Bioinformatics, 32(18), 2016, 2737–2743
doi: 10.1093/bioinformatics/btw314
Advance Access Publication Date: 3 June 2016
Original Paper

OXFORD

Genome analysis

Group-combined *P*-values with applications to genetic association studies

Xiaonan Hu[1,2,†], Wei Zhang[3,†], Sanguo Zhang[1,2], Shuangge Ma[4] and Qizhai Li[3,*]

[1]School of Mathematical Sciences, University of Chinese Academy of Sciences, [2]Key Laboratory of Big Data Mining and Knowledge Management and [3]Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and [4]Department of Biostatistics, Yale University, New Haven, CT, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on January 18, 2016; revised on May 6, 2016; accepted on May 13, 2016

Abstract

Motivation: In large-scale genetic association studies with tens of hundreds of single nucleotide polymorphisms (SNPs) genotyped, the traditional statistical framework of logistic regression using maximum likelihood estimator (MLE) to infer the odds ratios of SNPs may not work appropriately. This is because a large number of odds ratios need to be estimated, and the MLEs may be not stable when some of the SNPs are in high linkage disequilibrium. Under this situation, the *P*-value combination procedures seem to provide good alternatives as they are constructed on the basis of single-marker analysis.
Results: The commonly used *P*-value combination methods (such as the Fisher's combined test, the truncated product method, the truncated tail strength and the adaptive rank truncated product) may lose power when the significance level varies across SNPs. To tackle this problem, a group combined *P*-value method (GCP) is proposed, where the *P*-values are divided into multiple groups and then are combined at the group level. With this strategy, the significance values are integrated at different levels, and the power is improved. Simulation shows that the GCP can effectively control the type I error rates and have additional power over the existing methods—the power increase can be as high as over 50% under some situations. The proposed GCP method is applied to data from the Genetic Analysis Workshop 16. Among all the methods, only the GCP and ARTP can give the significance to identify a genomic region covering gene DSC3 being associated with rheumatoid arthritis, but the GCP provides smaller *P*-value.
Availability and implementation: http://www.statsci.amss.ac.cn/yjscy/yjy/lqz/201510/t20151027_313273.html
Contact: liqz@amss.ac.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the past two decades, more than ten thousands genetic markers have been identified to be associated with many human complex diseases using genome-wide association studies where single-marker analysis was adopted. In the post-genome era, the use of multi-marker analysis where several markers are analyzed in the same model provides a powerful solution in large-scale genetic association studies.

To detect the association between a gene or a genomic region (which may host tens of hundreds of single nucleotide polymorphisms—SNPs) and complex human diseases in large scale genetic studies, the traditional method of logistic regression may not work

appropriately. This is because it employs the likelihood function and uses the maximum likelihood approach to estimate the odds ratios, and a high-dimensional problem needs to be optimized. The global solution is difficult to find, and also in some cases when the SNPs are in high linkage disequilibrium (LD), the Hessian matrix derived from the likelihood function may be singular and thus the solution is not stable. A useful alternative strategy is to analyze the SNPs one by one and then combine the *P*-values together.

For combining the *P*-values, one popular choice is the Fisher's method (Fisher, 1932), also referred to as the Fisher's combined test (FCT). If the m *P*-values to be combined are independent, the FCT statistic follows the Chi-squared distribution with $2m$ degrees of freedom. If these m *P*-values are correlated, the re-sample methods are recommended to obtain the empirical distribution of the FCT. Otherwise the type I error rates may be inflated (Zaykin *et al.*, 2002). The FCT has been widely applied in genetic association studies (Hess and Iyer, 2007; Li *et al.*, 2014; Zheng *et al.*, 2012). However, it may lose power when the degrees of freedom are large and most of the SNPs are not significant. To tackle this problem, combining the truncated *P*-values have been recommended. Dudbridge and Koeleman (2003) proposed to combine the k smallest *P*-values as the test statistic. However the power of this approach is very sensitive to the value of k—an inappropriate k results in combing some non-significant *P*-values and power loss, especially when m is large (Dudbridge and Koeleman, 2003). Yu et al. (2009) suggested using multiple candidate truncation points and proposed an adaptive rank truncated product method (ARTP). This method can efficiently and flexibly accumulate association evidence across SNPs and remove the subjectivity of choosing k to some extent. Instead of combining the k smallest *P*-values, Zaykin proposed a truncated product method (TPM) by using the product of only those *P*-values smaller than a specified threshold (Zaykin *et al.*, 2002). By choosing a proper cut-point such as 0.05, the TPM increases power. Recently, Jiang proposed a truncated tail strength statistic (TTS) (Jiang *et al.*, 2011), which was proved to have higher power than the tail strength method (Taylor and Tibshirani, 2006) and the FCT under certain scenarios. Chen proposed a sequential test (Chen *et al.*, 2013), which was shown to have almost the same power with the ARTP.

