Tail Strength to Combine Two p Values: Their Correlation Cannot Be Ignored

To the Editor: The population-based case-control study is a useful approach to evaluating genetic association with many common and complex diseases. In general, one first uses the generalized linear model to fit the data and then uses an asymptotic test to detect the true association. In addition to this regression-based analysis, when Hardy-Weinberg equilibrium (HWE) holds in the population, testing HWE in cases has been used for indicating the association. Because the regression-based analyses (including the trend test and the likelihood-ratio test) are generally more powerful than testing HWE in cases, they are often employed in case-control studies. Less attention is paid to testing HWE in cases.

In the July 2008 issue of *The American Journal of Human Genetics*, Wang and Shete[1] proposed a novel approach of using the tail strength to combine the p value of the likelihood-ratio test (LRT) for association and the p value of an exact test for the deviation from HWE in cases. Taylor and Tibshirani[2] originally proposed the tail strength as a measure of the overall strength of association for a large number of hypotheses in microarray analyses and genome-wide association studies (GWAS). Compared to Fisher's combination of p values[3], which weights each p value equally, the tail strength weights each ordered p value by its expectation under the null hypothesis. The tail strength can be used for combining independent and dependent p values and is not restricted to any special genetic model underlying the data. Wang and Shete[1] combined the two p values by using the tail strength and extended the original tail strength by using the medians of the ordered p values as weights. They derived asymptotic null distributions for the tail strengths by applying the additive model and using the mean and median as weights, respectively. Their results showed significant improvement in terms of the power when the tail strengths were used. They also showed that the type I errors were under control, although we notice that almost all reported type I errors in their tables are less than the nominal levels.

Normally, when the tail strength is used as a test statistic, its asymptotic null distribution is approximated by Monte-Carlo simulation procedures. Simulation-based approaches to determining the tail probabilities or p values of complex statistics have limitations for applications in GWAS.[4,5] In this situation, deriving their asymptotic distributions is important. Although Wang and Shete[1] derived the asymptotic null distributions and critical values for their tail-strength statistics, they assumed in their derivations that the two p values were *independent* even though in the introduction section they mentioned that they would use the tail strength to combine two *dependent* p values. When the two p values are correlated, their asymptotic null distributions may be inappropriate. Using two test statistics different from those in Wang and Shete,[1] Zheng and Ng[6] noticed that the correlation between the p values of the trend test and testing HWE between cases and controls (HWDTT[7]) could also vary from the recessive (REC) model to the additive (ADD) model, the multiplicative (MUL) model, and the dominant (DOM) model. As we mentioned before, Wang and Shete[1] considered the tail strengths based only on the ADD model. However, the performance of testing HWE in cases would also vary across the genetic models. For example, it is known that testing HWE cannot detect association under the MUL model even though testing HWE has been used for detecting association.[8–10]

Therefore, the performance of the tail strength of Wang and Shete[1] can be potentially affected by two factors that were either ignored or not examined in their article. One is the correlation between the two p values of the LRT and the test for HWE in cases, and the other is the unknown underlying genetic models. In this letter, using Monte-Carlo simulation procedures, we study the correlations between the p values of the LRT and the exact test for Hardy-Weinberg proportion in cases under the four genetic models. If the two p values are indeed correlated, we examine the performance of the tail-strength statistics of Wang and Shete[1] under the null and alternative hypotheses. The analytical formula of the correlation, if any, between the LRT and the exact test for HWE used in Wang and Shete[1] is difficult to obtain. Therefore, we consider the combination of the p values of the trend test and chi-square test for HWE between cases and controls (HWDTT), from which the asymptotic correlation between the two p values has been obtained.[6,7] This new tail strength with the correlation is denoted by TSC. We further derive its asymptotic null distribution and critical value (see Appendix A). Comparison between our TSC and that of Wang and Shete[1] is obtained by Monte-Carlo simulations under the null and alternative hypotheses. We also denote the tail strengths based on the mean and median in Wang and Shete[1] by TS and TSM, respectively.

