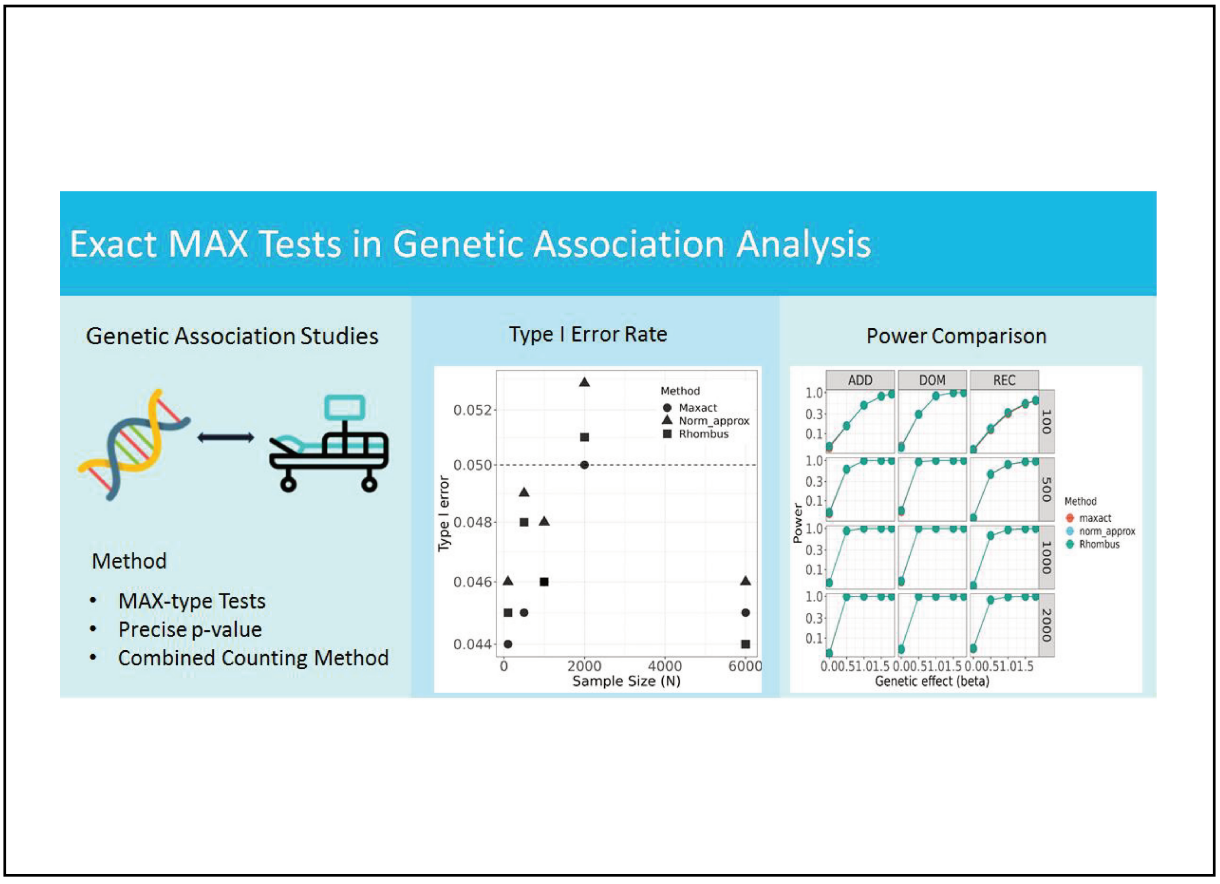
Exact MAX tests in genetic association analysis

Yi Li[1], Jianan Tian[1], Chenliang Xu[2], Yaning Yang[1], and Min Yuan[3] ✉

[1]*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China;*
[3]*Department of Health Data Science, School of Health Management, Anhui Medical University, Hefei 230032, China*

✉Correspondence: Min Yuan, E-mail: myuan@ustc.edu.cn

Graphical abstract



Type I error rate and power comparison of the proposed method MaXact, normal approximation, and Rhombus methods under various scenarios.

Public summary

- A precise and efficient algorithm is presented to compute the *p* value for the MAX-type test statistic utilizing a combined counting method.
- The suggested approach effectively manages the type I error rate and attains power comparable to that of the permutation method.
- The proposed method significantly reduces computational complexity, even with relatively large sample sizes.

