**Two-Phase SSU and SKAT in Genetic Association Studies**Yuan Xue1,2, Ding Juan3, Jinjuan Wang1,4, Sanguo Zhang1,2, Qizhai Li4,\*

1 School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

2Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

3 School of Mathematics and Statistics, Guangxi Normal University, Guangzhou, China

4LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

\*Correspondence to: Qizhai Li, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, 55, Beijing 100190, China. E-mail: liqz@amss.ac.cn, Tel: 86-10-82541839.

**Abstract**

The SSU and SKAT are two good alternative tests in genetic association studies for case-control data. Both of them are derived via assuming a dose-response model between the risk of disease and genotypes. However, the real genetic mode of inheritance is impossible to know in practice. Thus, both tests will lose power as shown in simulations when the genetic model is misspecified. To make both of them be suitable in broad situations, we propose two-phase SSU (tpSSU) and two-phase SKAT (tpSKAT), where the Hardy-Weinberg equilibrium test is used to choose the genetic model in the first phase and the SSU and SKAT are constructed under the selected genetic model in the second phase. The extensive computer simulations show that the proposed tpSSU and tpSKAT is more robust than the original SSU and SKAT. At last application to rheumatoid arthritis case-control data from Genetic Analysis Workshop 16 further shows their performances.

**Keywords:** Multiple-markers analysis; genetic model; Hardy-Weinberg equilibrium; power; SNP.

**Introduction**

In the recent decade, large-scale genetic study, especially genome-wide association studies (GWAS) has led to the discovery of genetic variants for human complex diseases and traits. A standard case-control GWAS inevitably genotypes a large number of single nucleotide polymorphisms (SNPs). A simple and popular approach is to analyze single SNP first, then use a stringent level to select the deleterious SNPs. Although single-SNP analysis has been proven valid in detecting many disease-susceptibility variants, it can be inefficient because the adjustment to control the false positive rates is conservative, and also the genetic effect is weak, often with the estimated relative risk ranging from 0.095 (=ln1.1) to 0.405 (=ln1.5). So it is necessary to develop many effective methods **(Altshuler et al. 2008)**. In genetic studies, the true causal SNP is rarely genotyped and the genotyped SNPs are expected to be in linkage disequilibrium (LD) with it, and the LD patterns are quite variable in the genome (**Daly et al. 2001;** **Crawford et al. 2004)**. The single-SNP analysis in genotyped SNPs in LD with the causal SNP can serve as an imperfect surrogate of the causal SNP.

To overcome the drawback of single-SNP analysis, multiple-SNPs approach considering the joint effect of many SNPs has been developed. A group of SNPs is analyzed together to obtain a global test statistic for the combined effects. There are many procedures in the literature (**Pan 2009; Han and Pan 2010; Neale et al. 2011; Wu et al. 2011; Ballard et al. 2012; Yoo et al. 2015; Yoo et al. 2017**). Here, we only list some of them. Based on the comparison of genotype differences between cases and controls, **Xiong et al. (2002)** proposed to use Hoteling’s test, which is equivalent to applying a multivariate regression to simultaneously test all SNPs, such as likelihood ratio test, Wald test and score test in the context of logistic regression (**Wang and Elston 2007**). Based on the differences of LD patterns between cases and controls, **Zaykin et al. (2006)** came up with a linkage disequilibrium contrast test and **Wang et al. (2007)** gave its modified version. To contrast both genotypes scores and LD patterns, **Pan (2009; 2014)** proposed the sum of squared score (SSU) and their weighted versions, and **Wu et al. (2010)** proposed a sequence kernel association test (SKAT). It has been shown that SSU is equivalent to SKAT with a linear kernel being used (**Basu and Pan 2011; Wu et al. 2011**).

All the above methods have a common point, that is, the association of genetic variants with a disease or a trait is assessed under the additive genetic mode of inheritance. According to the dose of risk allele in regression model, the genotypes are usually encoded as 0, 1 and 2 corresponding to the number of risk allele. The genetic model is a functional relationship of the risk measurement of a given genotype, which describes the effect of genotypes on phenotypes (**Zheng et al. 2016**). In reality, it is impossible to know the true inheritance pattern. For example, **Moltke et al. (2014)** reported a genetic variant p.Arg684Ter being associated with 2-h plasma glucose levels and type 2 diabetes in a recessive model. **Nik-Zainal et al. (2016)** found five genes including MED23, FOXP1, MLLT4, XBP1, and ZFP36L1, acting in breast cancer in a recessive fashion. Misspecifying genetic model might lead to a decline in power of a test. So, it is urgent to develop a robust test that is free of genetic model and has a satisfactory power over a range of genetic models.