Here we report the main results from our simulation study. In the simulation, we assumed HWE holds in the population. In each replicate, 500 cases and 500 controls were generated under the null hypothesis with the baseline penetrance fixed at 0.02 (the probability of disease with a genotype of zero risk alleles), and minor-allele frequency (MAF) increases from 0.1 to 0.5 in increments of 0.1. We used a total of 10,000 replicates to estimate the null correlations between the two p values, the type I error rates, and power. The nominal levels 0.01 and 0.05 were used. For LRT statistics, we considered 1-degree-of-freedom tests. Therefore, for each genetic model under

Table 1. Simulated Null Correlations of the Two p Values of Wang and Shete[1] and the Asymptotic Type I Errors with Nominal Level 0.01

MAF	Model	Simulated Null Correlations	TS	TSM	TSC
0.1	REC	0.2702	0.0262	0.0272	0.0067
	ADD	−0.0049	0.0056	0.0057	0.0081
	DOM	0.0238	0.0072	0.007	0.009
0.2	REC	0.2327	0.0248	0.0251	0.0123
	ADD	0.0018	0.0092	0.0092	0.0108
	DOM	0.0328	0.0129	0.0128	0.0116
0.3	REC	0.1672	0.0268	0.026	0.0131
	ADD	0.0187	0.0104	0.0101	0.0112
	DOM	0.0505	0.017	0.017	0.0092
0.4	REC	0.1454	0.0225	0.0225	0.0118
	ADD	−0.0149	0.0076	0.0074	0.0083
	DOM	0.0716	0.0157	0.0153	0.0092
0.5	REC	0.1037	0.0197	0.0201	0.0128
	ADD	−0.0047	0.0074	0.0077	0.0103
	DOM	0.0919	0.0174	0.0175	0.0081

TS uses means as weights, TSM uses medians as weights, and TSC is the proposed test with the correlations. Three genetic models, which are only used for constructing the optimal LRTs (for TS and TSM) and optimal-trend tests (for TSC), are considered.

Table 2. Simulated Null Correlations of the Two p Values of Wang and Shete[1] and the Asymptotic Type I Errors with Nominal Level 0.05

MAF	Model	Simulated Null Correlations	TS	TSM	TSC
0.1	REC	0.2702	0.0841	0.0832	0.0436
	ADD	−0.0049	0.0394	0.0403	0.0477
	DOM	0.0238	0.0409	0.0408	0.0462
0.2	REC	0.2327	0.0765	0.0772	0.0476
	ADD	0.0018	0.0443	0.0441	0.0557
	DOM	0.0328	0.0513	0.0524	0.0472
0.3	REC	0.1672	0.0727	0.072	0.0482
	ADD	0.0187	0.0438	0.0438	0.0468
	DOM	0.0505	0.0519	0.0524	0.0531
0.4	REC	0.1454	0.0695	0.0678	0.0458
	ADD	−0.0149	0.0519	0.0516	0.0521
	DOM	0.0716	0.0574	0.0569	0.0494
0.5	REC	0.1037	0.0722	0.0704	0.0481
	ADD	−0.0047	0.0474	0.0483	0.0509
	DOM	0.0919	0.0647	0.0633	0.0475

TS uses means as weights, TSM uses medians as weights, and TSC is the proposed test with the correlations. Three genetic models, which are only used for constructing the optimal LRTs (for TS and TSM) and optimal-trend tests (for TSC), are considered.

the alternative hypothesis (REC, ADD/MUL, and DOM), an optimal test is available for the LRT or trend test. In the simulation, we consider three LRTs and three trend tests, optimal for the three genetic models. Therefore, a total of nine tail strengths were considered in the simulation: TS, TSM, and TSC each have three model choices depending on the targeted genetic model. The results of the null correlations between the two p values in Wang and Shete[1] and corresponding type I errors are reported in Table 1 for the nominal level 0.01 and in Table 2 for the nominal level 0.05.

The results in Tables 1 and 2 follow similar patterns. Thus, we focus on Table 1. The simulated null correlations between the p value of LRT and the p value of the exact HWE test in cases indicate that the null correlations are not zero when the LRT is optimal for the REC or DOM models, but they are close to zero for the ADD (MUL) model. Hence, the type I errors of the TS and TSM of Wang and Shete[1] are under control when the LRT is optimal for the ADD (MUL) model but are largely inflated when the LRTs are optimal for the REC and DOM models, especially for the REC model. Note that Wang and Shete[1] only considered the LRT optimal for the ADD model. Therefore, their type I errors were under control. On the other hand, the type I errors of TSC, which takes care of the correlations, are close to the nominal level regardless of the targeted genetic models.