Citation: Li Y, Tian J N, Xu C L, et al. Exact MAX tests in genetic association analysis. *JUSTC*, **2024**, 54(12): 1205. CSTR: 32290.14.JUSTC-2023-0153. DOI: 10.52396/JUSTC-2023-0153

JUSTC

Article

http://justc.ustc.edu.cn

Received: November 03, 2023; Accepted: December 11, 2023

Exact MAX tests in genetic association analysis

Yi Li[1], Jianan Tian[1], Chenliang Xu[2], Yaning Yang[1], and Min Yuan[3] ✉

[1]*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China;*
[3]*Department of Health Data Science, School of Health Management, Anhui Medical University, Hefei 230032, China*

✉Correspondence: Min Yuan, E-mail: myuan@ustc.edu.cn

Cite This: *JUSTC*, **2024**, 54(12): 1205 (6pp)

Read Online

Abstract: The three common genetic models (or modes of inheritance) in association analysis are the dominant, additive, and recessive models. It is known that the Cochran-Armitage trend test (CATT) which correctly incorporates information from genetic models, is more powerful than the commonly used Pearson's chi-square test. However, the true genetic model is usually unknown in practice, and the power of the CAT test could be substantially reduced with a wrongly specified genetic model. To achieve a power that is close to that of a correctly specified CAT test, it is natural to apply trend tests under different possible genetic models and to report the most significant test result. This results in a MAX-type testing procedure, and it was found that this test is usually more powerful than the Pearson's chi-square test. Although the significance (i.e., p value) of the MAX-type test can be accessed by either large sample approximation or permutation methods, requirements for sample size or simulation replicates are demanding with respect to accuracy and efficiency. This paper proposes an approach to calculate the exact p values of MAX-type tests based on the combinatorial counting method. The simulation results show that the exact method is more accurate than the large sample approximation methods and more computationally efficient than the permutation method, and our method can be readily applied to genome-wide association studies (GWASs). The proposed method is built in an R package, MaXact, which is available at the https://github.com/Myuan2019/MaXact/.

Keywords: exact test; MAX test; normal approximation; permutation

CLC number: R311 **Document code:** A

2020 Mathematics Subject Classification: 62P10

1 Introduction

The case-control design is widely used in genetic association studies, particularly for large scale studies. To test the association between the disease and the candidate marker using case-control data, several tests can be used, including Pearson's chi-square test, and Cochran-Armitage's linear trend test (CATT)[1–3]. When the underlying genetic model is known, the corresponding CATT is the optimal test in terms of power. However, when the underlying genetic model is unknown, CATT with the incorrect genetic model will suffer substantial power loss, especially when the dominant model is misspecified as the recessive model and vice versa[4]. In practice, the underlying genetic model is rarely known and tests that are robust to model uncertainty are desirable. One of the robust tests is Pearson's chi-square test. An alternative process is to apply CATT under different genetic models and use the most significant model as the final result. Such approaches include MAX2, MAX3, and other MAX-type tests. MAX3 takes the maximum of the dominant, additive and recessive CATT or, equivalently, the minimum of the p values of the three tests. It has been shown that MAX3 is more powerful than Pearson's chi-square test and is close to the optimal CATT test with a correct genetic model[5–8].

As a robust test, MAX3 has been widely employed in genome-wide association studies (GWASs). The MAX3 test has been used to examine the effect of SNPs within functional IncRNAs on the occurrence and development of hepatocellular carcinoma (HCC) in studies conducted by Zhang et al.[9], and hepatitis B virus-related hepatocellular carcinoma (HBV-related HCC) in research by Liu et al[10]. Furthermore, the MAX3 test has been applied to explore the relationships among the ADIPO pathway[11], the LEP signaling pathway[12], and type 2 diabetes mellitus (T2DM) in the Chinese Han population. Additionally, several extensions of the MAX3 test have been proposed to accommodate a wider range of scenarios. The MAX3 test has been extended to transmission disequilibrium test (TDT) by Zheng et al.[13] and Joo et al[14]. So and Sham[15] devised an approach based on Monte-Carlo simulation for the computation of the covariance matrix of the asymptotic distribution of MAX3, and derived the p value through trivariate integration, accommodating quantitative or binary traits as well as covariates. Zang et al.[16] expanded the applicability of MAX3 by adapting it to assess gene-environment interactions. Chen and Zang[17] proposed CMAX3 (covariate-adjusted MAX3), an extension of MAX3 that integrates the score test derived from logistic regression to account for the covariate effect.

CSTR: 32290.14.JUSTC-2023-0153
DOI: 10.52396/JUSTC-2023-0153
JUSTC, **2024**, 54(12): 1205

In association analysis, the null distribution of MAX3 is needed to determine the critical value and calculate the p value. However, it is difficult to derive the analytical form of the distribution of MAX3 under the null hypothesis of no association. An asymptotic distribution was derived to approximate the null distribution of MAX3. González et al.[18] studied the normal approximation methods for the MAX3 test, and obtained the p value of MAX3 from a trivariate normal approximation. Li et al.[19] gave a formula called 'Rhombus' to obtain an approximation of the p value of the MAX3 test. However, the accuracy of the approximation depends on the sample size, and how large the sample size must be for the approximation to be satisfactorily accurate is unknown in most cases. Moreover, asymptotic theory is not always available for every test statistic, especially in cases where the regularity conditions for classical theory are not satisfied.The permutation method can provide an accurate approximation if the number of permutation replicates is large enough. However, genome-wide association studies usually involve scanning millions of SNPs simultaneously. The inflation of the Type Ⅰ error rate due to multiple testing procedures should be appropriately considered. Correction approaches, such as Bonferroni correction, can be applied to correct for the inflation of the type I error rate. For genome-wide association studies with 100,000 to 500,000 SNPs, the significance level for a single SNP is approximately $1\mathrm{E}-07$ after correction, while the total type I error rate is controlled at 0.05. Therefore, the permutation method requires intensive simulation replicates to achieve such a small significance level. An efficient method for computing the exact p value of MAX tests is still missing.