The Hardy-Weinberg equilibrium (HWE) principal plays an important role in genetic epidemiologic studies. Checking whether the HWE holds in the control sample is a key before doing an association test. There are many factors that can result in deviation from HWE, such as genotyping error, population stratification and so on **(Hosking et al. 2004; Wigginton et al. 2005; Schaid et al. 2006; Zhang et al. 2015)**. The HWE test can be used to determine the genetic model of a SNP **(Zheng and Ng 2008; Zhang and Li 2016; Zheng et al. 2016; Hu et al. 2017)**. In this work, we propose a two-phase procedure for multiple-SNPs analysis. In the first phase, we use the HWE test to choose the genetic model for each SNP. In the second phase, based on the chosen genetic model, the SSU and SKAT are constructed to do association studies. We use a permutation procedure to access its statistical significance. Extensive computer simulations are conducted to evaluate the performances of the proposed method by comparing with the original SSU and SKAT by taking the genotype as a dose effect. They show that the proposed method is more powerful than SSU and SKAT under most of the considered situations, especially when the real genetic model is recessive. Finally, application to rheumatoid arthritis case-control data further shows its performance.

**Methods**

Suppose that there are *n* individuals randomly sampled in a source population and *m* SNPs are genotyped on each subject in a case-control genetic study. Denote the disease status of the individual by where means a case and indicates a healthy control, . Without loss of generality, an allelic SNP with alleles *A* and *a* is considered here and let be the genotype values of the subject, where denotes the number of allele *A,* , and denotes the transpose of a matrix or a vector. For the SNP, we define as the number of cases, as the number of subjects with genotype *aa*, *Aa* and *AA* in cases, respectively, and as the number of subjects with genotype *aa*, *Aa* and *AA* in the whole sample, respectively. Denotebe *k*-dimensional covariates. The proposed tests consist with two phases with phase I choosing the genetic model and phase II constructing the association tests under the chosen model.

**Phase I: Genetic model selection**

**Song and Elston (2006)** proposed a Hardy-Weinberg equilibrium test (HWET) based on the data of both cases and controls. Recently, **Zheng et al. (2016)** showed that the HWET constructed only in cases has a higher power to detect the genetic model than that constructed in both cases and controls. The HWET on the basis of case data for SNP () is

where Then, the genetic model can be determined as: if , the recessive model is selected; if , the dominant model is selected; otherwise (), the additive model is assumed.

Next, we recode the genotype values based on the determined genetic model. For the SNP, the genotypes are recoded as (0,0,1) (0,0.5,1) and (0,1,1) corresponding to recessive, additive and dominant models, respectively. Denote the recoded genotype values by

**Phase II: SSU and SKAT**

The logistic regression model is used to evaluate the association between genotypes and the disease as

where Logit is a logistic function with is the parameter of interest and is the parameter corresponding to covariate we denote The null hypothesis is which is an *m*-dimensional column vector with all zero units.

Following **Pan (2009; 2010)**, we can construct the SSU, called tpSSU. If there is no covariate in the model, that is to say,, the score vector and its covariance matrix based on logistic regression model are

where and . If , the full score vector under is

,

in which **,**  are the maximum likelihood estimates (MLEs) of the regression coefficient estimated under the null model. The covariance matrix of can be consistently estimated by the expected Fisher information matrix,

where To test , the score vector , which is the first *m* components of , and the covariance matrix of is , where is the upper-left submatrix of with size .Once we get and . The SSU is constructed as (here, we called it tpSSU since the genotype values are recoded based on the selected model in Phase I)

where is the component of , and is the diagonal element of .

Similarly, we can construct the SKAT, called tpSKAT. From **Wu et al. (2010),** the tpSKAT is constructed as

where *,*  is the kernel function used in the statistic, we use the IBS kernel functionin our work.

**Statistical Significance Calculations**

To get the statistical significance of tpSSU and tpSKAT, we have to know the distribution of them, but it is really hard to get the joint distribution of and , so we use a permutation procedure to get the p-values of both two-phase tests. The detailed procedures are as follows:

1. Set a large number of , for example, .
2. Determine the genetic model of each SNP. For the SNP, we construct the HWET based on . The genetic model can be determined by if , the recessive model is selected; if , the dominant model is selected; otherwise (), the additive model is assumed.
3. The genotype score (0,1,2) is recoded as (0,0,1) (0,0.5,1) and (0,1,1) corresponding to the chosen recessive, additive and dominant models, respectively. Denote the recorded genotype values by . Then calculate the and based on Denote them by respectively.
4. Permute the original datatimes. For we compute the tpSSU and tpSKAT and denote them by respectively.
5. Calculate the p-values as

where the indicates the number of elements in a set.

**Results**

**Simulation Settings**

We carry out simulation studies following **Wang and Elston (2007)** and **Pan (2009)** considering and , and the sample size is with 500 cases and 500 controls.