Taking a closer look at the truncated combined *P*-value procedures, we found that the *P*-values are divided into two groups and that the group with smaller *P*-values is used to construct the test statistics. To develop a more powerful test for identifying genetic association, our proposal is to divide the *P*-values into three or more groups. In each group, a test statistic is constructed. The test statistics are then combined to form an omnibus test. The proposed grouping strategy has sound basis. Biologically, SNPs in a certain genomic region can be divided into multiple blocks, with high LD within blocks and low LD between blocks. In one block, if there is a SNP associated with a disease, then it is likely that all SNPs in this block are associated with the disease because of LD. Statistically, such SNPs form the groups of interest.

In fact, any *P*-value combination procedures can be seen as different combinations using functions. The FCT uses $f(p_1,p_2,\ldots,p_m)=\prod_{i=1}^{m}p_i$, and the TPM employs $g(p_1,p_2,\ldots,p_m)=\prod_{i=1}^{m}p_i^{I(p_i\le\tau)}$. The ARTP can also be regarded as $h(p_1,p_2,\ldots,p_m)=\min_{1\le j\le J}\widehat{q}(k_j)$, where $\widehat{q}(k_j)$ is the estimated *P*-value for the statistic $\prod_{i=1}^{k_j}p_{(i)}$ and $k_j, j=1,\ldots,J$ are a set of prespecified candidates. Since there does not exist the uniformly optimal function of combining *P*-values. As an extension of the TPM with setting $J=1$ in our method, where the number of group is one, we use two functions log and the cumulative distribution function of two degrees of freedom to combine the *P*-values. When the *P*-values are independent and each *P*-value has the significance, the FCT is more powerful than the GCP, and if these *P*-values are correlated and few *P*-values show significances, the GCP is more powerful than the FCT. Extensive computer simulations show that the proposed test is more powerful than the existing combined tests while maintaining a good control of type I error rates. In the analysis of the Genetic Analysis Workshop 16 data, the proposed method can successfully detect the association between a genomic region covering gene DSC3 and rheumatoid arthritis, which might be missed by some other existing *P*-values combination methods.

2 Methods

Suppose that there are m *P*-values in a genetic association study, and these *P*-values are obtained from testing the associations between the m SNPs and phenotype. The existing methods such as the TPM and TTS divide them into two groups and combine the *P*-values in the group with smaller values. A natural extension is to divide the *P*-values into more than two groups and combine the *P*-values at the group level and form group-level test statistics.

Let $0<\xi_1<\xi_2<\cdots<\xi_J<1$ be the J thresholds. Then the group-combined *P*-value test statistic (for short, GCP) is constructed as follows

$$\mathrm{GCP}=\prod_{j=1}^{J}\left[1-F_j\left(\sum_{i=1}^{m}-2\ln p_i I_{\{\xi_{j-1}<p_i\le\xi_j\}}\right)\right],$$

where $\xi_0=0$, and $F_j(\cdot)$ is the cumulative distribution function of $\sum_{i=1}^{m}-2\ln p_i I_{\{\xi_{j-1}<p_i\le\xi_j\}}$ for $j=1,2,\cdots,J$.

Statistical properties of the GCP are established in the Supplemental Material. The GCP builds on single-marker analysis, which analyze one SNP from a panel of SNPs each time and combine them together. To calculate *P*-value of the GCP, we use the bootstrap procedure, where the correlations among SNPs in each replicate are kept. So it considers the correlations among SNPs and can be regarded as an alternative way to do multi-marker analysis. On the other hand, we can use the standard multi-marker analysis procedures such as LASSO (Tibshirani, 1996) or graphical LASSO (Friedman *et al.*, 2008) to select SNPs in a logistic regression model. To obtain the *P*-value of the GCP, one needs to conduct a multiple integration using the asymptotical distribution. Alternatively, we propose using the following algorithm to estimate the empirical *P*-value.

Algorithm:

Step 1: Set a large enough B, for example $B=10,000$. Calculate the *P*-values for each SNP using the observed data, denote it by $p_1^{(0)},\ldots,p_m^{(0)}$.

Step 2: For b from 1 to B, permutate individual trait values of the original data under the null hypothesis and calculate m *P*-values, denote them by $p_1^{(b)},\ldots,p_m^{(b)}$.

Step 3: Evaluate the empirical cumulative distribution of $F_j(\cdot)$ (denote it by $\widehat{F}_j(\cdot)$) for $j=1,2,\cdots,J$ using the *P*-values in Step 2.