In this paper, we develop an efficient algorithm, MaXact, for calculating the exact p value of the MAX3 test. The MaXact algorithm uses a combinatorial counting method to calculate the exact p value and is similar to the algorithm proposed by Boulesteix[20] for collapsing contingency tables. Through simulation studies and real data analysis, we show that the MaXact method is comparable to the existing approximation methods in terms of computing efficiency but provides exact p value instead of approximations. On the other hand, the MaXact method is much more computationally efficient than the permutation method. The remainder of this paper is organized as follows. The MaXact method is described in the next section. The results of simulation and real data analysis are reported in Section 3.1 and 3.2. Some discussion remarks are provided in Section 3.3. Section 4 provides the concluding remarks.

2 Materials and methods

Suppose that the candidate gene of interest has two different alleles denoted by A and a, which can form three possible genotypes AA, Aa, and aa. There are R cases and S controls sampled from the population. The total sample size is N. Genotype counts for cases/controls are denoted as r_0/s_0, r_1/s_1, and r_2/s_2 for AA, Aa, and aa, respectively. The genotype counts for cases and controls are summarized in Table 1.

CATT proposed by William Cochran and Peter Armitage, is widely used to detect the association between column and row variables in contingency table analysis. CATT modifies the Pearson chi-squared test by incorporating information of the possible ordering of genetic effects. For example, an individual's risk of disease increases with the number of risk alleles. Therefore, genotypes AA, Aa, and aa can be ordered as "low", "medium", and "high" with respect to the risk of disease if allele a is assumed to be the risk allele. Scores in increasing order can be assigned to these three categories. CATT with scores $\boldsymbol{x}=(x_1,x_2,x_3)$ for genotypes (AA,Aa,aa) has a general form[3],

Table 1. Genotype counts for cases and controls.

	AA	Aa	aa	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

$$Z_{\boldsymbol{x}}=\frac{N^{1/2}\sum_{i=0}^{2}x_i(Sr_i-Rs_i)}{\left(RS\left(N\sum_{i=0}^{2}x_i^2n_i-\left(\sum_{i=0}^{2}x_in_i\right)^2\right)\right)^{1/2}},$$

Since CATT is invariant under a linear transformation of scores $\boldsymbol{x}$, we can confine scores in the unit interval, i.e., $\boldsymbol{x}=(0,\theta,1)$ with $0\leqslant\theta\leqslant1$[8]. θ corresponds to the genetic model or the mode of inheritance. Note that in Table 1, $N=R+S$, $n_i=r_i+s_i$ for $i=0,1,2$, we can rewrite the CATT indexed by θ as

$$Z_\theta=\frac{\frac{\theta n_1+n_2}{N}-\frac{\theta s_1+s_2}{S}}{\left(\frac{R}{NS}\left(\frac{\theta^2n_1+n_2}{N}-\left(\frac{\theta n_1+n_2}{N}\right)^2\right)\right)^{1/2}}. \tag{1}$$

The optimal choices of score θ for the recessive, additive, and dominant models are $\theta=0$, 0.5, and 1, respectively[3], and Z_θ has an asymptotically standard normal distribution under the null hypothesis that there is no association between the marker and the disease. When the genetic model is unknown, one tends to test the null hypothesis of no association by applying Z_θ for $\theta=0,0.5$, and 1 and report the most significant one. This results in the (two-sided) MAX3 test defined as follows.

$$Z_{\max3}=\max(|Z_0|,|Z_{0.5}|,|Z_1|),$$

The p value of the test statistic $Z_{\max3}$ conditional on R,S,n_0,n_1, and n_2 can be calculated as the probability of obtaining a more extreme observation under the null hypothesis. Let p_3 denote this probability, then,

$$p_3=P_{H_0}(Z_{\max3}\geqslant d)=1-P_{H_0}(Z_{\max3}<d)=$$
$$1-P_{H_0}(|Z_0|<d,|Z_{0.5}|<d,|Z_1|<d),$$

where d is the observed value (realization) of $Z_{\max}$ for a given data set. The results from Freidlin et al.[4] and Zheng et al.[8] showed that when the underlying model is unknown, the MAX3 test is robust with respect to model specification and efficient in terms of power.

