Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Power analysis of principal components regression in genetic association studies*

Yan-feng SHEN[1], Jun ZHU[†‡2]

([1]*Department of Mathematics, Zhejiang University, Hangzhou 310027, China*)

([2]*Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*)

[†]E-mail: jzhu@zju.edu.cn

Received Dec. 16, 2008; Revision accepted July 8, 2009; Crosschecked Sept. 8, 2009

Abstract: Association analysis provides an opportunity to find genetic variants underlying complex traits. A principal components regression (PCR)-based approach was shown to outperform some competing approaches. However, a limitation of this method is that the principal components (PCs) selected from single nucleotide polymorphisms (SNPs) may be unrelated to the phenotype. In this article, we investigate the theoretical properties of such a method in more detail. We first derive the exact power function of the test based on PCR, and hence clarify the relationship between the test power and the degrees of freedom (DF). Next, we extend the PCR test to a general weighted PCs test, which provides a unified framework for understanding the properties of some related statistics. We then compare the performance of these tests. We also introduce several data-driven adaptive alternatives to overcome difficulties in the PCR approach. Finally, we illustrate our results using simulations based on real genotype data. Simulation study shows the risk of using the unsupervised rule to determine the number of PCs, and demonstrates that there is no single uniformly powerful method for detecting genetic variants.

Key words: Complex trait, Association study, Principal components, Power
doi:10.1631/jzus.B0830866 **Document code:** A **CLC number:** Q39; O213

INTRODUCTION

Genetic association studies have been used for identifying genetic variants responsible for complex human diseases or traits (Risch and Merikangas, 1996). The recent availability of huge numbers of single nucleotide polymorphisms (SNPs) makes it possible to use this exciting approach in a systematic way. However, the detection of genetic variants of complex traits still faces difficult challenges. One of these challenges is to develop powerful statistical approaches that can make full use of all the information from the SNP data.

In the present study, we focus on assessing whether multiple correlated SNPs in a candidate gene or region influence the trait of interest. Many statistical methods have been developed to analyze SNP data in recent years. Single-locus analysis is the most direct approach: it performs a separate test at each genotyped SNP and takes the maximum of the resulting single-locus statistics to make statistical inference. However, this simple approach may be inefficient because a single locus may not have much information for predicting a causative variant. Schaid *et al*.(2002) showed that several loci within a single gene can produce a large interaction effect on the trait of interest. Therefore, it is reasonable to believe that the combined information across many SNPs may intuitively improve the test power.

Multi-locus association analysis is directly dependent on either haplotypes or genotypes. Currently only genotypes are observed; therefore the first step of haplotype-based methods is to estimate the

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2004CB117306) and the Hi-Tech Research and Development Program (863) of China (No. 2006AA10A102)

haplotypes for each individual by using some statistical procedures such as the expectation-maximization (EM) algorithm (Excoffier and Slatkin, 1995). Then, an association test between the inferred haplotypes and the trait of interest can be considered. Although haplotype-based methods directly exploit patterns of linkage disequilibrium (LD) in a region, these methods have some intrinsic drawbacks. For example, a large number of model parameters are often involved in the test statistic. In fact, the dimension of the statistical model increases exponentially with the increasing number of markers. To reduce the degrees of freedom (DF), strategies such as tagSNP-based haplotype have been proposed (Tzeng *et al*., 2006).

Genotype-based methods make use of genotype data directly to avoid estimating the haplotypes, and therefore do not entail a very large number of DF. Genotype-based tests can outperform haplotype-based approaches (Clayton *et al*., 2004; Chapman *et al*., 2003). One genotype-based method uses Hotelling's T^2 test (Xiong *et al*., 2002), where the number of DF is the same as the number of SNPs. However, multiple markers within a region are in LD and hence are often highly correlated, so some of the DF of Hotelling's T^2 test are wasted, resulting in a loss of power. To reduce the DF and the impact of collinearity, some new statistical methods have been developed in recent years. One such approach is to select a subset of tagSNPs as regressors to test association (Chapman *et al*., 2003). Wang and Elston (2007) provided a weighted Fourier transformation test to reduce the DF and improve the power. More recently, several studies applied principal components regression (PCR) to test for association of the set of SNPs with the phenotype (Gauderman *et al*., 2007; Wang and Abbott, 2008). This approach uses the first few principal components (PCs) directly to assess genetic association.

In general, both haplotype-based and genotype-based approaches can be divided into two basic stages. In the first stage, we want to extract important information from multiple markers and hence reduce the dimensions of the model parameters. Different procedures have been proposed for this purpose such as estimating haplotypes, Fourier transformation and principal components analysis. Several methods, e.g., Hotelling's T^2 test, that use the genotype data directly, may bypass this step. Up to now, the most popular strategies used in this stage can be regarded as unsupervised learning procedures since we use only the information from markers. The second stage is to construct the test statistic based on the important components selected from the first stage. Tests based on regression models are popular for assessing the relationship between the phenotype and these important components (Schaid *et al*., 2002; Kwee *et al*., 2008). In contrast to the first stage, the second stage can be regarded as a supervised learning procedure because information from the phenotype and the markers is used simultaneously. It is worth emphasizing that components identified as important in the first stage may not necessarily be of importance for testing the association between multiple markers and the phenotype of interest. Therefore, it is very important to optimize these two-stage statistical approaches.