Step 4: For b from 0 to B, calculate the GCP using $\widehat{F}_j(\cdot)$ $(j=1,2,\cdots,J)$ and $p_1^{(b)},\ldots,p_m^{(b)}$, denote it by GCP_b.

Step 5: The *P*-value of GCP is $\frac{\#\{\mathrm{GCP}_b>\mathrm{GCP}_0|b=1,2,\cdots,B\}}{B}$.

3 Results

3.1 Simulation settings

We conducted simulation studies to examine the performance of proposed GCP and to compare with the existing FCT, TPM, TTS and ARTP procedures. For the TPM, we chose 0.05 as the cut point. For the ARTP, the candidate points $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$, as recommended by Chen (Chen *et al.*, 2013), were used. We considered $m = 200$ and 500. The null hypothesis is that these m SNPs are not associated with the disease, and the alternative hypothesis is that at least one SNP is associated with the disease. Among these m SNPs, at most L are allowed to be associated with the phenotype, and $L\ (= m^{1-\gamma})$ is set following Mukherjee (Mukherjee *et al.*, 2015), where γ is used to determine whether the signals are dense or sparse ($\gamma \leq \frac{1}{2}$ corresponds to a dense region and $\gamma \geq \frac{1}{2}$ corresponds to a sparse region). Here we set $\gamma = 0.4$. To reduce computational cost, we generated m correlated P-values, corresponding to the significance levels of m SNPs, using the Gaussian-copula or t-copula directly, and the structure of the variance-covariance matrix is $\Lambda = \left(\rho^{|i-j|}\right)_{m\times m}$ for $0 \leq \rho \leq 1$. Under the null hypothesis, to generate P-values from a m-dimensional Gaussian-copula (Embrechts *et al.*, 2003), we first generated $\mathbf{X} \widehat{=} (X_1, X_2, \cdots, X_m)^\tau$ (where τ denotes the transpose of a vector or a matrix) from a m-dimensional normal distribution with mean $\mathbf{0}_m$ (a column vector with all elements being 0) and variance-covariance matrix Λ. That is $\mathbf{X} \sim N_m(\mathbf{0}_m, \Lambda)$. Then we obtained $(p_1, p_2, \ldots, p_m)$ by transforming each component in $\mathbf{X}$ by its marginal distribution. That is, $(p_1, p_2, \ldots, p_m) = (\Phi(X_1), \Phi(X_2), \ldots, \Phi(X_m))$, where $\Phi(\cdot)$ denotes the standard normal distribution function. For the t-copula, a similar procedure was implemented. Under the alternative hypothesis, we first generated the P-values following the above strategy and then transferred them to the p_i quantile of a Beta distribution with parameters 0.3 and 30. The probability density function of this distribution is shown in Figure 1. The transferred quantiles were then taken as the observations. We considered three values of ρ, including 0.2, 0.5 and 0.8. L takes value from $\{0, 1, 2, \cdots, 42\}$, where $L = 0$ corresponds to the null hypothesis. $L = 42$ is the maximum value for $m = 500$ considering $\gamma = 0.4$ and the maximum $L = 24$ for $m = 200$ is also in this set. The empirical power is calculated based on 10 000 replicates. The nominal significance level (α) is set as 0.05.

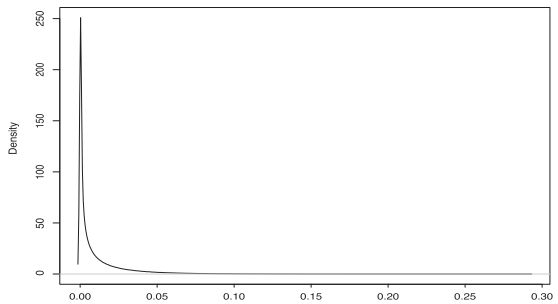

Fig. 1. The probability density function of Beta distribution with two parameters 0.3 and 30

3.2 Empirical type I error rates and power

3.2.1 Group number and threshold

It is a key to choose J and ξ_j, $j = 1, 2, \cdots, J$. An appropriate threshold is needed to obtain a powerful test. To assist choosing the proper threshold, we conducted simulations with several candidates and provided the results in Tables 1 and 2. The optimal selection of the groups and the thresholds depends on the P-values structure and therefore the structure of the data. In Table 1, it can be seen that in the majority of the considered scenarios, the GCP with $J = 2$ and $\xi_1 = 0.001$, $\xi_2 = 0.05$ perform better when $\rho = 0.5, 0.8$ for $m = 200$, but it with $J = 3$ and $\xi_1 = 0.001$, $\xi_2 = 0.01$ and $\xi_3 = 0.05$ works better when $\rho = 0.2$ and the number of markers associated with the diseases is more than 3. So the choices of $J = 2$ and $J = 3$ are both appropriate for a low linkage disequilibrium scenario. To reduce the computational complexity, we recommend $J = 2$. Synthesizing all the results in Tables 1 and 2, it can be seen that the performance for $J = 2$ is better than $J = 3$ in majority of the scenarios. As the uncertainty of the number of markers associated with the diseases in the real applications, from the standpoint of robustness, we also recommend the choice of $J = 2$. Next we use the idea of the degrees of freedom (DF) to explain why we chose $J = 2$. It is known that when a