We generate the genotypes from the following steps. First, a latent vector with length of 10 is generated from a multivariate normal distribution with two different covariance structures: a compound symmetry (CS) with the correlation , and AR-1 with the correlation matrix between components and . The CS correlation structure suggests that each SNP provides similar information on the disease locus, and in the AR-1, the pairwise correlation represents that the LD is primarily a function of marker distance. Second, the latent vector is dichotomized to yield genotypes on 10 SNPs with allele frequencies randomly between 0.2 and 0.8, while the allele frequency for the disease-causing SNP is fixed at 0.3. Third, the genotypes (0,1,2) are recoded as (0,0,1) (0,0.5,1) and (0,1,1) corresponding to recessive, additive and dominant model, respectively. Denote the recoded genotypes is , and the disease-causing genotypes is recoded as , where is the number of disease-causing SNPs. And the covariate is generated from a standard normal distribution. The disease status of subjects is generated from a logistic regression model

where we chose . In order to evaluate the type I error, we set all being 0. For power calculation, we set ranging from 0.18 to 0.96, and are equal or opposite sign when . Finally, following the case-control design, we sample cases (with ) from the case group and controls (with ) from the control group.

Comparing with SSU and SKAT, we show the performance of the two-phase tests (tpSSU and tpSKAT) under five scenarios as described in Table 1. For each scenario, 1000 replicated datasets to estimate the empirical type I error and power with setting the significant level of 0.05 are conducted. For the permutation procedure, .

Table 1 Trait scenarios for simulation studies

|  |  |  |  |
| --- | --- | --- | --- |
| Scenarios | Number of causal SNP | Genetic model |  |
| S1 | 1 | recessive | 0, 0.18, 0.34, 0.47, 0.59, 0.69, 0.79, 0.88, 0.96 |
| S2 | 2 | recessive | 0, 0.18, 0.34, 0.47, 0.59, 0.69 (equal) |
| S3 | 2 | recessive | 0, 0.18, 0.34, 0.47, 0.59, 0.69, 0.79,0.88, 0.96 (opposite) |
| S4 | 1 | dominate | 0, 0.18, 0.34, 0.47, 0.59, 0.69 |
| S5 | 1 | additive | 0, 0.18, 0.34, 0.47, 0.59, 0.69 |

**Type I Error Rates**

Figure 1 shows the empirical type I error rates of the tpSSU, tpSKAT, SSU, and SKAT. Obviously, all the values are very close to the nominal significance level of 0.05, which means that all the considered methods can control the type I error rates well. For example, when *m*=10, with the AR-1 covariance structures, the empirical type I error rates of the tpSSU, tpSKAT, SSU, and SKAT are 0.054 0.053, 0.055, and 0.056, respectively.

**Power Comparison**

The empirical power results of Scenario S1 to S5 are shown in Figures 2-6. It can be easy to catch that the tpSSU and tpSKAT are more robust than the SSU and SKAT. When the genetic model of causal SNPs is additive, all of the methods have similar results, and when the genetic model of causal SNP is dominate, the tpSSU and tpSKAT are more powerful than the SSU and SKAT, especially when the genetic model of causal SNP is recessive, the tpSSU and tpSKAT have a great power promotion. Because of the similar results, for each scenario, we will give detail results of *m*=10.

Figure 2 shows the empirical power of Scenario S1. It can be seen that sometimes the tpSSU and SKAT has more than 20% power improvements than the SSU and SKAT. For example, under Scenario S1, when  is 0.69 with the CS covariance structure, the empirical power of the tpSSU, tpSKAT, SSU, and SKAT are 0.564, 0.484, 0.353, and 0.364, respectively; when is 0.88 with the AR-1 covariance structure, the power of the tpSSU, tpSKAT, SSU, and SKAT are 0.823, 0.76, 0.619, and 0.635, respectively.

Figure 3 shows the empirical power of Scenario S2. From the figure, we can get the same result as above when there are two causing SNPs. For instance, when *β* is 0.47 with the CS covariance structure, the empirical power of the tpSSU, tpSKAT, SSU, and SKAT are 0.677, 0.589, 0.489, and 0.505, respectively; when *β* is 0.47 with the AR-1 covariance structure, the power of the tpSSU, tpSKAT, SSU, and SKAT are 0.63, 0.534, 0.432, and 0.437, respectively.

Figure 4 shows the empirical power of Scenario S3. It can be easy to find that the power of all methods is less than that shown in S2, which the two causal SNPs have the same sign. And under the scenario S3, we can get the same results, the tpSSU and tpSKAT also has a significant improvement of power than the SSU and SKAT. For example, With the CS covariance structure, when is 0.88, the power of the tpSSU, tpSKAT, SSU, and SKAT are 0.57, 0.435, 0.264, and 0.315, respectively; when the is 0.79, the power of the tpSSU, tpSKAT, SSU, and SKAT are 0.71, 0.612, 0.501, and 0.507, respectively.

Figure 5 shows the empirical power results when the genetic model of causing SNP is dominate. We can find that there also has about 5-8% power improvement. For example, with the CS covariance structure, when is 0.59, the power of the tpSSU, tpSKAT, SSU, and SKAT are 0.893, 0.939, 0.817, and 0.888, respectively.

Figure 6 shows the empirical power results when the genetic model of causing SNP is additive. We can see that all of four methods have similar results. For example, with the CS covariance structure, when is 0.69, the power of the tpSSU, tpSKAT, SSU, and SKAT are 0.722, 0.703, 0.699, and 0.698, respectively.