Calculation of p value of $Z_{\max3}$ needs to know its distribu-

CSTR: 32290.14.JUSTC-2023-0153
DOI: 10.52396/JUSTC-2023-0153
JUSTC, **2024**, 54(12): 1205

tion under the null hypothesis. However, the null distribution of $Z_{\max 3}$ is completely unknown. Several methods have been proposed to approximate the null distribution of $Z_{\max 3}$ using large sample theory (González et al.[18] and Li et al.[19]). The normal approximation methods of González et al.[18] and Li et al.[19] are based on the fact that when the null hypothesis is true and the sample size is large enough, the joint distribution of the trend test statistics $\boldsymbol{Z}=(Z_0,Z_{0.5},Z_1)$ can be approximated by the normal distribution $\mathrm{N}_3(0,\boldsymbol{\Sigma})$, where the 3×3 correlation matrix $\boldsymbol{\Sigma}$ can be estimated from the data (see Refs.[18] and [8] for details). Then, the normally approximated p value of González et al.[18] is given by

$$\hat{p}_{3,\mathrm{norm}}=\int_{-d}^{d}\int_{-d}^{d}\int_{-d}^{d}\phi(z;\mathbf{0},\boldsymbol{\Sigma})\mathrm{d}z,$$

where $\phi(\boldsymbol{Z};\mathbf{0},\boldsymbol{\Sigma})$ is the probability density function of $N_3(\mathbf{0},\boldsymbol{\Sigma})$. Li et al.[19] gave a formula called 'Rhombus' to approximate $\hat{p}_{3,\mathrm{norm}}$, which is denoted as $\hat{p}_{3,\mathrm{rhom}}$. On the other hand, the permutation method can be used to calculate the p values of $Z_{\max 3}$. By switching the labels "case" and "control", $Z_{\max 3}$ for each new data set can be calculated and the percentage of new $Z_{\max 3}$s that are greater than the observed value of $Z_{\max 3}$ can be calculated, which is exactly the p value of $Z_{\max 3}$. The p value from the permutation method is denoted by $\hat{p}_{3,\mathrm{perm}}$.

In what follows, we describe our method, MaXact, for calculating the exact p value of the two-sided MAX3 test. Conditioning on the total counts of cases and controls (row sums), R and S, as well as the total counts of all genotypes (column sums), n_0, n_1, and n_2, the trend test statistic Z_θ in (1) and $Z_{\max 3}$ depends only on variables (s_1,s_2). First, note that the counts (s_1,s_2) in Table 1 satisfy the following natural constraint:

$$\begin{aligned}\Omega_\infty=\{&(s_1,s_2)\mid s_0+s_1+s_2=\\&S,\ 0\leqslant s_i\leqslant S,\ 0\leqslant n_i-s_i\leqslant R,\ \text{for } i=0,1,2\}.\end{aligned}\tag{2}$$

On the other hand, from (1), the event $\{Z_{\max 3}<d\}$ is equivalent to

$$\Omega_d=\{(s_1,s_2)\in\Omega_\infty\mid L_d(\theta)<\theta s_1+s_2<U_d(\theta),\ \text{for } \theta=0,0.5,1\}.\tag{3}$$

where

$$\begin{aligned}L_d(\theta)&=\frac{(\theta n_1+n_2)S}{N}-d\left(\frac{RS}{N}\left[\frac{\theta^2 n_1+n_2}{N}-\left(\frac{\theta n_1+n_2}{N}\right)^2\right]\right)^{1/2},\\U_d(\theta)&=\frac{(\theta n_1+n_2)S}{N}+d\left(\frac{RS}{N}\left[\frac{\theta^2 n_1+n_2}{N}-\left(\frac{\theta n_1+n_2}{N}\right)^2\right]\right)^{1/2}.\end{aligned}$$

Under the null hypothesis, each possible configuration (s_1,s_2) in Table 1 has a probability

$$P(s_1,s_2)=\frac{\binom{n_0}{S-s_1-s_2}\binom{n_1}{s_1}\binom{n_2}{s_2}}{\binom{N}{S}},\ (s_1,s_2)\in\Omega_\infty,\tag{4}$$

as a result of enumerating all possible and equiprobable paths. Then, the p value of the MAX3 test can be computed by

$$p_3=1-\sum_{(s_1,s_2)\in\Omega_d}P(s_1,s_2).$$

We remark that if the hypothesis is one-sided, then the one-sided MAX3 test is taken to be $Z'_{\max 3}=\max\{Z_0,Z_{0.5},Z_1\}$, and the exact p value can be calculated similarly by setting $L_d(\theta)=-\infty$. Another MAX-type test that has been used in the literature, MAX2, is defined as $Z_{\max 2}=\max(|Z_0|,|Z_1|)$. Its p value can be calculated similarly. One only needs to change Ω_d in (3) to

$$\Omega'_d=\{(s_1,s_2)\in\Omega_\infty\mid L_d(\theta)<\theta s_1+s_2<U_d(\theta),\ \text{for } \theta=0,1\},$$

where the subscript d is the realization of the statistic $Z_{\max}$.

3 Results and discussion

3.1 Simulation

In this section, we examine the MaXact method for the MAX3 test via simulation studies. Specifically, we illustrate the superiority of MaXact over the normal approximation methods of González et al.[18] and Li et al.[19] in terms of computational accuracy and its superiority over the permutation method in terms of computational efficiency.