In the present paper, we focus our attention on the test procedure based on PCR. Although principal components (PCs) analysis is a popular and efficient statistical method for reducing high dimensionality, one of the crucial problems is to determine the number of PCs to retain for constructing the test statistic. Some authors (Wang and Abbott, 2008; Gauderman *et al*., 2007) have suggested choosing the first few PCs that account for 80%~90% of the total variation in the original SNPs. However, these first few PCs may be unrelated to the outcome. Our goal is to explore the theoretical performance of this approach in more detail. The results clarify the relationship between the test power and the DF, and hence indicate the risk of using the unsupervised rule to select the number of PCs. Next, we introduce a weighted PCs test, which is a general form of many popular test statistics, and compare the test performance of these test statistics. We also provide several alternatives for bypassing the issue of PC number determination. Finally, we demonstrate our results using simulations based on real LD structure.

METHODS

Notation

Assume that there are n unrelated observations with p markers. Let $\boldsymbol{x}_i=(x_{1i}, x_{2i}, \ldots, x_{ni})^{\mathrm{T}}$ be an $n\times1$ vector of all observations at locus i, where x_{ji} is coded as 0, 1, or 2 for the number of copies of the minor

allele. Let $\boldsymbol{X}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ be an $n\times p$ design matrix and $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ be an $n\times 1$ vector of the quantitative trait. Finally, we assume that both the columns of $\boldsymbol{X}$ and $\boldsymbol{y}$ are centered.

Principal components regression

The singular value decomposition (SVD) of matrix $\boldsymbol{X}$ has the form $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n\times p$ and $p\times p$ orthogonal matrices, respectively, and $\boldsymbol{D}$ is a $p\times p$ diagonal matrix with diagonal elements $d_1\geq d_2\geq\ldots\geq d_p\geq 0$. The sample covariance matrix is given as $\boldsymbol{S}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/(n-1)$, and from the SVD of centered matrix $\boldsymbol{X}$ we have

$$\boldsymbol{S}=\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{\mathrm{T}}/(n-1). \tag{1}$$

Let $\boldsymbol{v}_i$ be the ith column of $\boldsymbol{V}$, for $i=1, 2, \ldots, p$, so from Eq.(1) $\boldsymbol{v}_i$ satisfies $\boldsymbol{S}\boldsymbol{v}_i=d_i^2\boldsymbol{v}_i/(n-1)$. Now the ith PC can be defined as $\boldsymbol{z}_i=\boldsymbol{X}\boldsymbol{v}_i$. There are some important properties of PCs. First, the sample variance of $\boldsymbol{z}_i$ equals $\lambda_i=d_i^2/(n-1)$, and hence the PCs are ordered by the decreasing order of explained variances $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p$. Thus, the first few PCs capture more information than the others in contributing to the variation in $\boldsymbol{X}$. Second, the sample covariance of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ is zero, for $i\neq j$. This means that the transformed variables (PCs) are orthogonal and uncorrelated. More details about PCs analysis can be found in books about multivariate statistics (Mardia *et al*., 1979; Jolliffe, 2002).

We assume that phenotype values depend on genotype data through the following linear model:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\beta}$ is a $p\times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an $n\times 1$ vector of normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{I}_n$ with $\boldsymbol{I}_n$ the $n\times n$ identity matrix. Note that the intercept is not included in the model Eq.(2) because the data are centered.

The basic idea of PCR is that we use the first few PCs to replace the original genotypic variables in model Eq.(2), i.e.,

$$\boldsymbol{y}=\sum_{i=1}^{m}\boldsymbol{z}_i\alpha_i+\boldsymbol{e}, \tag{3}$$

for some $m\leq p$. Here α_i is the regression coefficient of $\boldsymbol{z}_i$, and $\boldsymbol{e}$ is a vector of normal distribution. Note that if $m=p$, we would have simply the original model Eq.(2) as a full model. For $m<p$, we have a reduced regression model. Clearly, the mean of $\boldsymbol{z}_i$ is zero, so model Eq.(3) does not have an intercept term.