**Application**

Rheumatoid arthritis (RA) is a chronic autoimmune disease that are generally triggered by infections and inflammatory mediators, with a prevalence of approximately 0.3–1% worldwide **(Liang et al. 2009; Chaudhari et al. 2016)**. We used the RA case-control data in Genetic Analysis Workshop 16 to illustrate the application of the four methods: tpSSU, tpSKAT, SSU, and SKAT. After quality control, there is 868 RA cases and 1,194 controls **(Plenge et al. 2007)**. The chromosome 10 is shown to be associated with RA **(Zhang et al. 2009; Okada et al. 2014).** We analyzed chromosome 10q21.22 region (49.64-49.79 Mb on chromosome 10), which containing SNPs. Set B= for permutation procedure, p-value less than 10-6 is reported for both tests tpSSU and tpSKAT. Meanwhile, the p-values reported by SSU and SKAT are 7.14 and 1.97 respectively. All the four tests show that the 10q21.22 region is significantly related with RA. The most significant SNP is rs2671692, which is one of 42 novel loci identified in a large GWAS study meta-analysis based on a total of >100,000 subjects of European and Asian ancestries **(Okada et al. 2014)**.

**Discussion**

Multiple-SNPs analysis has been well appreciated because of its potentially improved statistical power in genetic association studies. Most of the existing association studies were done by assuming an additive genetic effect, but the true genetic model is usually unknown in advance. It is well known that a test derived from an improper model might result in a substantial loss of power, especially when the genetic model is recessive. To overcome this drawback, a new test, which is free of genetic model assumption, is ideal. In this work, we proposed two tests, tpSSU and tpSAKT to handle this issue. Numerical studies showed that, compared with the test derived under the additive genetic model, the power of the tpSSU and tpSKAT has a power promotion when the genetic model causal SNP is recessive or dominate. Especially when the causal SNP belongs to the recessive model, our proposed method showed great efficiency.

A genetic model is a measure of the risk of a functional relationship. In reality, the genetic model is usually unknown. Using inappropriate genetic model will inevitably result in loss of efficiency especially when the true model is recessive (dominant) and a dominant (recessive) used. The Hardy-Weinberg equilibrium choosing the true genetic model shows a high accuracy rate, **(Zheng and Ng, 2008, Hu et al. 2017)**.

In general, there is no consistency of the most powerful tests for multiple parameters in the case of multi-SNPs association tests, nevertheless, the SSU and SKAT have shown its popularity in most situation **(Park et al. 2017)**. The SSU in our work is the weight version of SSU in **Pan (2009),** which is called the summation of weighted squared score in the original paper. Both tests are based on ignoring the covariance matrix of the score test statistic. In most situations, the summation of weighted squared score has a robust property. There are also some other types of SSU. **Pan and Shen (2011)** proposed the adaptive tests for association analysis of rare variants and **Pan et al. (2015)** proposed a general score-based statistic test. Expect for the IBS kernel function, the common kernel function for SKAT includes linear kernel function and quadratic kernel function **(Wu et al. 2010)**. It has shown that SSU is equivalent to kernel machine regression test with a linear kernel (**Basu and Pan 2011; Wu et al. 2011)**. And a weighted version of SKAT is also available and one can get more details about the setting of weight in **Wu et al. (2011)**.

In our work, we consider when the disease status is binary. It has been shown the recessive and dominant model are well defined for a quantitative as well **(Zheng et al. 2016),** the definition of the genetic model can be got by a transformed binary trait. It's very worth studying the further extension of the two-phase method to the ordinal or quantitative outcome.

**Funding:** This study was funded by Breakthrough Project of Strategic Priority Program of the Chinese Academy of Sciences (Grant No. XDB13040600), and National Science Foundation of China (Grant No. 11722113; Grant No. 11501134).

**Compliance with Ethical Standards**

**Conflict of Interest:** Yue Xue, Juan Ding, Jinjuan Wang, Sanguo Zhang, Qizhai Li declare that they have no conflicts of interest.

**Ethical Approval:** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed Consent:** Informed consent was obtained from all individual participants included in the study.

**Reference**

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322(5903): 881-888

Ballard DH, Cho J, Zhao H (2010) Comparisons of multi‐marker association methods to detect association between a candidate region and disease. Genet Epidemiol 34(3):201-212

Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35(7): 606-619

Chaudhari K, Rizvi S, Syed BA (2016) Rheumatoid arthritis: current and future trends. Nat Rev Drug Discov 15:305-306

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD et al (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74(4):610-622

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nature Genet 29(2):229

Derkach A, Lawless JF, Sun L (2012) Assessment of pooled association tests for rare genetic variants within a unified framework. arXiv preprint arXiv:1205.4079

Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70(1):42-54

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A et al (2004) Detection of genotyping errors by Hardy–Weinberg equilibrium testing. Eur J Hum Genet 12(5):395