We also conducted simulations to compare the powers of the TS, TSM, and TSC. For the TS and TSM, the correlations between the two p values were not incorporated. Thus, on the basis of results in Tables 1 and 2, their powers could be inflated under the REC and DOM models, but not under the ADD and MUL models. The powers are presented for the TS, TSM, and TSC (from left to right) under the REC model (Figure 1) and ADD model (Figure 2). The plots for the MUL and DOM models can be found in the Supplemental Data available online (Figures S1 and S2, respectively). The parameter values of the simulations under the alternative hypotheses are similar to those in Tables 1 and 2, except that the genotype relative risk (gamma2, which is defined as the ratio of penetrances with two risk alleles to those with zero risk alleles) ranges from 1 to 2, and the MAF is fixed at 0.3. The "asymptotic" and "simulated" powers in the figures were based on the critical values obtained from 10,000 parametric bootstrap samples and 10,000 permutations, respectively.

Figure 1 (under the REC model) shows that TS and TSM have similar powers and are more powerful than TSC. This could be due to the fact that TS and TSM had inflated type I errors as shown in Tables 1 and 2. On the other hand, Figure 2 shows that the powers of TS, TSM, and TSC are similar under the ADD model because the three statistics had similar type I errors. For the TS and TSM, the bootstrap and permutation procedures yield similar powers under the ADD, MUL, and DOM models but have slightly different powers under the REC model.

We also studied empirical powers of the TSC, the optimal trend test, a robust test MAX3[11], and classical Pearson's test for association under the four genetic models. The description and summary of our findings are given in Appendix B. The results show that the TSC has moderate power improvement under the REC model but loses significant power under the ADD and MUL models. This can be explained by the fact that testing HWE has little power under the ADD and MUL models.

In summary, the tail strength may improve power under some specific genetic models after correction for the correlation. However, when the underlying genetic model is unknown, the robust statistics are more preferable.[6,11]