We simulate case-control data under the null hypothesis and assume that Hardy-Weinberg equilibrium (HWE) holds in the population. When the HWE assumption is violated, the results are similar and thus not reported here. We assume that the frequency of allele a in the population is randomly sampled from a uniform distribution $U(0.1,0.5)$ and the number of cases and controls are equal ($R=S=N/2$). We simulate for various sample sizes N=100, 500, 1000, 2000, and 6000 and various numbers of permutation replicates $n=10^3$, 5×10^3, 10^4, 5×10^4, 10^5 for the permutation method, and perform $n_{\mathrm{sim}}=1000$ simulations for each scenario. We use the average Euclidean distance, denoted by $D(x,p_3)$, to measure the discrepancy between the exact p value p_3 and the other three methods, where x can be $\hat{p}_{3,\mathrm{norm}}$ or $\hat{p}_{3,\mathrm{rhom}}$, or $\hat{p}_{3,\mathrm{perm}}$.

The results under the null hypothesis are listed in Table 2. Table 2 shows that the accuracy measurement $D(\hat{p}_{3,\mathrm{perm}},p_3)$ decreases with the number of permutations n for the permutation method. The permutation method can produce p values arbitrarily close to the exact p values as long as the number of permutation replicates is large enough. On the other hand, $D(\hat{p}_{3,\mathrm{norm}},p_3)$ for the normal approximation method decreases as sample size N increases, which is justified by the fact that the approximation is more accurate when the sample size is larger. However, even if the sample size reaches 6000, $D(\hat{p}_{3,\mathrm{norm}},p_3)$ is still 8.43×10^{-3} according to Table 2, which

Table 2. Accuracy of the permutation method and approximation methods (up to 10^{-2}) for calculating the p value of MAX3 based on $n_{\mathrm{sim}}=1000$ replicates of simulations.

	Permutation replicates					Normal	Rhombus
N	10^3	5×10^3	10^4	5×10^4	10^5	approximation	approximation
100	1.229	0.538	0.392	0.179	0.121	8.438	9.048
500	1.294	0.563	0.392	0.180	0.124	3.232	10.587
1000	1.313	0.570	0.381	0.183	0.127	2.256	10.971
2000	1.315	0.558	0.411	0.184	0.129	1.577	11.366
6000	1.301	0.598	0.428	0.184	0.130	0.843	11.748

CSTR: 32290.14.JUSTC-2023-0153
DOI: 10.52396/JUSTC-2023-0153
JUSTC, **2024**, 54(12): 1205

may not be tolerable in a genome-wide association study. For the Rhombus approximation method, $D(\hat{p}_{3,\text{rhom}}, p_3)$ is larger than that of the normal approximation method, and it does not decrease with sample size since the Rhombus formula only gives an upper bound of the true p value.

We then compare the computing expenses of the four methods using the statistical software R implemented on a personal computer (Intel(R) Xeon(R) CPU X5355 @ 2.66GHz). All four methods are applied to the same data set generated under the null hypothesis. The results in Table 3 show that the MaXact method is the most computationally efficient when the sample size is smaller than 6000, and is slightly less computationally efficient than the Rhombus method and more efficient than the normal approximation method when the sample size reaches 6000. Each replicate of the permutation method takes less computing time than the other three methods, but since the permutation method usually runs thousands or more replicates of permutation, the permutation method is the least efficient among all the methods. We can also see that the computational efficiency of the MaXact method decreases with sample size N, but even when the sample size is $N = 6000$, its computational efficiency is still comparable to that of the Rhombus method (MaXact takes 5.06×10^{-4} s, and the Rhombus method takes 3.69×10^{-4} s).

From Table 3, for a genome-wide association study with sample size 2000 and 1 million SNPs, assuming that almost all SNP markers are unlinked with the disease under study, the exact method takes approximately 4.52 min for single-SNP screening using the MAX3 test, which is more computationally efficient than using the normal approximation method or the Rhombus method.

Fig. 1 and Fig. 2 compare Type Ⅰ error and power under different genetic models across varying sample sizes and genetic effects. In general, Maxact demonstrates superior control over Type Ⅰ error at the nominal level while maintaining comparable power to the approximation-based method. Additionally, we conducted supplementary simulations to assess the influence of case-control ratios on the results. The results suggest that our Maxact method maintains robustness across a range of minor allele frequencies (MAFs) and case-control ratios.