Hu X, Duan X, Pan D, Zhang S, Li Q (2017) A model-embedded trend test with incorporating Hardy-Weinberg equilibrium information. J Syst Sci Complex 30(1):101-110

Liang X, Gao Y, Lam TK, Li Q, Falk C, Yang XR et al (2009) Identifying rheumatoid arthritis susceptibility genes using high-dimensional methods. BMC Proc 3 (Suppl 7):S79

Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M et al (2014) A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature 512(7513):190

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M et al (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7(3):e1001322

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X et al (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534(7605):47

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K et al (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506(7488): 376

Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33(6):497-507

Pan W, Kim J, Zhang Y, Shen X, Wei P (2014) A powerful and adaptive association test for rare variants. Genetics 197(4):1081-1095

Pan W, Kwak IY, Wei P (2015) A powerful pathway-based adaptive test for genetic association with common or rare variants. Am J Hum Genet 97(1):86-98

Pan W, Shen X (2011) Adaptive tests for association analysis of rare variants. Genet Epidemiol 35(5):381-388

Park JY, Wu C, Basu S, McGue M, Pan W. (2018) Adaptive SNP-Set Association Testing in Generalized Linear Mixed Models with Application to Family Studies. Behav Genet 1-12

Plenge, RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B et al (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. N Engl J Med 357:1199–1209

Schaid DJ, Batzler AJ, Jenkins GD, Hilderbrandt MA (2006) Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am J Hum Genet 79:1071–1080

Song K, Elston RC (2006) A powerful method of combining measures of association and Hardy–Weinberg disequilibrium for fine‐mapping in case‐control studies. Stat Med 25(1):105-126

Wang T, Elston RC (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet 80(2):353-360

Wang T, Zhu X, Elston RC (2007) Improving power in contrasting linkage-disequilibrium patterns between cases and controls. Am J Hum Genet 80(5):911-920

Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887–893

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82-93

Wu M C, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ et al (2010) Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86(6):929-942

Xiong M, Zhao J, Boerwinkle E (2002) Generalized T2 test for genome association studies. Am J Hum Genet 70(5):1257-1268

Yoo YJ, Kim SA, Bull SB (2015) Clique-based clustering of correlated SNPs in a gene can improve performance of gene-based multi-bin linear combination test. Biomed Res Int Article ID 852341

Yoo YJ, Sun L, Poirier JG, Paterson AD, Bull SB (2017) Multiple linear combination (MLC) regression tests for common variants adapted to linkage disequilibrium structure. Genet Epidemiol 41(2):108-121

Zaykin DV, Meng Z, Ehm MG (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 78(5):737-746

Zhang M, Lin Y, Wang L, Pungpapong V, Fleet JC, Zhang D (2009) Case-control genome-wide association study of rheumatoid arthritis from genetic analysis workshop 16 using penalized orthogonal-components regression-linear discriminant analysis. BMC Proc 3(S7):S17.

Zhang W, Li Q (2016) Incorporating Hardy–Weinberg Equilibrium Law to Enhance the Association Strength for Ordinal Trait Genetic Study. Ann Hum Genet 80(2):102-112

Zhang W, Zhang Z, Li X, Li Q (2015) Fitting proportional odds model to case-control data with incorporating hardy-weinberg equilibrium. Sci Rep 5:17286

Zheng G, Ng HKT (2008) Genetic model selection in two-phase analysis for case–control association studies. Biostatistics 9(3):391-399

Zheng G, Zhang W, Xu J, Yuan A, Li Q, Gastwirth JL (2016) Genetic risks and genetic model specification. J Theor Biol 403:68-74

**The captions for illustrations**

Figure 1. The empirical Type-I error rates of the tpSSU, tpSKAT, SSU, and SKAT with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

Figure 2. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S1 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

Figure 3. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S2 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

Figure 4. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S3 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

Figure 5. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S4 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

Figure 6. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S5 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

