RESEARCH NOTE

Association test with the principal component analysis in case–parents studies

LI YU-MEI and XIANG YANG*

Mathematics Department of Huaihua University, Huaihua, Hunan 418008, People's Republic of China

[Yu-Mei L. and Yang X. 2015 Association test with the principal component analysis in case–parents studies. *J. Genet.* **94**, 347–349]

Introduction

In this report, we proposed a statistic with the principal component analysis (PCA) to test association between multiple markers and the disease-susceptibility gene in case–parents data set. The proposed statistic is based on a difference vector calculated by comparing the genotypes of affected offspring with the hypothetical siblings who carry the parental genotypes not present in the affected offspring. The statistic here focuses on all the principal components and is asymptotically distributed as a χ^2 distribution with one degree of freedom. Simulation studies showed that, when the number of markers is not very large, the proposed statistic has higher power than the APRICOT test which is based on the same method of the PCA.

The rapid advancement of genotyping technologies and the availability of enormous quantities of genotype or haplotype data provide an unprecedented opportunity for identifying genes underlying complex traits. When multiple markers are available, haplotype-based methods and genotype-based methods are commonly used for conducting association between complex traits and a series of possibly linked markers. Owing to the information from individual markers as well as the linkage disequilibrium (LD) structure between the markers, the haplotype-based association study is considered to be a potentially superior strategy (Akey *et al.* 2001). However, haplotype-based methods are challenged by a large number of distinct haplotypes, which results in a large number of degrees of freedom, and some haplotype-based methods need estimating haplotype phases only when genotype data at multiple marker loci are collected. On the other hand, genotype-based methods have the advantage of not requiring phase information and in many situations they have higher power than haplotype-based methods (Chapman *et al.* 2003; Xu *et al.* 2006; Rakovski *et al.* 2007; Yu and Wang 2011).

Population-based case–control studies and family-based studies are widely used for association analysis. In comparison to case–control studies, family-based studies are more attractive for their robustness to population stratification. Several genotype-based methods with the use of case–parent data have been proposed for testing association. McIntyre *et al.* (2000) proposed a maxTDT test, in which the usual TDT statistic for each locus is calculated and the maximum is taken as the test statistic. Based on a difference vector calculated by comparing the genotypes of affected offspring with their corresponding 'complements', (i.e., the hypothetical siblings who carry the parental genotypes not present in the affected offspring), three statistics - a paired Hotelling's T^2 test, a max_Z^2 test and the adaptive principal component test (APRICOT) PCT_L are proposed by Fan *et al.* (2005), Shi *et al.* (2007) and Lee (2002), respectively. APRICOT test is developed with the PCA by an adaptive procedure, in which one calculates the principal components of variance–covariance matrix of the difference vector and uses the first few principal components with larger variances determined by a threshold, *c*.

Here, we propose a new statistic, denoted as T_{PC}, with the PCA for testing association. The test statistic T_{PC}, similar to APRICOT test, is also based on the difference vector in case–parents data set, but focuses on all the principal components without examining the optimal choice of the threshold value. Moreover, the statistic T_{PC}, without reference to the number of the markers, is asymptotically distributed as a χ^2 distribution with one degree of freedom. We will assess and compare the performance of the proposed test with the APRICOT test by using simulation studies.

Methods

Consider p diallelic markers, each with alleles 'A' and 'a'. Assume that there are n case–parents trios with the genotypes known for each member of the triads. Let M_i, F_i and C_i

*For correspondence. E-mail: xymail@126.com.

Keywords. association test; linkage disequilibrium; principal component analysis; case–parent trios.

be the number of copies of allele ‘A’ carried by the mother, father, and affected offspring, respectively, at marker locus i. The number of copies of allele ‘A’ carried by the corresponding ‘complement,’ the hypothetical sibling who carries the parental genotypes not present in the affected offspring is $M_i + F_i - C_i$. $2C_i - F_i - M_i$ is then the paired difference in genotypes between the affected offspring and the complement. Let $x_i = 2C_i - F_i - M_i$. The vector of differences for p markers is then p-dimensional random vector, $X = (x_1, x_2, \cdots, x_p)^T$. Let Σ be the covariance matrix of X. Denote the p eigenvalue–eigenvector pairs for Σ by $(\lambda_1, e_1), \cdots, (\lambda_p, e_p)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the kth $(k = 1, \cdots, p)$ principal component is $Y_k = e_k^T X$ with variance of $\mathrm{Var}(Y_k) = \lambda_k$. Note that the p principal components $Y_1, \cdots, Y_p$ are uncorrelated with each other, i.e. $\mathrm{Cov}(Y_k, Y_l) = 0$ $(k, l = 1, \cdots, p;\ k \neq l)$. Let S be the sample covariance matrix for the observed case–parent trios data $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T \quad (i = 1, \cdots, n)$. Assume that the p sample eigenvalue–eigenvector pairs for S are $(\hat{\lambda}_1, \hat{e}_1), \cdots, (\hat{\lambda}_p, \hat{e}_p)$, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Then the kth $(k = 1, \cdots, p)$ sample principal component is $y_k = \hat{e}_k^T X$, and the kth sample principal component for the ith case–parent triad is $y_{ik} = \hat{e}_k^T X_i((i = 1, \cdots, n))$.