3.2 A real example

In a genome-wide association study, strong signals (small p values) are of central importance. To compare MaXact with other methods in calculating small p values, we analyzed the top 28 significant SNPs detected by Sladek et al.[21] in a study of type 2 diabetes mellitus (T2DM). They used the same two-sided MAX3 statistic, T_{MAX3}, as we did, but had to run the very time-consuming permutations to obtain the corresponding p values. The SNPs with sufficiently small p values ($p \leqslant 1 \times 10^{-4}$) are listed in Table 4. We applied the MaXact method and other methods mentioned above to these 28 SNPs. Since the true p values are very small, we performed the permutation method with a large number of replicates, and the results show that the p values obtained from the permutation method are very close to those computed from the MaXact method. In addition, the normal approximation tends to produce smaller p values, while the Rhombus approximation produces larger ones. Computation time of p values for these 28 SNPs was approximately 0.009 s for the MaXact method, 0.087 s for the normal approximation method and 0.011 s for

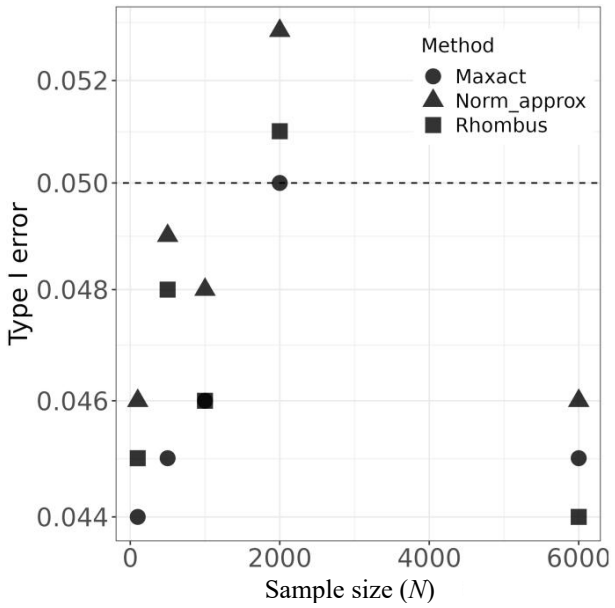


Fig. 1. Type Ⅰ error rates of the MaXact, normal approximation, and Rhombus methods under various scenarios.

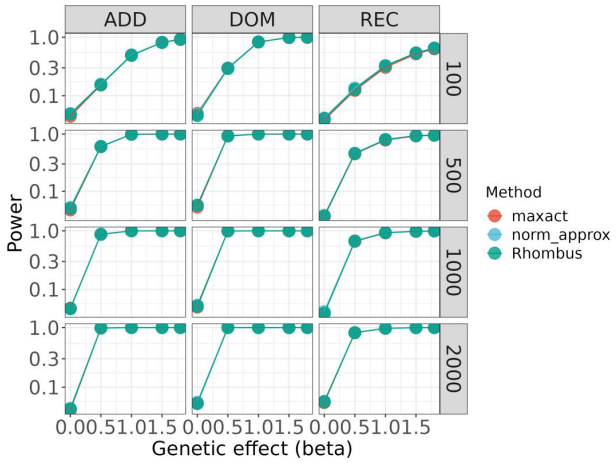


Fig. 2. Powers of the MaXact, normal approximation, and Rhombus methods under various scenarios.

Table 3. Average time cost (in 10^{-4} s*) of the MaXact method, permutation method, and approximation methods in calculating the p value of MAX3.

N	MaXact	Normal approximation	Rhombus approximation	Permutation ($n = 1$)
100	1.82	41.38	3.94	1.10
500	1.96	41.79	3.71	1.46
1000	2.08	40.02	3.45	2.01
2000	2.71	40.19	3.73	2.89
6000	5.06	41.97	3.69	7.38

*Software: R; Computer: Intel(R) Xeon(R) CPU X5355 @ 2.66GHz.

CSTR: 32290.14.JUSTC-2023-0153
DOI: 10.52396/JUSTC-2023-0153
JUSTC, **2024**, 54(12): 1205

Table 4. *P* values of the most significant SNPs in the T2DM association study.