Table 1. The empirical type I error rates ($L = 0$) and powers of the GCP under different thresholds when $m = 200$ and the data were generated using Gaussian-copula. The first panel is for $\rho = 0.2$, the second panel is for $\rho = 0.5$ and the last panel is for $\rho = 0.8$

L	0	1	2	3	4	5	6	7	8
$\xi_1 = 0.01,\ \xi_2 = 0.1$	0.051	0.122	0.257	0.420	0.601	0.742	0.853	0.928	0.970
$\xi_1 = 0.001,\ \xi_2 = 0.05$	0.048	0.225	0.447	0.599	0.748	0.845	0.913	0.950	0.972
$\xi_1 = 0.01,\ \xi_2 = 0.05$	0.047	0.120	0.262	0.460	0.629	0.775	0.893	0.943	0.974
$\xi_1 = 0.001,\ \xi_2 = 0.01,\ \xi_3 = 0.1$	0.046	0.196	0.395	0.585	0.739	0.848	0.915	0.956	0.977
$\xi_1 = 0.001,\ \xi_2 = 0.01,\ \xi_3 = 0.05$	0.050	0.186	0.406	0.610	0.761	0.860	0.924	0.964	0.978
L	0	1	2	3	4	5	6	7	8
$\xi_1 = 0.01,\ \xi_2 = 0.1$	0.046	0.088	0.176	0.273	0.448	0.579	0.692	0.804	0.862
$\xi_1 = 0.001,\ \xi_2 = 0.05$	0.049	0.177	0.362	0.500	0.621	0.727	0.806	0.859	0.901
$\xi_1 = 0.01,\ \xi_2 = 0.05$	0.042	0.103	0.181	0.308	0.484	0.615	0.718	0.814	0.869
$\xi_1 = 0.001,\ \xi_2 = 0.01,\ \xi_3 = 0.1$	0.055	0.146	0.311	0.458	0.590	0.718	0.788	0.854	0.909
$\xi_1 = 0.001,\ \xi_2 = 0.01,\ \xi_3 = 0.05$	0.048	0.147	0.289	0.462	0.625	0.708	0.812	0.866	0.919
L	0	1	2	3	4	5	6	7	8
$\xi_1 = 0.01,\ \xi_2 = 0.1$	0.049	0.070	0.094	0.140	0.179	0.266	0.325	0.394	0.471
$\xi_1 = 0.001,\ \xi_2 = 0.05$	0.045	0.098	0.179	0.296	0.418	0.482	0.545	0.601	0.651
$\xi_1 = 0.01,\ \xi_2 = 0.05$	0.045	0.067	0.096	0.143	0.175	0.239	0.290	0.385	0.453
$\xi_1 = 0.001,\ \xi_2 = 0.01,\ \xi_3 = 0.1$	0.053	0.085	0.156	0.216	0.304	0.378	0.470	0.569	0.629
$\xi_1 = 0.001,\ \xi_2 = 0.01,\ \xi_3 = 0.05$	0.049	0.091	0.157	0.202	0.305	0.378	0.504	0.569	0.623

L denotes the number of disease-associated SNPs.

Table 2. The empirical type I error rates ($L=0$) and powers of the GCP under different thresholds when $m=500$ and the data were generated using Gaussian-copula. The first panel is for $\rho=0.2$, the second panel is for $\rho=0.5$ and the last panel is for $\rho=0.8$