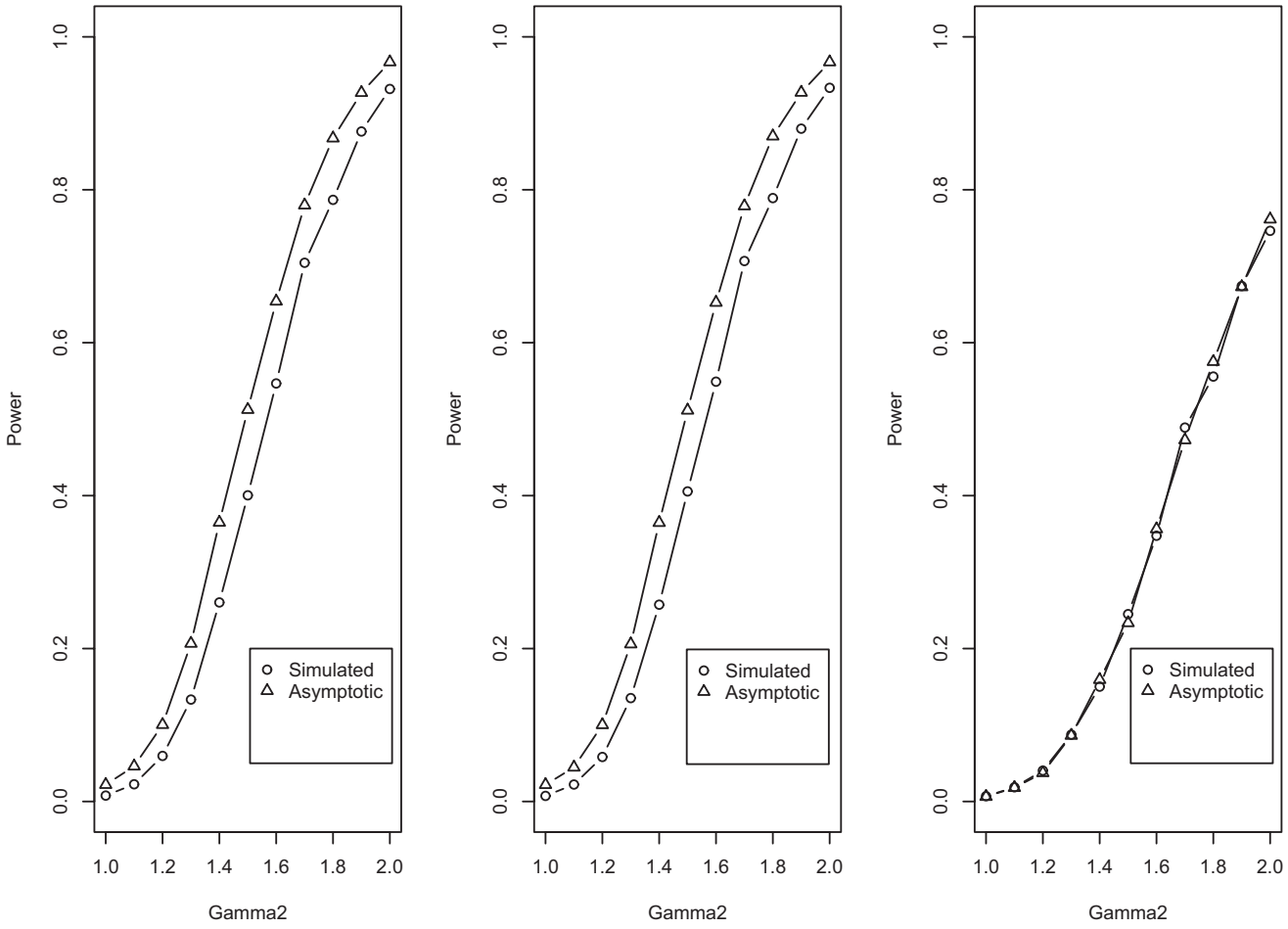

Figure 1. The Asymptotic and Simulated Powers under the REC Model
The tests from left to right are TS, TSM, and TSC. Gamma2 is the ratio of penetrances with two risk alleles to no risk alleles.

Appendix A

The Asymptotic Null Distribution of the TSC with the Correlation

Denote the HWDTT by Z_*, which is a statistic testing HWE between cases and control and was proposed by Song and Elston.[7] Denote the trend test as Z_x, where $x = 0$, 0.5, and 1 for the REC, ADD (MUL), and DOM models, respectively.[11–13] Under the null hypothesis H_0, (Z_*, Z_x) follows the bivariate normal distribution $N(0, \Sigma_1)$ with the density function f_1, where $\Sigma_1 = \begin{pmatrix} 1 & \rho_x \\ \rho_x & 1 \end{pmatrix}$, and $(-Z_*, Z_x)$ follows the bivariate distribution $N(0, \Sigma_2)$ with the density function f_2, where $\Sigma_2 = \begin{pmatrix} 1 & -\rho_x \\ -\rho_x & 1 \end{pmatrix}$. The expressions for ρ_x were given in Zheng and Ng for different x values.[6] The following derivation can be modified to the tail strength of any two correlated p values.

The p value of Z_* is $P_* = 2\Phi(-|z_*|)$, and the p value of Z_x is $P_x = 2\Phi(-|z_x|)$, where Φ is the cumulative distribution function of the standard normal $N(0, 1)$, and z_* and z_x are observed statistics. Then the joint distribution of P_* and P_x is:

$$\begin{aligned}
F(x_1,x_2) &= Pr(P_* < x_1, P_x < x_2) \\
&= Pr\left(\Phi(-|Z_*|) < \frac{x_1}{2}, \Phi(-|Z_x|) < \frac{x_2}{2}\right) \\
&= Pr\left(|Z_*| > \Phi^{-1}\left(1-\tfrac{x_1}{2}\right), |Z_x| > \Phi^{-1}\left(1-\tfrac{x_2}{2}\right)\right) \\
&= Pr\left(Z_* < \Phi^{-1}\left(\tfrac{x_1}{2}\right), Z_x < \Phi^{-1}\left(\tfrac{x_2}{2}\right)\right) \\
&\quad + Pr\left(Z_* < \Phi^{-1}\left(\tfrac{x_1}{2}\right), -Z_x < \Phi^{-1}\left(\tfrac{x_2}{2}\right)\right) \\
&\quad + Pr\left(-Z_* < \Phi^{-1}\left(\tfrac{x_1}{2}\right), Z_x < \Phi^{-1}\left(\tfrac{x_2}{2}\right)\right) \\
&\quad + Pr\left(-Z_* < \Phi^{-1}\left(\tfrac{x_1}{2}\right), -Z_x < \Phi^{-1}\left(\tfrac{x_2}{2}\right)\right) \\
&= 2\int_{-\infty}^{\Phi^{-1}\left(\frac{x_1}{2}\right)} \int_{-\infty}^{\Phi^{-1}\left(\frac{x_2}{2}\right)} f_1(y_1,y_2)\,dy_1 dy_2 \\
&\quad + 2\int_{-\infty}^{\Phi^{-1}\left(\frac{x_1}{2}\right)} \int_{-\infty}^{\Phi^{-1}\left(\frac{x_2}{2}\right)} f_2(y_1,y_2)\,dy_1 dy_2.
\end{aligned}$$

