European Journal of Human Genetics (2014) 22, 148–151

www.nature.com/ejhg

SHORT REPORT

A modified two-stage approach for family-based genome-wide association studies

Weijun Ma[1], Ying Zhou*[,1], Yajing Zhou[1], Lili Chen[1] and Zhen Gu[2]

Genome-wide association studies can provide researchers some reference on gene mapping of complex trait, a key point of which is how to improve the power of association test. Recently, two-stage approaches are widely used to genome-wide association analysis. In the first stage, a screening test is used to select markers, and in the second stage, a family-based association test is performed based on a smaller set of the selected markers. Here, we modify an existing two-stage approach and propose a new test statistic for the association analysis. Simulation studies are conducted to compare the type I error rates and powers of the proposed approach with those of the existing two-stage approaches. Simulation results show that the new two-stage approach has greater power than the other two-stage approaches to some extent.
European Journal of Human Genetics (2014) **22,** 148–151; doi:10.1038/ejhg.2013.105; published online 22 May 2013

Keywords: association test; complex disease; FDR; gene mapping; two-stage approach

INTRODUCTION

Association analysis for single-gene disease is easy to perform in general. But if the target trait is controlled by multiple loci (eg, complex diseases), the genes are generally more difficult to detect in practice. It can be said that the methods for detecting loci of complex trait are not enough so far, and the powers of the existing methods are relatively limited. Seeking powerful methods in current association analysis is an important issue. Aiming at complex trait, multiple tests are usually involved in association analysis, and how to control type I error rate in analysis is its critical step. There are many marker loci in human genome. When the genome-wide association analysis being conducted, the general approaches for controlling type I error rate may be more conservative, which will lead to limited power.[1]

Recently, many researchers are committed to how to improve the power of test in genome-wide association analysis. In this respect, two-stage and multi-stage approaches have been proposed in succession,[2–9] and detecting loci of interest is divided into different stages. The two-stage approaches proposed in Steen *et al*[7] and Feng *et al*[8] are more representative. Steen *et al*[7] proposed a two-stage approach for family-based genome-wide association study. In the first stage, a screening test is used to select markers, and in the second stage, a family-based association test is performed based on a smaller set of the selected markers. The two-stage approach is more powerful than the traditional family-based association tests. Feng *et al*[8] extended the approach so that the test statistic can incorporate parental information and can be applied to arbitrary pedigree structure. Their results show that the two-stage approach that incorporates phenotypes of the founders has correct type I error rates, and is more powerful than the two-stage approach that only incorporates information of children. Motivated by Feng *et al*'s method, Gu[9] proposed a new test approach (ie, TTFPBSA). The difference between the two methods is that the TTFP proposed by Feng *et al*[8] uses information of the founders in the pedigrees in the first stage, but not in the second stage, while the TTFPBSA proposed by Gu[9] incorporates information of founders into both the screening test and the association test. Although more information is used in Gu's method, this operation will lead to the dependence of the two test statistics in the two stages, and therefore affect the power of the two-stage approach.[10,11]

In this paper, based on Steen *et al*'s idea,[7] we propose a new two-stage approach, by modifying the existing statistics for family-based association study. In the new approach, the information of the founders in the pedigrees is incorporated into the screening statistic in the first stage, but not into the association statistic in the second stage, to guarantee the independence between them. Our simulation results suggest that the proposed method performs well in the power of the association test, and outperforms current methods.

METHODS

Suppose there are n pedigrees, and M SNP loci in their genome can be genotyped, with the alleles denoting by 0 and 1 each.

We use the following notation: N_i, number of nuclear families in the ith pedigree; n_{ij}, number of children in the jth nuclear family of the ith pedigree; Y_{ijF}, Y_{ijM} and Y_{ijk}, respectively, denote the trait value of the father, the mother and the kth child in the jth nuclear family of the ith pedigree; X_{ijF}, X_{ijM} and X_{ijk}, respectively, denote the genetic score of the father, the mother and the kth child in the jth nuclear family of the ith pedigree; $\overline{Y}$, the average trait value of all the individuals; and $\bar{X}$, the average genotypic score of all the individuals.