L	0	1	2	3	4	5	6	7	8
$\xi_1=0.01,\ \xi_2=0.1$	0.043	0.098	0.167	0.253	0.387	0.497	0.631	0.731	0.811
$\xi_1=0.001,\ \xi_2=0.05$	0.050	0.146	0.267	0.453	0.588	0.680	0.778	0.858	0.906
$\xi_1=0.01,\ \xi_2=0.05$	0.046	0.099	0.164	0.267	0.397	0.509	0.646	0.745	0.827
$\xi_1=0.001,\ \xi_2=0.01,\ \xi_3=0.1$	0.047	0.131	0.246	0.405	0.560	0.647	0.765	0.846	0.9
$\xi_1=0.001,\ \xi_2=0.01,\ \xi_3=0.05$	0.048	0.131	0.243	0.417	0.569	0.663	0.772	0.854	0.906
L	0	1	2	3	4	5	6	7	8
$\xi_1=0.01,\ \xi_2=0.1$	0.051	0.080	0.120	0.197	0.267	0.372	0.457	0.547	0.634
$\xi_1=0.001,\ \xi_2=0.05$	0.050	0.115	0.227	0.375	0.473	0.569	0.652	0.728	0.789
$\xi_1=0.01,\ \xi_2=0.05$	0.050	0.076	0.123	0.201	0.273	0.392	0.474	0.562	0.657
$\xi_1=0.001,\ \xi_2=0.01,\ \xi_3=0.1$	0.050	0.110	0.184	0.315	0.431	0.531	0.606	0.691	0.772
$\xi_1=0.001,\ \xi_2=0.01,\ \xi_3=0.05$	0.048	0.102	0.187	0.324	0.434	0.544	0.617	0.710	0.784
L	0	1	2	3	4	5	6	7	8
$\xi_1=0.01,\ \xi_2=0.1$	0.046	0.066	0.082	0.100	0.132	0.169	0.191	0.244	0.314
$\xi_1=0.001,\ \xi_2=0.05$	0.047	0.070	0.130	0.207	0.290	0.350	0.399	0.468	0.515
$\xi_1=0.01,\ \xi_2=0.05$	0.045	0.065	0.082	0.104	0.135	0.175	0.193	0.251	0.314
$\xi_1=0.001,\ \xi_2=0.01,\ \xi_3=0.1$	0.047	0.072	0.107	0.153	0.208	0.263	0.296	0.381	0.443
$\xi_1=0.001,\ \xi_2=0.01,\ \xi_3=0.05$	0.046	0.071	0.106	0.155	0.207	0.262	0.295	0.393	0.449

L denotes the number of disease-associated SNPs.

Table 3. The empirical type I error rates of the FCT, the TPM, the TTS, the ARTP and the GCP

	$\rho=0.2$		$\rho=0.5$		$\rho=0.8$	
m	200	500	200	500	200	500
FCT	0.048	0.048	0.048	0.049	0.042	0.051
TPM	0.056	0.049	0.048	0.048	0.044	0.052
TTS	0.051	0.051	0.049	0.049	0.044	0.050
ARTP	0.045	0.050	0.052	0.056	0.047	0.047
GCP	0.051	0.045	0.047	0.051	0.050	0.051
	$\rho=0.2$		$\rho=0.5$		$\rho=0.8$	
m	200	500	200	500	200	500
FCT	0.050	0.049	0.053	0.052	0.055	0.049
TPM	0.052	0.049	0.049	0.053	0.050	0.049
TTS	0.057	0.049	0.048	0.050	0.052	0.050
ARTP	0.048	0.050	0.053	0.050	0.050	0.046
GCP	0.050	0.048	0.050	0.052	0.053	0.051

The first panel is for Gaussian-copula and the second panel is for t-copula.

random variable X follows the uniform distribution on $[0,1]$, $-2\ln(X)$ follows the Chi-squared distribution with 2 *DF*s. So, when $J=2$, the pseudo *DF*s of the final test statistic could be 4 and when $J=3$, it increases to 6. So the pseudo *DF*s might increase with the increasing of the number of groups. And also the computation is intensive when J is large. To reduce the pseudo *DF* and computational intensity, we adopt $J=2$. The form of the used GCP is as follows.

$$\mathrm{GCP}=\left[1-F_1\left(\sum_{j=1}^{m}-2\ln p_j I_{\{0<p_j\le 0.001\}}\right)\right]$$

$$\left[1-F_2\left(\sum_{j=1}^{m}-2\ln p_j I_{\{0.001<p_j\le 0.05\}}\right)\right].$$

3.2.2 Methods comparison

The empirical type I error rates are shown in Table 3. It can be seen that all of the methods considered can control the type I error rates properly. The empirical values are very close to the nominal level of 0.05. For example, when $\rho=0.2$, $m=200$ and the data are generated from the Gaussian-copula, the empirical type I error rates of the FCT, TPM, TTS, ARTP and GCP are 0.048, 0.056, 0.051, 0.045 and 0.051, respectively. When $\rho=0.8$, $m=500$ and the data are generated using the t-copula, the values of power of the FCT, TPM, TTS, ARTP and GCP are, respectively, 0.049, 0.049, 0.050, 0.046 and 0.051.