Thus, its density function can be written as

$$\begin{aligned}
f(x_1,x_2) &= \frac{\partial F(x_1,x_2)}{\partial x_1 \partial x_2} \\
&= \sum_{i=0}^{1} \exp\left[-\frac{\left\{\Phi^{-1}\left(\frac{x_1}{2}\right)\right\}^2 + \left\{\Phi^{-1}\left(\frac{x_2}{2}\right)\right\}^2 + (-1)^i 2\rho_x \left\{\Phi^{-1}\left(\frac{x_1}{2}\right)\right\}\left\{\Phi^{-1}\left(\frac{x_2}{2}\right)\right\}}{2(1-\rho_x^2)}\right] \\
&\quad \times \left\{2\sqrt{1-\rho_x^2}\,\exp\left[-\frac{\left\{\Phi^{-1}\left(\frac{x_1}{2}\right)\right\}^2 + \left\{\Phi^{-1}\left(\frac{x_2}{2}\right)\right\}^2}{2}\right]\right\}^{-1}.
\end{aligned}$$

Therefore, the ordered p values have the cumulative function given by

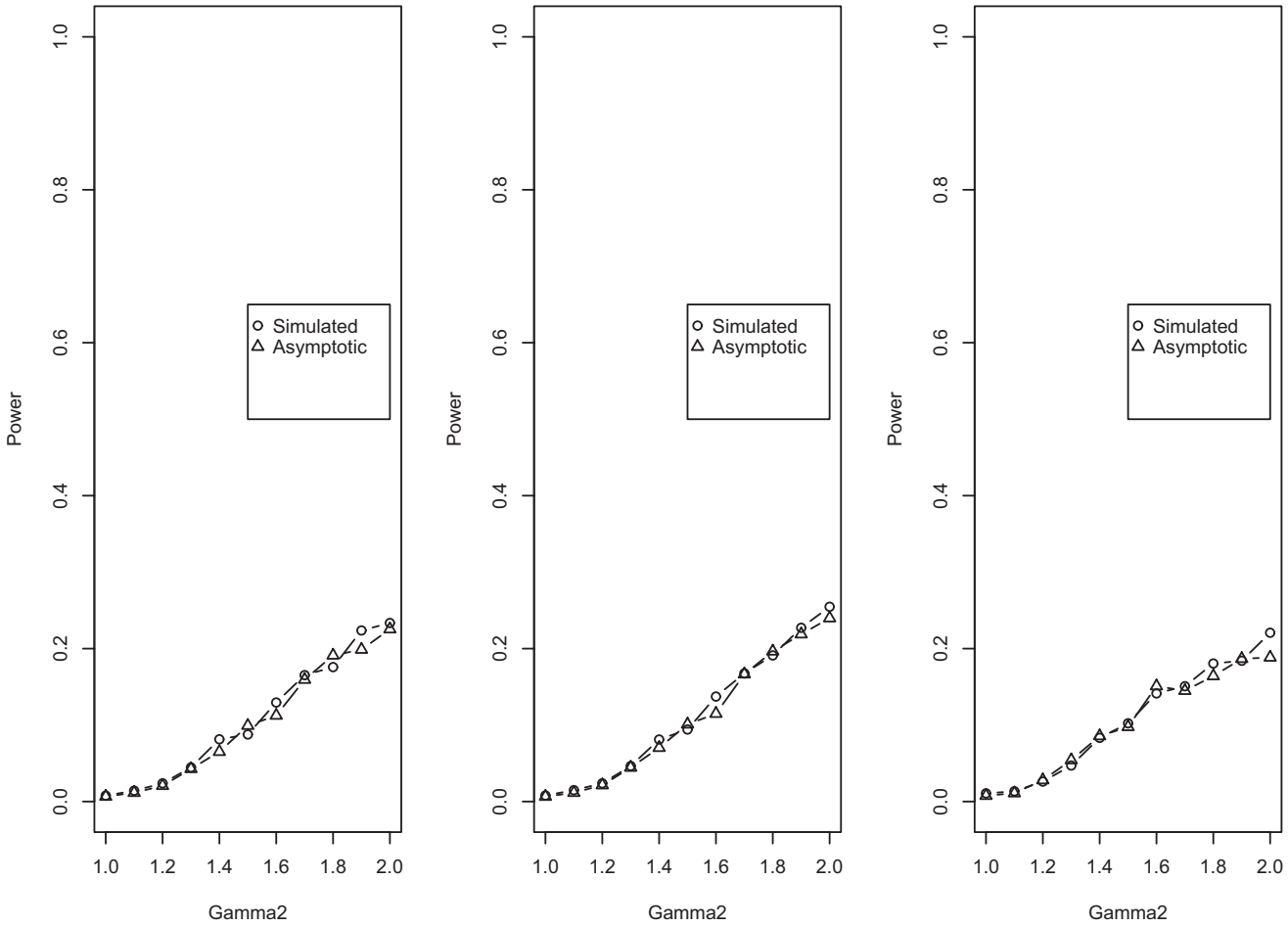

Figure 2. The Asymptotic and Simulated Powers under the ADD Model
The tests from left to right are TS, TSM, and TSC. Gamma2 is the ratio of penetrances with two risk alleles to zero risk allele.

$$\begin{aligned} F(x_{(1)},x_{(2)}) &= Pr(P_{(1)} < x_{(1)}, P_{(2)} < x_{(2)}) \\ &= Pr(P_{(2)} < x_{(2)}) - Pr(P_{(1)} \geq x_{(1)}, P_{(2)} < x_{(2)}) \\ &= Pr(P_* < x_{(2)}, P_x < x_{(2)}) - Pr(x_{(1)} \leq P_*, P_x < x_{(2)}) \\ &= \int_0^{x_{(2)}} \int_0^{x_{(2)}} f(y_1,y_2)\,dy_1 dy_2 \\ &\quad - \int_{x_{(1)}}^{x_{(2)}} \int_{x_{(1)}}^{x_{(2)}} f(y_1,y_2)\,dy_1 dy_2. \end{aligned}$$

The density function of the ordered p values is given by

$$\begin{aligned} g(x_{(1)},x_{(2)}) &= \frac{\partial F(x_{(1)},x_{(2)})}{\partial x_{(1)},\partial x_{(2)}} \\ &= f(x_{(1)},x_{(2)}) + f(x_{(2)},x_{(1)}),\ 0 \leq x_{(1)} \leq x_{(2)} \leq 1. \quad \text{(A1)} \end{aligned}$$

Once we obtain the above joint distribution $g(x_{(1)}, x_{(2)})$, we can use the results of Wang and Shete[1] to obtain the asymptotic null distribution for TSC:

$$\mathrm{TSC} = \frac{1}{2}\left\{ (1 - P_{(1)} \times 3) + \left(1 - P_{(2)} \times \frac{3}{2}\right) \right\}.$$

The density function of TSC is given by

$$\begin{aligned} f_{\mathrm{TSC}}(u) &= \int_{\frac{1}{6}(1-4u)}^{\frac{4}{9}(1-u)} \tfrac{4}{3} g(u,v)\,dv, \quad \text{when } u \in [-1.25, 0.25]; \\ &= \int_0^{\frac{4}{9}(1-u)} \tfrac{4}{3} g(u,v)\,dv, \quad \text{when } u \in [0.25, 1], \end{aligned}$$

where g is given in Equation (A1). We also consider a test for departure from HWE only by using cases in Appendix B. In this case, the above formulas can also be used except that the correlation needs to be modified accordingly.

Appendix B

Power Comparison between the Optimal-Trend Test, MAX3, Pearson's Test, and the TSC Tests

We compared the performance of several test statistics under the alternative hypotheses with the genotype relative risk 1.5, the disease prevalence 0.1, and 500 cases and 500 controls. The nominal level was 0.05. All critical values were obtained from the simulation with 100,000 replicates. The estimated powers were obtained from 10,000 replicates.

We considered four different genetic models: REC, ADD, MUL, and DOM models. Under each model the optimal-trend test was used.[11–13] These optimal-trend tests may not be realistic when the underlying genetic model is unknown. Thus, for comparison, we included two robust tests: MAX3, proposed by Freidlin et al.[11], and the classical Pearson's test with 2 degrees of freedom. For the tail strength, we considered TSC (the tail strength with the correlation). Two TSCs

were considered. One is discussed in the text (denoted by TSC2, where HWDTT is used), and the second one only uses cases to detect departure from HWE (denoted by TSC1).