SNP	r_0	r_1	r_2	s_0	s_1	s_2	MaXact	Permutation ($n \geqslant 10^8$)*	Normal approximation	Rhombus approximation
rs7900150	129	326	229	198	325	143	1.3×10^{-8}	1.3×10^{-8}	8.9×10^{-9}	1.4×10^{-8}
rs7100927	129	328	229	198	326	143	1.3×10^{-8}	1.4×10^{-8}	9.0×10^{-9}	1.4×10^{-8}
rs1193179	340	288	58	423	202	44	1.0×10^{-6}	9.8×10^{-7}	7.5×10^{-7}	1.0×10^{-6}
rs932206	134	285	267	158	333	178	3.7×10^{-6}	3.7×10^{-6}	2.6×10^{-6}	3.7×10^{-6}
rs1978717	300	308	75	364	260	36	4.0×10^{-6}	4.0×10^{-6}	3.0×10^{-6}	4.0×10^{-6}
rs11084127	300	311	75	363	266	36	5.3×10^{-6}	5.2×10^{-6}	4.5×10^{-6}	6.1×10^{-6}
rs1111875	77	298	310	122	316	231	7.8×10^{-6}	8.0×10^{-6}	5.2×10^{-6}	7.1×10^{-6}
rs11084128	299	302	76	363	264	36	6.2×10^{-6}	6.3×10^{-6}	5.4×10^{-6}	7.2×10^{-6}
rs282705	24	239	423	60	264	345	6.3×10^{-6}	6.2×10^{-6}	5.5×10^{-6}	7.3×10^{-6}
rs1836002	300	311	75	364	268	37	9.2×10^{-6}	9.1×10^{-6}	6.8×10^{-6}	9.1×10^{-6}
rs3740878	25	273	386	65	249	353	1.8×10^{-5}	1.8×10^{-5}	1.4×10^{-5}	1.8×10^{-5}
rs11037909	25	274	387	65	251	353	1.8×10^{-5}	1.8×10^{-5}	1.4×10^{-5}	1.9×10^{-5}
rs8101509	303	297	80	344	285	33	2.1×10^{-5}	2.1×10^{-5}	1.5×10^{-5}	2.0×10^{-5}
rs2499953	646	39	1	660	9	0	6.3×10^{-6}	6.2×10^{-6}	1.9×10^{-5}	2.2×10^{-5}
rs6670163	34	204	448	45	266	358	2.0×10^{-5}	2.0×10^{-5}	1.9×10^{-5}	2.5×10^{-5}
rs945384	614	69	3	640	28	1	2.0×10^{-5}	2.1×10^{-5}	3.2×10^{-5}	3.8×10^{-5}
rs1113132	25	271	390	63	251	355	3.9×10^{-5}	3.9×10^{-5}	3.2×10^{-5}	4.1×10^{-5}
rs2278419	319	294	69	368	270	27	4.3×10^{-5}	4.1×10^{-5}	3.1×10^{-5}	4.0×10^{-5}
rs7651936	156	326	204	186	351	131	4.4×10^{-5}	4.4×10^{-5}	3.1×10^{-5}	4.1×10^{-5}
rs10211998	26	189	466	37	249	380	3.9×10^{-5}	4.1×10^{-5}	3.1×10^{-5}	4.0×10^{-5}
rs5756371	28	194	462	41	251	376	4.0×10^{-5}	3.9×10^{-5}	4.0×10^{-5}	5.1×10^{-5}
rs13064991	15	177	494	27	233	409	2.8×10^{-5}	2.8×10^{-5}	3.1×10^{-5}	3.9×10^{-5}
rs1256517	471	184	17	527	116	15	5.2×10^{-5}	5.1×10^{-5}	5.0×10^{-5}	6.2×10^{-5}
rs6541240	83	253	350	101	303	265	6.3×10^{-5}	6.2×10^{-5}	4.8×10^{-5}	6.2×10^{-5}
rs6413504	126	340	215	185	313	163	7.0×10^{-5}	7.1×10^{-5}	5.1×10^{-5}	6.8×10^{-5}
rs2050831	36	184	466	29	258	381	7.5×10^{-5}	7.4×10^{-5}	6.8×10^{-5}	8.7×10^{-5}
rs873492	270	321	95	337	256	76	1.2×10^{-4}	1.2×10^{-4}	9.0×10^{-5}	1.1×10^{-4}
rs11078674	297	304	83	354	267	46	8.0×10^{-5}	7.9×10^{-5}	5.6×10^{-5}	7.3×10^{-5}

*$n = 10^{10}$ when $p_3 < 10^{-7}$; $n = 10^9$ when $10^{-7} \leqslant p_3 < 10^{-5}$; $n = 10^8$ when $p_3 \geqslant 10^{-5}$.

the Rhombus approximation. In contrast, the permutation method with 10^9 replicates took approximately 76 d on a single PC computer (8 d on 10 computers).

3.3 Discussion

In this paper, we proposed a simple but efficient procedure, MaXact, to compute the exact *p* value of a MAX test often used in genetic association studies. We focused on MAX3 in this study, and similar procedures could be implemented to obtain exact *p* values of other MAX-type tests. Our method was built in the R package "MaXact". The advantage of our method is that it calculates the exact *p* value of a MAX-type test efficiently. We have shown that even when the sample size is very large, the normal approximations may still have errors that cannot be ignored. On the other hand, the permutation method is accurate as long as the number of permutation replicates is large enough, which is computationally intensive and not feasible in GWASs. Our method outperforms the normal approximation since it calculates the exact *p* value of a MAX-type test with similar computing expenses. The computational efficiency of MaXact also greatly exceeds that of the permutation method, although the permutation method can produce results very close to the exact *p* values. Since highly significant markers are of particular interest in genome-wide association studies, we recommend using MaXact, instead of normal approximation and permutation methods, to obtain exact *p* values.

The incorporation of MaXact into real-world genetic association studies offers significant advantages for researchers seeking meaningful insights from complex genetic data, particularly in the context of large-scale genomic investigations. In genome-wide association studies, where numerous genetic markers are simultaneously evaluated, the efficiency of MaXact becomes a valuable asset, ensuring rapid and accurate analyses crucial for timely and meaningful results. Moreover, MaXact's provision of a reliable and efficient tool contributes to the progress of genetic research. Its ability to obtain accurate *p* values without compromising computational effi-

CSTR: 32290.14.JUSTC-2023-0153
DOI: 10.52396/JUSTC-2023-0153
JUSTC, **2024**, 54(12): 1205

ciency aligns with the dynamic landscape of genomics, where precision and speed are paramount.