Figures 2, 3 and 4 show the empirical power of the FCT, TPM, TTS, ARTP and GCP for $\rho=0.2$, $\rho=0.5$ and $\rho=0.8$, respectively. It can be easily seen that the GCP is much more powerful than the alternatives under most of the simulated scenarios. Sometimes, the improvement can be 50% or more. For example, when $\rho=0.2$, $m=200$, $L=3$, and the data are generated using the t-copula, the values of power of the FCT, TPM, TTS, ARTP and GCP are 0.069, 0.064, 0.056, 0.159 and 0.736, respectively. The improvement of the GCP over the alternatives is smaller under the Gaussian-copula than the t-copula. To explain why there are large power difference between the Gaussian-copula and t-copula, we have provided the

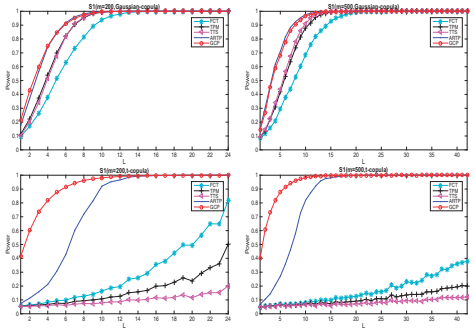

Fig. 2. The empirical powers of the FCT, the TPM, the TTS, the ARTP and the GCP for $m = 200$ or 500. The data were generated using Gaussian-copula/t-copula with the variance and covariance matrix Λ and $\rho = 0.2$. L denotes the number of diseased-associated SNPs

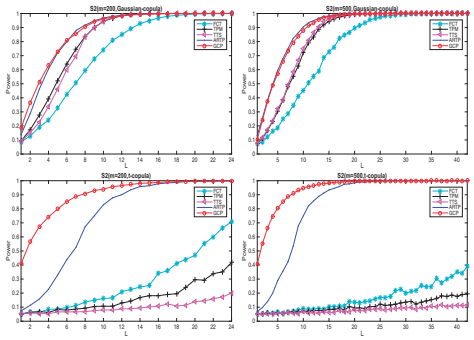

Fig. 3. The empirical powers of the FCT, the TPM, the TTS, the ARTP and the GCP for $m = 200$ or 500. The data were generated using Gaussian-copula/t-copula with the variance and covariance matrix Λ and $\rho = 0.5$. L denotes the number of diseased-associated SNPs

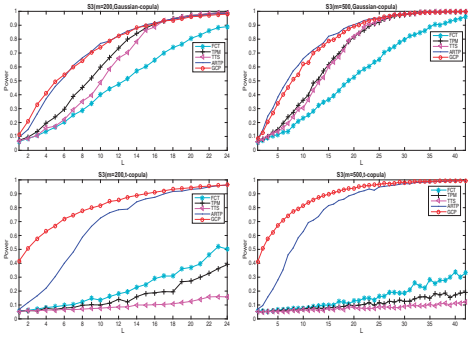

Fig. 4. The empirical powers of the FCT, the TPM, the TTS, the ARTP and the GCP for $m = 200$ or 500. The data were generated using Gaussian-copula/t-copula with the variance and covariance matrix Λ and $\rho = 0.8$. L denotes the number of diseased-associated SNPs

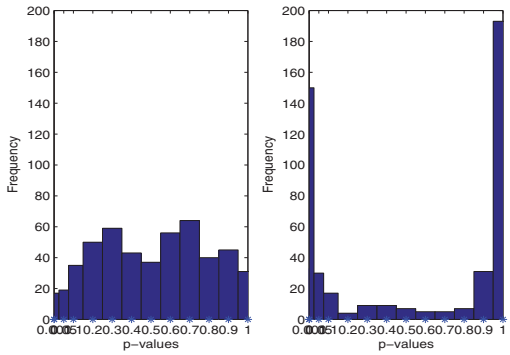

Fig. 5. The histogram of P-values when $m = 500$ and $L = 4$ under the alternative hypothesis. The left figure is for Gaussian-copula, and the right figure is for t-copula

histogram of the P-values under the alternative hypothesis in Figure 5. It can be seen that the number of P-values in two groups is very different under Gaussian-copula. In contrast, under t-copula, the two groups both have a certain amount of P-values and there is slight difference between them. It may result in the more difference of power between the GCP and other methods under t-copula than Gaussian-copula. It is also observed that when ρ increases, the difference between the GCP and alternatives decreases. This is also sensible. Consider the extreme case when $\rho = 1$, then all of the considered tests should have the same power.