The APRICOT test (Lee 2002): the APRICOT test in Lee (2002) is defined as $PCT_L = \max_{k \leq L} T_k$, which is based on the first L components, where $T_k = \left(\sum_{i=1}^{n} y_{ik}\right)^2 / (n\hat{\lambda}_k)$ is asymptotically distributed as a χ^2 distribution with one degree of freedom (df) and $\mathrm{L} = \max\{l : (\lambda_l - \lambda_{l+1})/\lambda_l > c\}$ is determined by the predetermined threshold, c.

The statistic T_{PC}: let $\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_{ik}$. It can be seen that $E(\bar{y}_k) = E(y_k)$, $\mathrm{Var}(\bar{y}_k) = \frac{1}{n}\hat{\lambda}_k$, and $\mathrm{Cov}(\bar{y}_k, \bar{y}_l) = 0$ $(k, l = 1, \cdots, p;\ k \neq l)$. Under the null hypothesis of no association between the set of markers and the disease-susceptibility gene, $E(X) = 0$, and then $E(\bar{y}_k) = 0$ for $k = 1, \cdots, p$. Then, under the null hypothesis of no association, the following statistic T_{PC} is asymptotically a central $\chi^2_{(1)}$ distribution:

$$
\begin{aligned}
T_{PC} &= \left(\sum_{k=1}^{p} \bar{y}_k - E\left(\sum_{k=1}^{p} \bar{y}_k\right)\right)^2 \Bigg/ \mathrm{Var}\left(\sum_{k=1}^{p} \bar{y}_k\right), \\
&= \left(\sum_{k=1}^{p} \bar{y}_k\right)^2 \Bigg/ \left(\sum_{k=1}^{p} \mathrm{Var}(\bar{y}_k) + \sum_{k \neq l} \mathrm{Cov}(\bar{y}_k, \bar{y}_l)\right), \\
&= n\left(\sum_{k=1}^{p} \bar{y}_k\right)^2 \Bigg/ \left(\sum_{k=1}^{p} \hat{\lambda}_k\right).
\end{aligned}
$$

Note that, when $p = 1$, $T_{PC} = PCT_L$. It can be seen that the APRICOT test focuses only on the first few principal components with larger variances, but the statistic T_{PC} focuses on all the principal components without examining the optimal choice of the threshold value.

Results

In order to assess the performance of T_{PC} in finite samples, the simulation study is performed under a wide range of parameter values. The simulations are implemented similar to those described in Lee (2002). Consider a candidate region of 100 kb where p dense marker loci locate according to a uniform distribution. Here, p is chosen from 20 to 100 with increments 20. The marker frequencies are uniformly generated between 0.1 and 0.9, and the Lewontin disequilibrium coefficients between two adjacent markers were uniformly generated between -0.9 and 0.9 (Lee 2002). Assume that a biallelic disease-susceptibility gene with alleles D and d is located within the candidate region according to a uniform distribution. The frequency of disease allele D is set to be 0.05. Let R_1 denote the relative risk with one copy of the disease allele D, and R_2 denote the relative risk with two copies compared with zero copies. Consider one of three settings of population history in Lee (2002): the allele D was introduced into a homogeneous population 2000 generations ago by a mutational process with the mutation occurring in an individual with a specific haplotype—the ‘ancestral haplotype’. The way of generating markers and haplotypes is same as that described in Lee (2002). For the null hypothesis of no association between the set of markers and the disease-susceptibility gene, we let $\mathrm{R}_1 = \mathrm{R}_2 = 1$. For the alternative hypothesis of association, we let $\mathrm{R}_1 = \mathrm{R}_2 = 2$, $\mathrm{R}_1 = 1, \mathrm{R}_2 = 2$, and $\mathrm{R}_1 = 2, \mathrm{R}_2 = 3$, which corresponded to a dominance model, a recessive model and an unrestricted model, respectively. The sample size (the number of case–parent triads) is taken as 200. We performed 10,000 simulations and obtained 10,000 values for T_{PC} and PCT_L. For the given significance level, α (0.05), the actual power/type I error rate is then estimated as the proportion of rejecting the null hypothesis, in 10,000 simulations performed when the alternative/null hypothesis holds.