Stage I: screening test

A screening test is employed to select L significant markers among the M marker loci. For each of the M markers, we test the null hypothesis H_0: no association.

The screening test statistic we used is given by

$$T_{\text{screen}} = \frac{\sum_{i=1}^{n} U_i}{\sqrt{\sum_{i=1}^{n} U_i^2}},$$

[1]Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin, China; [2]Department of Basic Research, East University of Heilongjiang, Harbin, China
*Correspondence: Dr Y Zhou, Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China. Tel: +86 451 88197423; Fax: +86 451 86604399; E-mail: yzhou@aliyun.com
Received 1 May 2012; revised 24 March 2013; accepted 19 April 2013; published online 22 May 2013

where

$$U_i = \sum_{j=1}^{N_i} U_{ij},$$

and

$$U_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y})(\bar{X}_{ij} - \bar{X}) + (Y_{ijF} - \bar{Y})(X_{ijF} - \bar{X}) I_{ijF} + (Y_{ijM} - \bar{Y})(X_{ijM} - \bar{X}) I_{ijM}$$

where $I_{ijF} = 1$, if the father of the jth family in the ith pedigree is a founder of this pedigree, and $I_{ijF} = 0$, otherwise; I_{ijM} is similarly defined for the mother; $\bar{X}_{ij} = (X_{ijF} + X_{ijM})/2$, if parental genotypes are available, and $\bar{X}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$, otherwise.

In fact, in this stage we choose the same test statistic as the one in Feng *et al*.[8] Under the null hypothesis of no association, the screening test statistic follow a standard normal distribution. From the M tests, we obtain L markers with the smallest P-values, where L is a pre-specified number, and we will discuss the value of L later.

Stage II: association test

We conduct multiple tests for the L selected markers, thus we can further detect the gene loci associated with the target trait.

The new association test statistic we proposed is as follows

$$T_{\text{association}} = \frac{\sum_{i=1}^{n} V_i}{\sqrt{\sum_{i=1}^{n} V_i^2}}$$

where

$$V_i = \sum_{j=1}^{N_i} V_{ij},$$

$$V_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij})(X_{ijk} - \bar{X}_{ij})$$

and

$$\bar{X}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}, \ \bar{Y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}.$$

Under the null hypothesis of no association, the association statistic $T_{\text{association}}$ asymptotically follows a standard normal distribution. We use the association statistic to test each of the L selected markers and declare a marker is significant at a level of α, if the P-value of the $T_{\text{association}}$ at this marker is less than the threshold $\delta_{L\alpha}$, which is determined by the procedures for controlling the false discovery rate (FDR).[8,12]

To control the FDR at a level of α, the cutoff can be chosen $\delta_{L\alpha}$ as follows: let $p_{(1)}, p_{(2)}, \ldots, p_{(L)}$ be the ordered P-values when we apply the $T_{\text{association}}$ to the L selected markers, then $\delta_{L\alpha} = \max\{ p_{(i)}: p_{(i)} \leq i\alpha/L\}$.

The new approach is called RTTFP. Note that the two test statistics in the two stages has the relationship between covariance between groups and covariance within groups, which is similar to the relationship between sum of squares between groups and sum of squares within groups in analysis of variance. Therefore, the two test statistics are independent to each other, and correspondingly the two-stage tests are independent to each other.

SIMULATION STUDIES

Simulation design

We randomly generated genotype and phenotype data of pedigrees as given in Figure 1. For the data, we considered the type I error rates and the powers of the three methods: the TTFP, the TTFPBSA and the RTTFP. At the same time, to demonstrate it is not suitable to incorporate information of the founders in the second stage of the TTFP, we also compare the statistic in which founder information are incorporated in the second stage of the TTFP. We named the method by TTFFP.