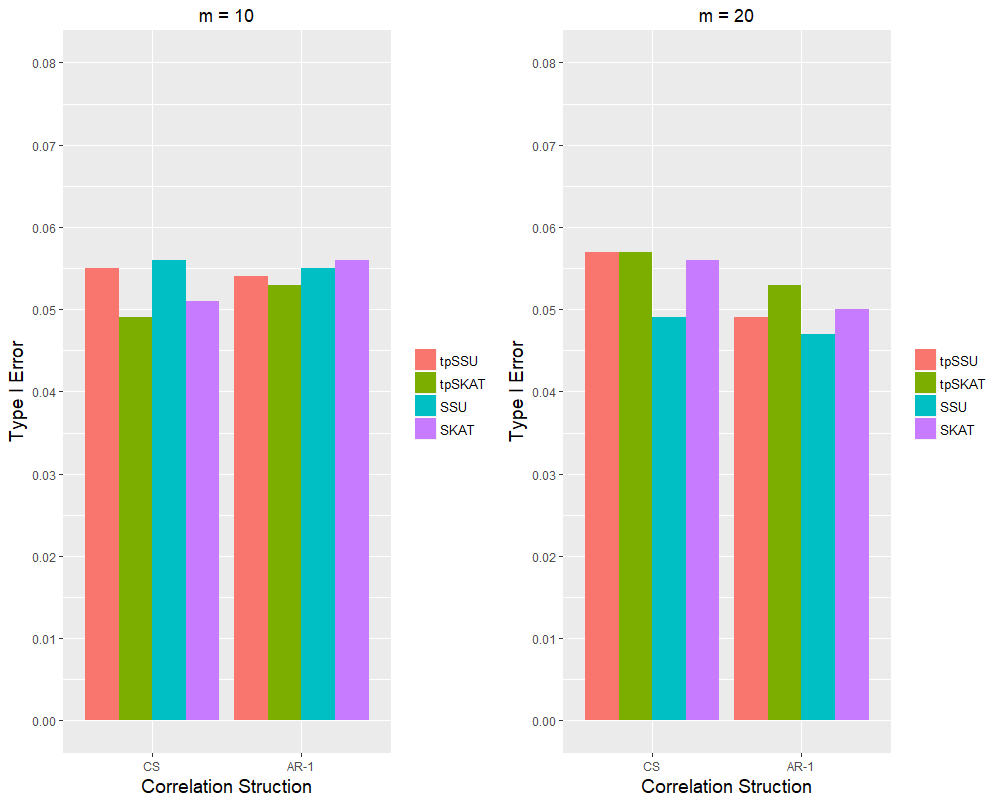


Figure 1．The empirical Type-I error rates of the tpSSU, tpSKAT, SSU, and SKAT with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

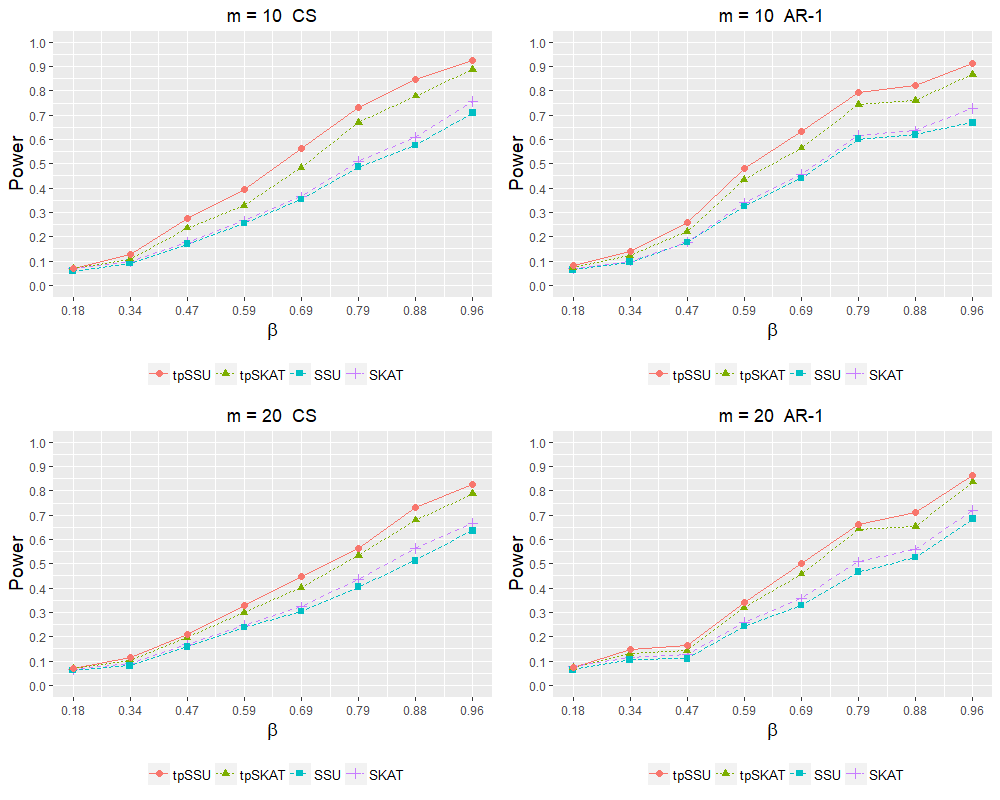


Figure 2. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S1 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