Zang et al.[16] extended the MAX3 method to accommodate covariates and gene-environment interactions, exploring the asymptotic distribution of the generalized MAX3 test under the null hypothesis. Their approach involves incorporating the genetic model into logistic regression to derive the MAX3 test. In contrast, the MaXact method introduced in this article is a combination-based approach utilizing contingency table data representing gene models and disease states. While this method can be extended to categorical covariates, it is not applicable to continuous covariates.

4 Conclusions

In conclusion, the practical application of MaXact elevates it from a methodological advancement to a valuable resource for researchers navigating the intricacies of genetic association studies. The precision, efficiency, and adaptability of MaXact position it as a tool capable of shaping more robust and reliable findings, exerting a positive and impactful influence on the trajectory of genetic research.

Acknowledgements

This work was supported by the Natural Science Foundation of Anhui Province (2008085MA09) and the National Natural Science Foundation of China (11671375).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yi Li is currently a graduate student in the Department of Statistics and Finance, University of Science and Technology of China, under the supervision of Prof. Yaning Yang. Her research mainly focuses on genetic association analysis and longitudinal data analysis.

Min Yuan is currently a Professor of Statistics at Anhui Medical University. She received her Ph.D degree from the University of Science and Technology of China in 2009. Her research interests include genome-wide association studies on Alzheimer's disease, longitudinal data analysis, and statistical modeling and applications in public health and biomedicine.

References

[1] Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics*, **1955**, *11* (3): 375–386.
[2] Cochran W G. Some methods for strengthening the common chi-square tests. *Biometrics*, **1954**, *10*: 417–451.
[3] Sasieni P D. From genotypes to genes: Doubling the sample size. *Biometrics*, **1997**, *53* (4): 1253–1261.
[4] Freidlin B, Zheng G, Li Z H, et al. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity*, **2002**, *53* (3): 146–152.
[5] Balding D J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **2006**, *7* (10): 781–791.
[6] Zang Y, Fung W K, Zheng G. Tail strength to combine two p values: Their correlation cannot be ignored. *American Journal of Human Genetics*, **2009**, *84* (2): 291–295.
[7] Zang Y, Fung W K, Zheng G. Simple algorithms to calculate asymptotic null distributions of robust tests in case-control genetic association studies in R. *Journal of Statistical Software*, **2010**, *33* (8): 1–24.
[8] Zheng G, Freidlin B, Gastwirth J L. Comparison of robust tests for genetic association using case-control studies. *IMS Lecture Notes: Monograph Series*, **2006**, *49*: 253–265.
[9] Zhang J G, Liu L, Lin Z F, et al. SNP-SNP and SNP-environment interactions of potentially functional *HOTAIR* SNPs modify the risk of hepatocellular carcinoma. *Mol Carcinog*, **2019**, *58* (5): 633–642.
[10] Liu Q, Liu G Y, Lin Z F, et al. The association of lncRNA SNPs and SNPs - environment interactions based on GWAS with HBV - related HCC risk and progression. *Molecular Genetics & Genomic Medicine*, **2021**, *9* (2): e1585.
[11] Yu H B, Hu W, Lin C W, et al. Polymorphisms analysis for association between ADIPO signaling pathway and genetic susceptibility to T2DM in Chinese han population. *Adipocyte*, **2021**, *10* (1): 463–474.
[12] Yu H B, Xu L, Liu H, et al. Association analysis of LEP signaling pathway with type 2 diabetes mellitus in Chinese Han population from South China. *BioMed Research International*, **2021**, *2021*: 5517364.
[13] Zheng G, Freidlin B, Gastwirth J L. Robust TDT-type candidate-gene association tests. *Annals of Human Genetics*, **2002**, *66* (2): 145–155.
[14] Joo J, Kwak M, Chen Z, et al. Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. *Statistics in Medicine*, **2010**, *29* (1): 158–180.
[15] So H C, Sham P C. Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. *Behavior Genetics*, **2011**, *41* (5): 768–775.
[16] Zang Y, Fung W K, Cao S, et al. Robust tests for gene–environment interaction in case-control and case-only designs. *Computational Statistics & Data Analysis*, **2019**, *129*: 79–92.
[17] Chen Z, Zang Y. CMAX3: A robust statistical test for genetic association accounting for covariates. *Genes*, **2021**, *12* (11): 1723.
[18] González J R, Carrasco J L, Dudbridge F, et al. Maximizing association statistics over genetic models. *Genetic Epidemiology*, **2008**, *32* (3): 246–254.
[19] Li Q, Zheng G, Li Z, et al. Efficient approximation of p-value of maximum of corrected test with application to genome-wide association studies. *Annals of Human Genetics*, **2008**, *72* (3): 397–406.
[20] Boulesteix A L. Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal*, **2006**, *48* (3): 451–462.
[21] Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **2007**, *445*: 881–885.

CSTR: 32290.14.JUSTC-2023-0153
DOI: 10.52396/JUSTC-2023-0153
JUSTC, **2024**, 54(12): 1205