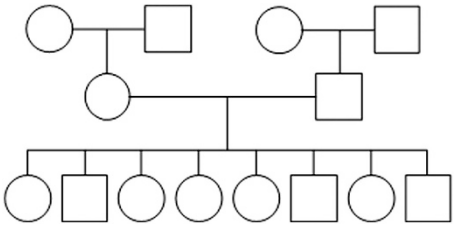

Figure 1 The pedigree structure used in the simulation studies.

To better compare these four methods, we apply the similar sampling design in Feng *et al*.[8] To assess the type I error rates, we generate data under the null hypothesis of no association. First, we generate genotypes under the Hardy–Weinberg equilibrium and linkage equilibrium. This means that we generate each allele and each marker independently. The frequency of the minor allele at each marker is randomly sampled between 0.1 and 0.4. Under these conditions, we can generate genotype data of all individuals for 50 pedigrees. Second, we generate phenotype data of all individuals. Let $Y_1 = (y_F, y_M)$ denote the trait values of the parents and $Y_2 = (y_1, y_2, \ldots, y_m)$ denote the trait values of the m children. Assume (Y_1, Y_2) follow a normal distribution with a mean vector of zero and variance–covariance matrix of

$$\Sigma = \begin{pmatrix} 1 & 0 & \rho & \ldots & \rho \\ 0 & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \ldots & 1 \end{pmatrix}$$

It is easy to see from the above variance–covariance matrix that y_F and y_M are independent, parents with children and children with children are correlated with the correlation coefficient ρ. We can first generate data of Y_1, and then generate data of Y_2 conditional on the values of Y_1. In our simulations, we also set different cases of heritability ($h^2 = 0.03, 0.05, 0.07$).

To ascertain the appropriate number of the selected loci in the first stage, we vary the value of L from 1 to M ($M = 100, 1000$), and compute the corresponding FDR (or power) for each test. In each scenario, we use the average of 1000 replications to estimate the FDR for each method. The analysis results are shown in Figures 2–4 (for each M, see the sub-figure about FDR).

For power comparisons, we need to regenerate pedigree data under H_1. We generate genotype data as described previously and then the trait value of the kth member is given by the linear model below:

$$y_k = \beta x_k + \varepsilon_k,$$

where the value of β here is determined by the value of heritability h^2 and disease models.

After the simulated data of 50 pedigrees are generated, we respectively compute the test powers by the four methods, and use the average of 1000 replications to estimate the power for each method. The corresponding results of the powers are shown in Figures 2–4 (for each M, see the sub-figure about power).

Evaluation on type I error rates

The sub-figures about FDR in Figures 2–4 show that when significant level α is 0.05, the type I error rates of the TTFP, the TTFPBSA and the RTTFP can be basically controlled within a reasonable range, however, the type I error rates of the TTFFP cannot be controlled.