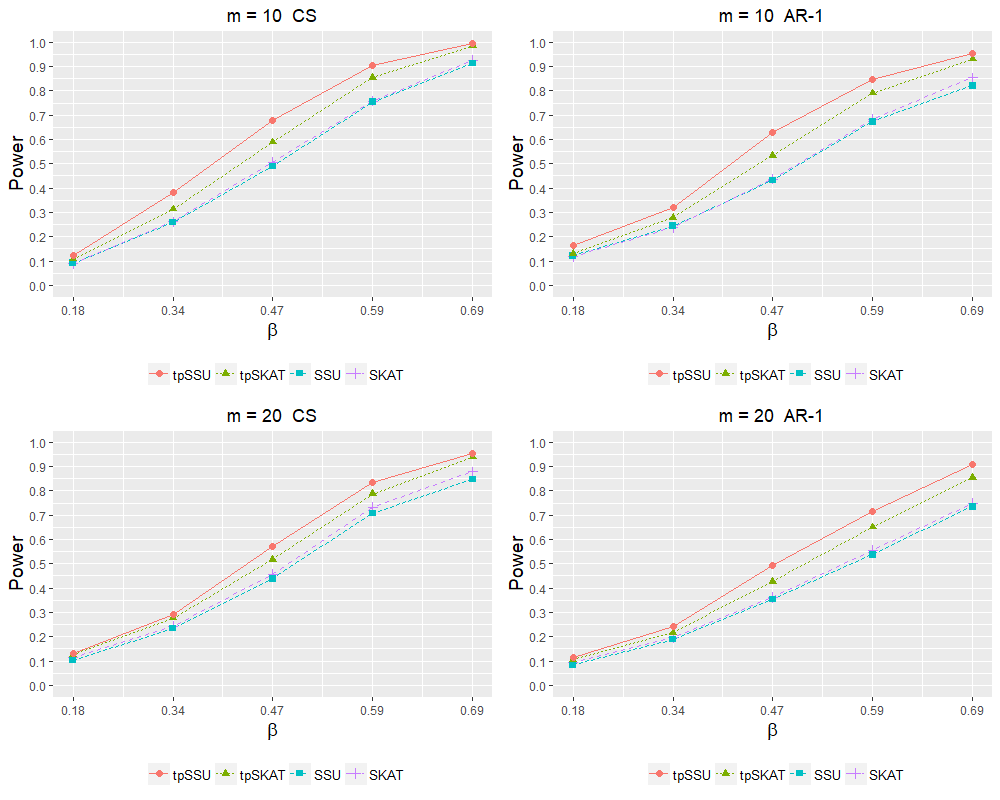


Figure 3. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S2 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

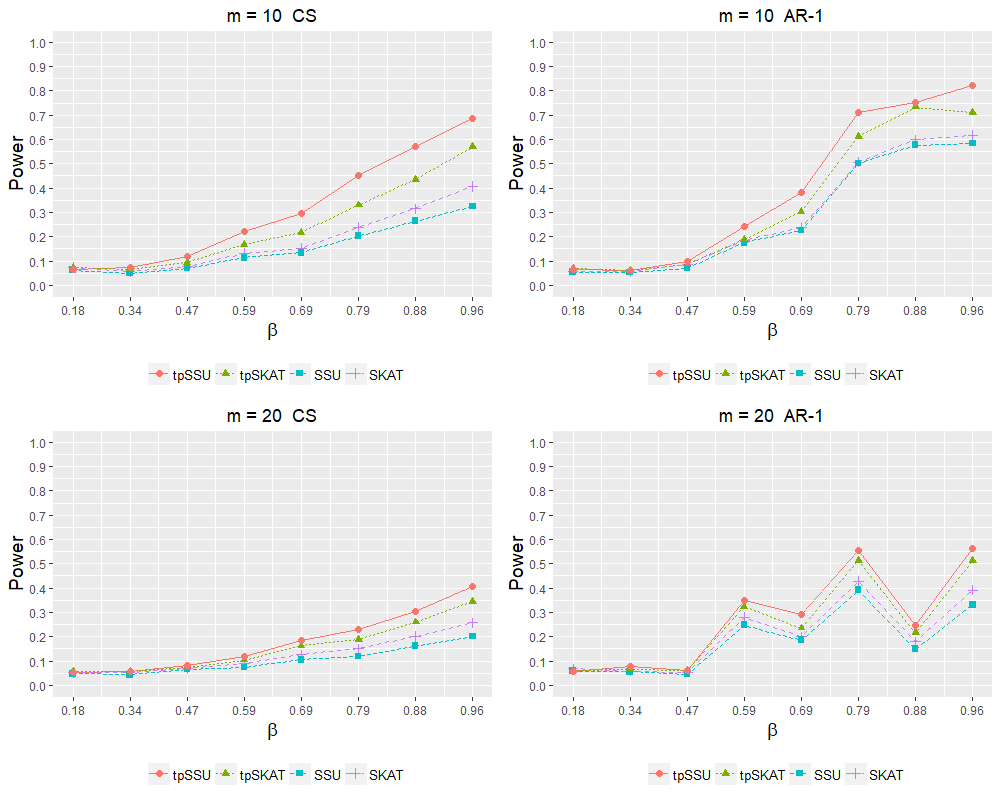


Figure 4. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S3 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