3.3 Application to detect the association between a genomic region covering gene DSC3 and rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic systemic diseases and the symptom is mainly based on inflammatory synovitis. RA has been found to be associated with genetic variants (Ellinghaus et al., 2011). Our goal is to examine whether there is an association between a genomic region covering gene DSC3 and RA, and the region has been reported to be associated with RA (Zhang *et al.*, 2009). The data from the Genetic Analysis Workshop 16 (GAW16) where the data contain records on 2062 subjects was used to get to this aim. We construct a gene region with 133 SNPs by extending the gene DSC3 before and after each with 50 SNPs, respectively. After the quality control of removing SNPs with the minor allele frequencies being less than 0.05 and missing rate being larger than 5%, there were 129 SNPs remained for our analysis. To account for the possible confounding of population stratification, we matched 12 898 structure inference SNPs (Yu *et al.*, 2008) with low local background LD with the data of GAW16 and selected 12 749 SNPs to construct four principal coordinates using the multi-dimensional scaling method (Li and Yu, 2008). We applied the Wald test based on the logistic regression model to obtain P-value for each SNP. The *ID* information is provided in the Supplemental Material. Then we applied the FCT, the TPM, the TTS, the ARTP and the GCP to analyze these data. 10 000 permutations were conducted to get the P-values of these tests. The P-values of the FCT, the TPM, the TTS, the ARTP and the GCP are 0.0952, 0.0608, 0.0732, 0.0236 and 0.0222, respectively. Using the ARTP and the GCP give the significance of the association under the nominal significance level of 0.05, while other methods miss this finding, and the GCP gives smaller P-values than the ARTP.

4 Discussion

In large-scale genetic association studies, usually tens of hundreds of SNPs are genotyped and tested. The traditional statistical framework, such as the logistic regression, needs to solve a multidimensional optimization problem on odds ratios. It is difficult to find the global solution, and sometimes the solution is not stable, or there may be multiple solutions, when the SNPs are in high LD. To tackle this problem, we have developed a GCP method. This method is built on single-marker analysis and can be easily applied to testing a large number of SNPs. Advancing from the existing alternatives, the

GCP integrates the information significance at different levels and improves power.

The proposed GCP can have applications far beyond this article. As the first example, we consider the analysis of different complex human diseases, which can be 'correlated'. Simultaneously analyzing genetic variants associated with multiple correlated diseases or traits can help investigators understand the genetic architecture of diseases and improve the power to identify deleterious genetic variants (Aschard *et al.*, 2014; Li *et al.*, 2014; Solovieff *et al.*, 2013). In such analysis, quite often more than 100 phenotypes and a genetic variant are studied at a time (Aschard *et al.*, 2014). The proposed GCP can be directly applied to this example. In addition, the proposed GCP can also be used to detect the associations between multiple genetic variants and multiple phenotypes. One can first use a multistep combined principal component method (Aschard *et al.*, 2014) or multivariate linear mixed model (Zhou and Stephens, 2014) to obtain the *P*-value for the association between each variant and multiple phenotypes and then employ the GCP to combine them.

To obtain the *P*-value under the GCP, potentially, a two-layer resampling procedure needs to be conducted. First, an inner layer of bootstrap or permutation procedure is required to obtain the distribution of test statistic at the group level. Then a second layer of bootstrap or permutation procedure is needed to obtain the distribution of the GCP. This can be computational intensive, especially, when the nominal significance level is small. One needs to conduct millions of bootstrap or permutation steps to obtain a valid *P*-value. To overcome this obstacle, we proposed a one-layer permutation procedure to obtain the *P*-value.

In this work, the simulations and data analysis have been focused on binary traits. It is noted that the GCP is flexible and can be applied to other types of phenotypes in large-scale genetic studies. If the trait is quantitative, we can first use the linear regression model to obtain the *P*-value for each SNP when the normal assumption holds, and then combine them together. When the normal distribution does not hold (even after transformation), one can use the nonparametric trend tests (Li *et al.*, 2013; Schaid *et al.*, 2005; Zhang and Li, 2015) to analyze one SNP at a time with or without considering the genetic models and then employ the proposed method to construct the omnibus test. If the phenotype is ordinal, the score test derived from the proportional odds model (McCullagh, 1980) can be first used to obtain the *P*-value for each SNP, and then the *P*-values can be combined using the proposed method.

From the simulations, the power difference between the GCP and other methods may be attributable to the number of markers associated with the diseases. For the dependence on the data structure, the GCP can obtain more power than other methods when each group has a certain amount of *P*-values. The distributions of *P*-values have been shown in Figure 5 and Figure S1 in the Online Supplementary Material. While, as the increasing of the number of associated markers, the power of all the methods will increase and then the power difference between the GCP and others may reduce. The simulation results also show that the GCP performs better than the FCT, TPM and TTS consistently. Towards the ARTP, the GCP has obtained a uniformly higher power when the data are generated using t-copula. For the Gaussian-copula, the GCP and ARTP have its own 'sweet' spots. But overall, the power difference between the GCP and ARTP is larger for t-copula than that for Gaussian-copula, please see Figures 2–4. Therefore, our method is more robust than the ARTP if the assumption of normality is not considered.