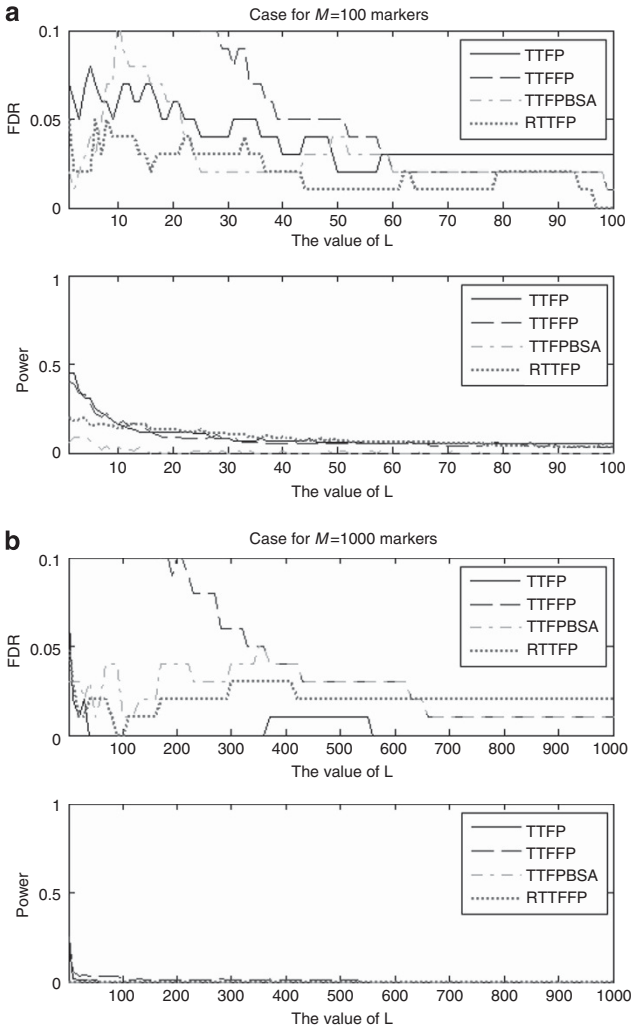

Figure 2 Type I error rates and powers of the four methods for pedigree data ($h^2 = 0.03$). Note: TTFP, the method proposed by Feng *et al*;[8] TTFFP, the method in which founder information are incorporated in the second stage of the TTFP; TTFPBSA, the method proposed by Gu;[9] and RTTFP, the new method proposed in this paper.

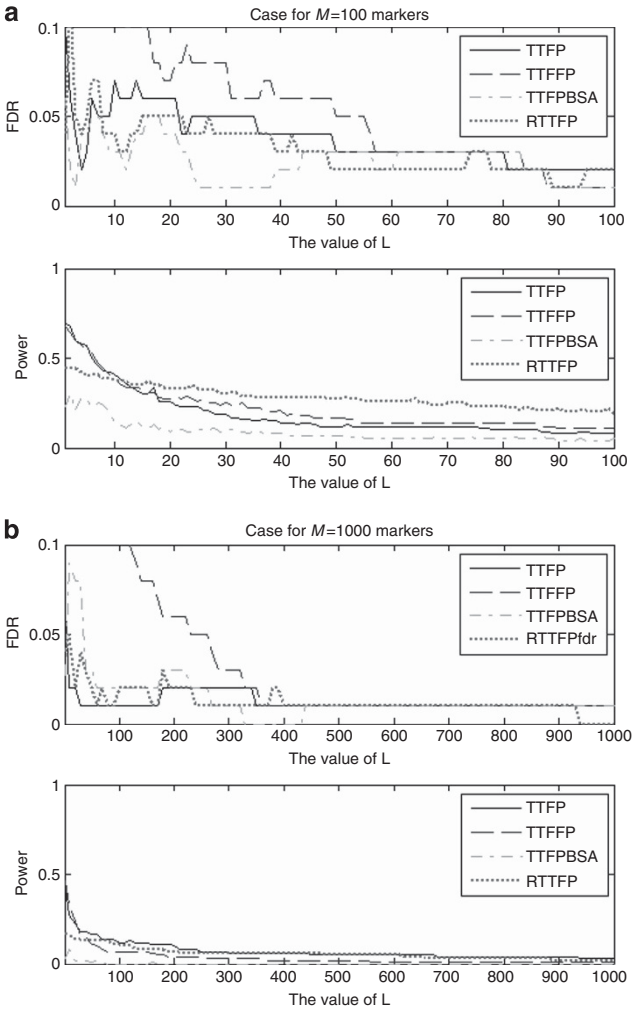

Figure 3 Type I error rates and powers of the four methods for pedigree data ($h^2 = 0.05$). See note in Figure 2 for abbreviation details.

Comparing with the case of $M = 100$, the values of FDRs for each method decrease when $M = 1000$ to some extent. Besides, when $M = 100$ and the value of L is < 20, the TTFP gives larger FDRs in fact. In addition, these figures also report that the FDRs of each method show similar trend for different cases of heritability.