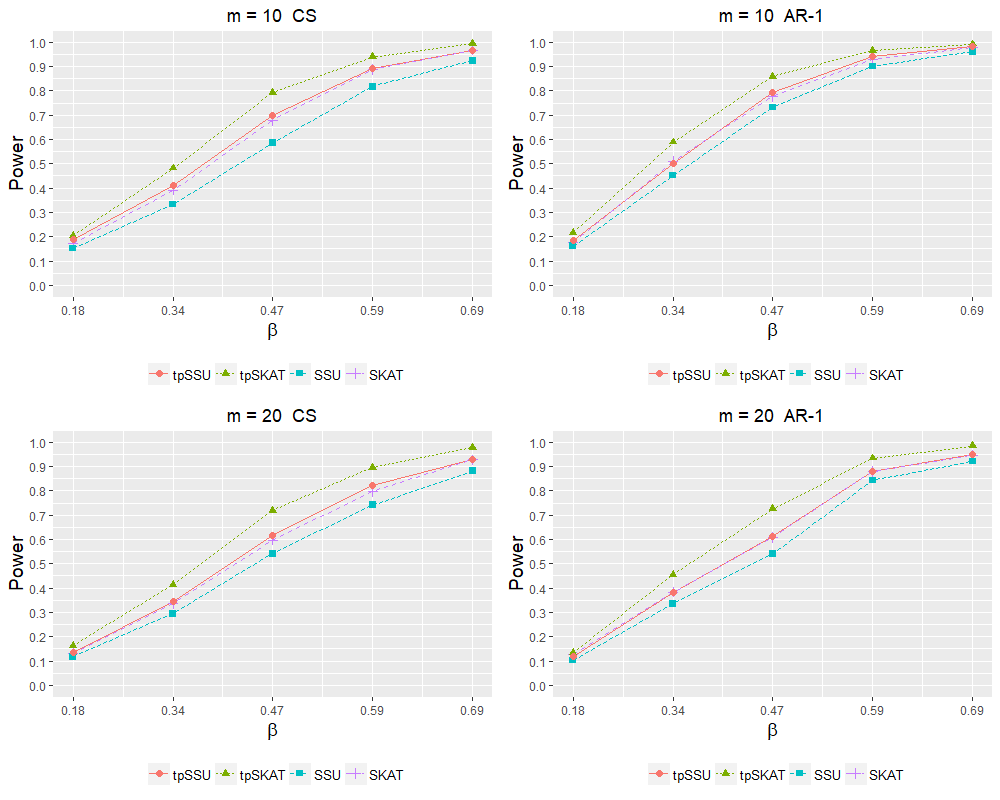


Figure 5. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S4 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).

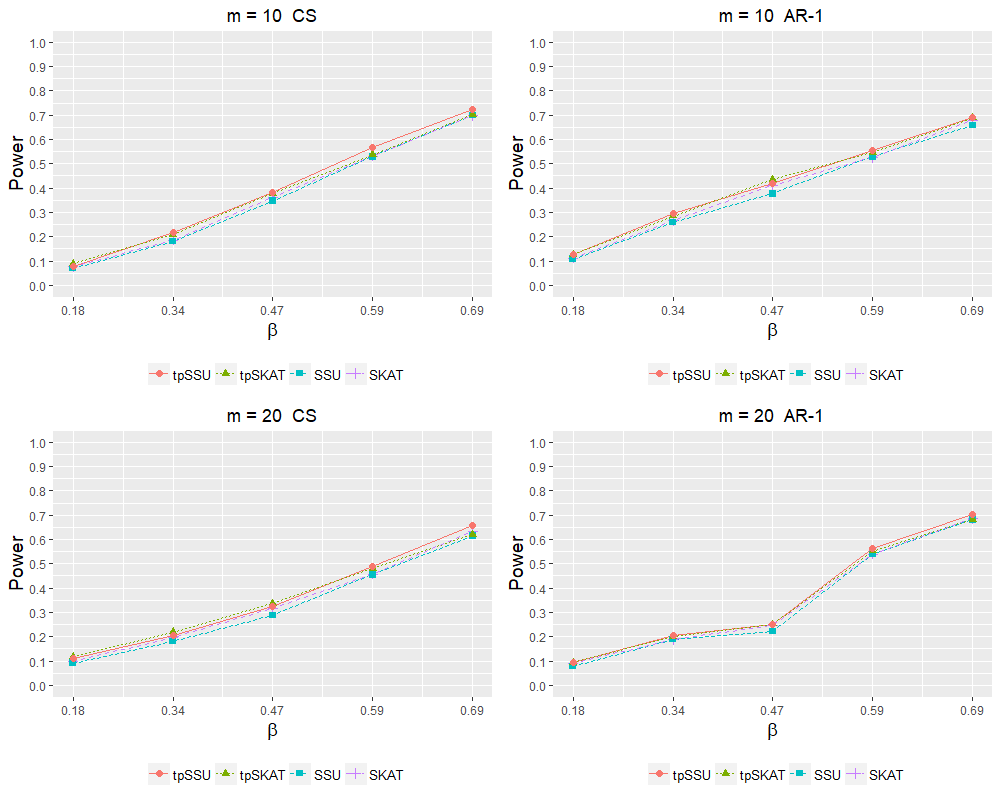


Figure 6. The empirical power of the tpSSU, tpSKAT, SSU, and SKAT under Scenarios S5 with the CS correlation structure (CS) and the AR-1 correlation structure (AR-1).