However, the proposed method has some drawbacks. First, as difficulty of deriving the exact distribution of the GCP, we need to use empirical cumulative distribution to obtain its *P*-value. In a large-scale genetic association study, for example, when m is large, say a million or bigger, the computation can be very intensive. Also, the distribution function of the GCP cannot be computed. Second, theoretically, the number of groups (J) does not depend on the number of SNPs. However, in practice, we don't know J, the data-driven selection of J might be slightly correlated with the number of SNPs. Third, the proposed method is constructed based on the individual *P*-values, which can be obtained by single-marker analysis. For the selection of the thresholds, we proposed to use 0.001 and 0.05. They are not optimal and depend on the structure of the data. To overcome it, one possible scheme is to adopt some alternative thresholds and use the maximum of the GCP on these thresholds as the final test statistics and use the bootstrap or permutation to assess its statistical significance.

Funding

This work was supported by the National Institutes of Health [NO1-AR-2-2263, RO1-AR-44422, Peter K. Gregersen, PI]; and the National Arthritis Foundation. The use of the data was approved by GAW16. Q. Li was supported in part by the National Science Foundation of China [Grant No. 11371353, 61134013]; and the Breakthrough Project of Strategic Priority Program of the Chinese Academy of Sciences [Grant No. XDB13040600].

Conflict of Interest: none declared.

References

Aschard,H. *et al.* (2014) Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.*, **94**, 662–676.

Chen,H.S. *et al.* (2013) A powerful method for combining p-values in genomic studies. *Genet. Epidemiol.*, **37**, 814–819.

Dudbridge,F. and Koeleman,B.P. (2003) Rank truncated product of P-values, with application to genomewide association scans. *Genet. Epidemiol.*, **25**, 360–366.

Ellinghaus,E. *et al.* (2011) Genome-wide meta-analysis of psoriatic arthritis identifies susceptibility locus at REL. *J. Invest. Dermatol.*, **132**, 1133–1140.

Embrechts,P. *et al.* (2003) Modelling dependence with copulas and applications to risk management. In: Rachev,S.T. (ed) *Handbook of Heavy Tailed Distributions in Finance*. Elsevier, North-Holland.

Fisher,R.A. (1932) *Statistical Methods for Research Workers*. 4th edn. Oliver and Boyd, London.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Hess,A. and Iyer,H. (2007) Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *BMC Genomics*, **8**, 96.

Jiang,B. *et al.* (2011) A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *J. Theor. Biol.*, **277**, 67–73.

Li,Q. *et al.* (2014) Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. *Biostatistics*, **15**, 284–295.

Li,Q. and Yu,K. (2008) Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet. Epidemiol.*, **32**, 215–226.

Li,Q. *et al.* (2013) Rank-based robust tests for quantitative-trait genetic association studies. *Genet. Epidemiol.*, **37**, 358–365.

McCullagh,P. (1980) Regression models for ordinal data. *J. R. Stat. Soc. B*, **42**, 109–142.

Mukherjee,R. *et al.* (2015) Hypothesis testing for high-dimensional sparse binary regression. *Ann. Stat.*, **43**, 352–381.

Schaid,D.J. *et al.* (2005) Nonparametric tests of association of multiple genes with human disease. *Am. J. Hum. Genet.*, **76**, 780–793.

Solovieff,N. *et al.* (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, 483–495.

Taylor,J. and Tibshirani,R. (2006) A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, **7**, 167–181.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Yu,K. *et al.* (2008) Population substructure and control selection in genome-wide association studies. *PLoS One*, **3**, e2551.

Yu,K. *et al.* (2009) Pathway analysis by adaptive combination of p-values. *Genet. Epidemiol.*, **33**, 700–709.

Zaykin,D.V. *et al.* (2002) Truncated product method for combining P-values. *Genet. Epidemiol.*, **22**, 170–185.

Zhang,W. and Li,Q. (2015) Nonparametric risk and nonparametric odds in quantitative genetic association studies. *Sci. Rep.-UK*, **5**, 12105.

Zhang,M. *et al.* (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**, S17,

Zheng,G. *et al.* (2012) Joint analysis of binary and quantitative traits with data sharing and outcome-dependent sampling. *Genet. Epidemiol.*, **36**, 263–273.

Zhou,X. and Stephens,M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.