The results from the simulations are reported in Table S1. The results show that TSC1 is usually more powerful than TSC2. Note that TSC1 is more powerful than the optimal trend test under the REC model when MAF is small to moderate. But TSC1 is much less powerful than the optimal trend test under the ADD and MUL models. This is because testing HWE has little power under these two models. TSC1 catches some power under the DOM model, but it is slightly less powerful than the optimal-trend test. On the other hand, when the genetic model is unknown, we cannot use the optimal-trend test. However, we compare the TSC1 with the robust test MAX3, which does not require that we know the genetic model. Table S1 shows that, except for the REC model, MAX3 is more powerful than TSC1.

Yong Zang,[1] Wing K. Fung,[1] and Gang Zheng[2,*]
[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China; [2]Office of Biostatistics Research, Division of Prevention and Population Sciences, National Heart, Lung, and Blood Institute, Bethesda, MD 20892-7913, USA
*Correspondence: zhengg@nhlbi.nih.gov

Supplemental Data

Supplemental Data include two figures and one table and are available with this article online at http://www.ajhg.org/.

Acknowledgments

We would like to thank Yaning Yang for some helpful discussions that brought our attention to the tail strength. The work of Y. Zang and W.K. Fung were partially supported by The Croucher Foundation and China Natural Science Foundation (no. 10701067).

Web Resources

The URL for data presented herein is as follows:

The R program (TSC.txt) used in the simulation can be downloaded from the website: www.statisticalsource.com/software/TSC1.txt.

References

1. Wang, J., and Shete, S. (2008). A test for genetic association that incorporates information about deviation from Hardy-Weinberg proportions in cases. Am. J. Hum. Genet. *83*, 53–63.
2. Taylor, J., and Tibshirani, R. (2006). A tail strength measure for assessing the overall univariate significance in a dataset. Biostatistics *7*, 167–181.
3. Elston, R.C. (1991). On Fisher method of combining p-values. Biometrical J. *33*, 339–345.
4. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature *445*, 881–885.
5. Conneely, K.N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of *p* values for multiple correlated tests. Am. J. Hum. Genet. *81*, 1158–1168.
6. Zheng, G., and Ng, H.K.T. (2008). Genetic model selection in two-phase analysis for case-control association studies. Biostatistics *9*, 391–399.
7. Song, K., and Elston, R.C. (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat. Med. *25*, 105–126.
8. Nielsen, D.M., Ehm, M.G., and Weir, B.S. (1998). Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am. J. Hum. Genet. *63*, 1531–1540.
9. Wittke-Thompson, J.K., Pluzhnikov, A., and Cox, N.J. (2005). Rational inferences about departure from Hardy-Weinberg equilibrium. Am. J. Hum. Genet. *76*, 967–986.
10. Wang, T., Zhu, X., and Elston, R.C. (2007). Improving power in contrasting linkage-disequilibrium patterns between cases and controls. Am. J. Hum. Genet. *80*, 911–920.
11. Freidlin, B., Zheng, G., Li, Z., and Gastwirth, J.L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum. Hered. *53*, 146–152.
12. Sasieni, P.D. (1997). From genotype to genes: doubling the sample size. Biometrics *53*, 1253–1261.
13. Zheng, G., Freidlin, B., Li, Z., and Gastwirth, J.L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. Biometrical J. *45*, 335–348.

DOI 10.1016/j.ajhg.2009.01.014.

Is the Tail-Strength Measure More Powerful in Tests of Genetic Association?

To the Editor: It is well known that Hardy-Weinberg equilibrium (HWE) is an important property in population genetics. Deviation from HWE among cases can provide evidence for a valid association.[1–4] Thus, it would be advisable to incorporate information from the HWE test for the improvement of power in detecting associated variants in genetic association studies. In the July 2008 issue of *The Journal*, Wang et al.[5] described a test statistic, the tail-strength (*TS*) measure,[6] for evaluation of the global null hypothesis, that the SNP was not associated with disease, which is a function of two p values: one from a logistic-regression test in a genetic association study and one from a HWE test in cases. The authors further extended the mean-based *TS* measure to a median-based measure (*TSM*) by measuring the deviation of each p value from its median value instead of its expected value. On the basis of simulation studies and real disease