The results showed type I error rates consistent with the nominal 0.05 level (results not shown). The estimated powers for the statistics T_{PC} and PCT_L are presented in table 1. Here, we let $c = 20\%$ for PCT_L. For a comparison, we also consider the power for $p = 1$. It is observed from the results that the powers both for T_{PC} and PCT_L increase with the number of markers increasing under three genetic models. T_{PC} has higher power than PCT_L when the number of markers

Table 1. The powers for the statistics T_{PC} and PCT_L.

	$\mathrm{R}_1 = \mathrm{R}_2 = 2$		$\mathrm{R}_1 = 1, \mathrm{R}_2 = 2$		$\mathrm{R}_1 = 2, \mathrm{R}_2 = 3$	
Marker	T_{PC}	PCT_L	T_{PC}	PCT_L	T_{PC}	PCT_L
1	0.143	0.141	0.131	0.130	0.137	0.136
20	0.683	0.625	0.651	0.577	0.668	0.582
40	0.749	0.710	0.711	0.676	0.733	0.694
60	0.783	0.788	0.720	0.719	0.775	0.776
80	0.810	0.899	0.782	0.863	0.807	0.880
100	0.845	0.937	0.813	0.905	0.839	0.921

$c = 20\%$ for PCT_L.

sample size is smaller than 60, but PCT_L has higher power than T_{PC} when the number of markers sample size is larger than 60. We also investigated the performance of T_{PC} when there exists population admixture. The results are similar to those under the homogeneous population (data not shown).

Discussion

In this report, we proposed a statistic with the PCA to test association between multiple markers and the disease-susceptibility gene. The proposed statistic is based on a difference vector calculated by comparing the genotypes of affected offspring with the hypothetical siblings who carry the parental genotypes not present in the affected offspring in case–parents data set. Lee (2002) has ever presented the APRICOT test with the PCA. Our method is different from the APRICOT test. The APRICOT test uses the first few principal components with larger variances, whereas our method focuses on all the principal components. Simulation studies showed that the proposed statistic here has higher power than the APRICOT test when the number of markers is not very large. Given that the usual genomewide association (GWA) study data would involve millions of markers, our method may be useful for testing associations between genetic variations within a candidate gene/region. However, it can also be used in a GWA study. In fact, we can adopt the strategy with a gene-centric GWA approach using our proposed statistic by focusing only on those markers located within a gene, i.e., we can deal first with the multiple variants within a gene and then with the multiple genes in the genome (Neale and Sham 2004). Further studies should examine the statistical properties of this approach. Note that our method is not valid when there are missing individual marker genotypes. In a future study, we will focus on the improvement of the principal-component-based method when SNP data are incomplete across loci.

Acknowledgement

The study was supported by National Natural Science Foundation of China (11301206).

References

Akey J., Jin L. and Xiong M. 2001 Haplotypes vs single marker linkage disequilibrium tests: what do we gain. *Eur. J. Hum. Genet.* **9**, 291–300.

Chapman J. M., Cooper J. D., Todd J. A. and Clayton D. G. 2003 Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31.

Fan R., Knapp M., Wjst M., Zhao C. and Xiong M. 2005 High resolution *T*2 association tests of complex diseases based on family data. *Ann. Hum. Genet.* **69**, 187–208.

Lee W. C. 2002 Testing for candidate gene linkage disequilibrium using a dense array of single nucleotide polymorphisms in case–parents studies. *Epidemiology* **13**, 545–551.

McIntyre L. M., Martin E. R., Simonsen K. L. and Kaplan N. L. 2000 Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genet. Epidemiol.* **19**, 18–29.

Neale B. M. and Sham P. C. 2004 The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* **75**, 353–362.

Rakovski C. S., Xu X., Lazarus R., Blacker D. and Laird N. M. 2007 A new multimarker test for family-based association studies. *Genet. Epidemiol.* **31**, 9–17.

Shi M., Umbach D. M. and Weinberg C. R. 2007 Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. *Am. J. Hum. Genet.* **81**, 53–66.

Xu X., Rakovski C., Xu X. P. and Laird N. 2006 An efficient family-based association test using multiple markers. *Genet. Epidemiol.* **30**, 620–626.

Yu Z. and Wang S. 2011 Contrasting linkage disequilibrium as a multilocus family-based association test. *Genet. Epidemiol.* **35**, 487–498.

Received 22 June 2014, in final revised form 2 February 2015; accepted 4 February 2015
Unedited version published online: 6 February 2015
Final version published online: 5 June 2015