Evaluation on powers

The sub-figures about power in Figures 2–4 clearly show that in most cases our RTTFP is more powerful among the four methods. Although the TTFP has a little higher powers for smaller L, it is accompanied with larger FDRs at the same time. The powers of the TTFPBSA are the lowest among all methods, so the method does not obtain higher powers on the basis of the TTFP. The powers of the TTFFP are very close to the ones of the TTFP. Unfortunately, however, the TTFFP could not control type I error rate, so it is difficult to use in practice. The powers of all methods show the trend of decrease with increasing of L. However, by contrast the powers of the RTTFP decrease stationarily as L increase.

Feng *et al*[8] considered the choice for L and suggested that a value between 10 and 20 is a good choice for L. From our simulation results, we obtain the similar conclusion. Figures 2–4 show that the RTTFP is more powerful when L equals to 10 or so.

The simulation results also show that the heritability is also an important factor that impacts the association test. Figures 2–4 show that powers of various methods gradually increase as the heritability increases. At the same time, in each heritability case (eg, $h^2 = 0.07$, see Figure 4), the powers of each method all decrease correspondingly with the value of M increasing from 100 to 1000. It is as expected, because identifying more markers for the same sample size in association analysis is more difficult from statistical viewpoint.

Besides, simulation results for data of nuclear families are similar to those for the pedigree data. In fact, the nuclear family situation is nothing but a special case of the pedigree structure, therefore, the corresponding conclusions should be consistent.

The proposed method can be applied to the trios data, in which case the two-step method will be reduced into one-step method (screening test). Through choosing appropriate value of L, we can also identity those significant trait loci. Of course, more children in each observed family will provide more association information, and therefore improve the power of test.

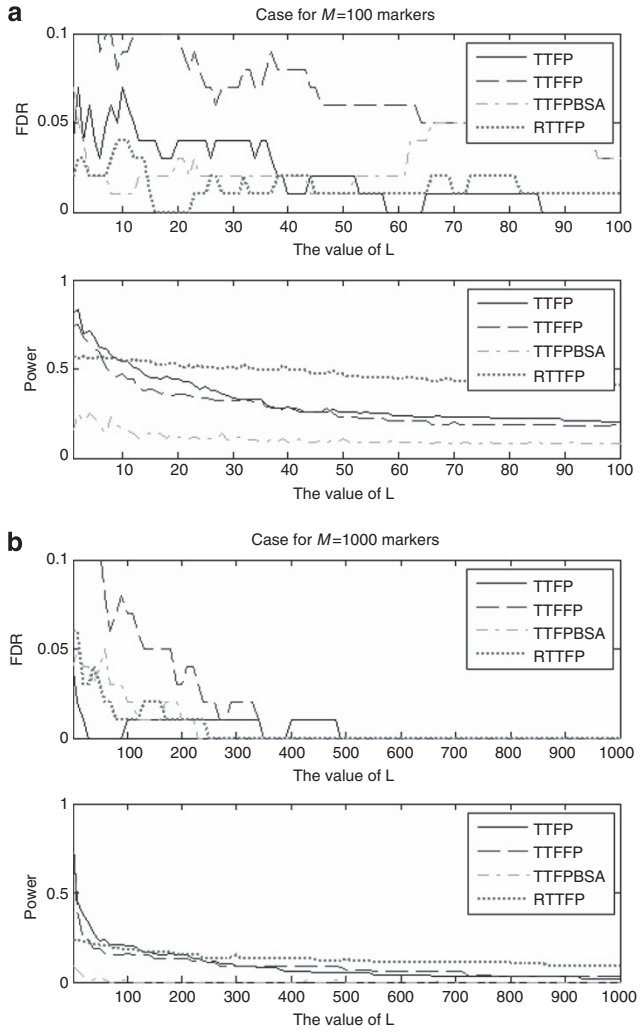

Figure 4 Type I error rates and powers of the four methods for pedigree data ($h^2 = 0.07$). See note in Figure 2 for abbreviation details.

DISCUSSION

Currently, researchers have made much progress in genome-wide association studies which provides certain reference for gene mapping of complex disease. To overcome the impact of population stratification on the association test, researchers have proposed many different methods. For example, one-stage approach, two-stage approach, and so on. Among them, one-stage approach tests all markers simultaneously, the power of which is generally not high. Therefore, people increasingly take advantage of the two-stage approaches in family-based association studies.[7,8]

The idea of adding information of the founders into a statistic seems feasible in association analysis, however, how to construct the new statistic is crucial. In this paper, we present a new association test approach which is an improvement on the two-stage method in literature.[8] It is fit for the gene mapping of complex trait in family-based genome-wide analysis. The new method is more powerful than the current methods to some extent, because we not only ensure the independence between two step tests, but also we reasonably use all the data information.

At the same time, we demonstrate the method proposed by Gu.[9] Although she seems to use more data information, the actual results of test are unsatisfactory. One main reason is that the two tests between the two steps are not independent. In each simulation case, we also consider the TTFFP method. The multiple test results show that the type I error rates of the TTFFP are difficult to be controlled. Even this method is more powerful, in practice we cannot apply it, as its screening process is too coarse.

Our simulation results show that when the number of markers in analysis is about 1000, the value of L equals to 10 or so is a good choice. At this time, the RTTFP method has higher power. Of course, this conclusion is directly obtained through the simulation analysis. In fact, we can utilize some model selection criteria such as the AIC criterion,[13] BIC criteria,[14] and so on, to get the theoretical optimal value of L. In all scenarios of our simulation, when L exceeds some value the RTTFP method has the highest power among all methods.

Our method also has shortcomings. For example, we do not embed an algorithm for finding the optimal value of L into the test, which is a common problem of the TTFP and the TTFPBSA. These issues will be explored in our future research to find more effective methods of mapping genes of complex traits in the genome-wide analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (no. 11201129), the Natural Science Foundation of Heilongjiang Province of China (A201207), the Scientific Research Foundation of Department of Education of Heilongjiang Province of China (nos 1253G044, and 12531508) and the Scientific Foundation of Heilongjiang University for Distinguished Young Scholars (no. JCL 201003).

1 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95–108.
2 Risch N: Searching for genetic determinants in the new millennium. *Nature* 2000; **405**: 847–856.
3 Zhang S, Zhang K, Li J, Sun FZ, Zhao H: Test of linkage and association for quantitative traits in general pedigree: the quantitative pedigree disequilibrium test. *Genet Epidemiol* 2001; **18**: 370–375.
4 Satagopan JM, Verbel DA, Venkatraman ES *et al*: Two-stage designs for gene-disease association studies. *Biometrics* 2002; **58**: 163–170.
5 Thomas D, Xie R, Gebregziabher M: Two-stage sampling designs for gene association studies. *Genet Epidemiol* 2004; **27**: 401–414.
6 Morley M, Molony CM, Weber T *et al*: Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004; **430**: 743–747.
7 Steen KV, McQueen MB, Herbert A *et al*: Genomic screening and replication using the same data set in family-based association testing. *Nat Genet* 2005; **37**: 683–691.
8 Feng T, Zhang S, Sha Q: Two-stage association tests for genome-wide association studies based on family data with arbitrary family structure. *Eur J Hum Genet* 2007; **15**: 1169–1175.
9 Gu Z: Multiple-stage approach for genome-wide association studies based on data with any family structure. M.Sc. Dissertation, Harbin: Heilongjiang University, 2010, pp 9–10.
10 Lange C, Lyon H, DeMeo D *et al*: A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. *Hum Hered* 2003; **56**: 10–17.
11 Lange C, Demeo DL, Silverman E *et al*: Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am J Hum Genet* 2003; **73**: 801–811.
12 Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; **57**: 289–300.
13 Akaika H: Fitting Autoregressive Models for Prediction. *Ann I Stat Math* 1969; **21**: 243–247.
14 Schwarz G: Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.