In genetic association analysis, our interest focuses primarily on testing whether these SNPs have association with the phenotype. It is equivalent to testing the hypothesis H_0: $\boldsymbol{\beta}=\mathbf{0}$; all regression coefficients are zero. If the null hypothesis is true, we can expect that the regression coefficients of model Eq.(3) should all be zero. To construct the test statistic of PCR, we first present some useful properties of PCR in lemma 1:

Lemma 1 Under the assumption of model Eq.(2), let SSR_i be the regression sum of squares of $\boldsymbol{z}_i$ for model Eq.(3) and RSS be the residual sum of squares for the full model Eq.(2). Then we have

(1) $SSR_i/\sigma^2=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y}/\sigma^2$ follows a χ^2 distribution with one DF and non-centrality parameter $\delta_i=d_i^2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\beta}/\sigma^2$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ ($i=1, 2, \ldots, p$), respectively.

(2) RSS/σ^2 follows a center χ^2 distribution with $n-1-rk$ DF, where rk is the rank of $\boldsymbol{X}$.

(3) SSR_1, SSR_2, ..., SSR_p, RSS are mutually independent.

The proof of lemma 1 is straightforward. For simplicity, δ_i will be called the non-centrality parameter corresponding to the ith PC. These non-centrality parameters that measure the association strength between the PCs and the trait of interest will play an important role in the following sections.

Now we propose the test statistic of PCR as followings:

$$F_{\mathrm{PCR},m}=\frac{SSR(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m)/m}{RSS/(n-1-rk)}, \tag{4}$$

where $SSR(\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m)$ denotes the regression sum of squares for model Eq.(3). Since the $\boldsymbol{z}_i$ is orthogonal, we have $SSR(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m)=\sum_{i=1}^{m}SSR_i$. From lemma 1, it can be shown that if H_0 is true, $F_{\mathrm{PCR},m}$ follows a $F_{m,n-1-rk}$ distribution. Note that our test statistic based on PCR is different from the test proposed by Wang and Abbott (2008). Here we use the residual sum of squares corresponding to a full model Eq.(2) to construct the test statistic rather than that corresponding

to a reduced model Eq.(3). Since *RSS* follows the same χ^2 distribution under both the null and alternative hypotheses, it is convenient to derive the power function of $F_{\mathrm{PCR},m}$, as discussed in the next subsection.

Power function of principal components regression test

In practice, it is of great importance to determine the number of PCs in the PCR model. This is equivalent to choosing the DF of model Eq.(3). In this subsection, we will gain insight into the relationship between the test power and the number of PCs, which also clarifies the relationship between the power and the DF.

The power of a statistical test is the probability of rejecting a false null hypothesis. Using the results of lemma 1, it is easy to show that under the alternative hypothesis, $F_{\mathrm{PCR},m}$ follows a non-central F distribution with DF m and $N=n-1-rk$ and non-centrality parameter $\sum_{i=1}^{m}\delta_i$. Therefore, the power function of $F_{\mathrm{PCR},m}$ is as follows:

$$\begin{aligned} &Power(m,N,\alpha,\boldsymbol{\Delta}) = \\ &P(F(m,N,\sum_{i=1}^{m}\delta_i) \geq F^{-1}_{(m,N)}(1-\alpha)), \end{aligned} \tag{5}$$

where α is significance level and $\boldsymbol{\Delta}=(\delta_1, \delta_2, \ldots, \delta_p)$. Next we will examine some important features of the power function Eq.(5).

First, we consider the situation when m and n are fixed. In this case, the DF of $F_{\mathrm{PCR},m}$ are fixed, and then the power is directly dependent on the non-centrality parameter of the F distribution $\sum_{i=1}^{m}\delta_i$. It is well known that for $a_1 \geq a_2 \geq 0$ and $x \geq 0$, we have $P(F(m,N,a_1)\geq x) \geq P(F(m,N,a_2)\geq x)$. To gain greater power of $F_{\mathrm{PCR},m}$, we naturally expect that the first few non-centrality parameters δ_i are larger than others. However, the values of non-centrality parameters δ_i are dependent not only on λ_i but also on $\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\beta}$. Although the PCs are arranged by the magnitude of the corresponding variances $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, the non-centrality parameters δ_i may not follow the same order. Therefore the test statistic $F_{\mathrm{PCR},m}$ using the first m PCs may not be the most powerful. In other words, although the first few PCs can explain most of the variation in the original variables, these PCs may not contribute to the variation in the phenotype. In this situation the test statistic based on these PCs has a poor power. In addition, when m is fixed, clearly the power of $F_{\mathrm{PCR},m}$ is optimal when the m PCs corresponding to the largest m non-centrality parameters δ_i are chosen to construct the test statistic.

Second, the power function is a function of the number of PCs m. Roughly speaking, when the number of PCs varies from m to $m+1$, the power may increase or decrease according to the association strength of the $(m+1)$th PC with the phenotype. Mathematically, if m and $\sum_{i=1}^{m}\delta_i$ are fixed, we can calculate a critical value θ that satisfies

$$\begin{aligned} &P(F(m,N,\sum_{i=1}^{m}\delta_i) \geq F^{-1}_{(m,N)}(1-\alpha)) = \\ &P(F(m+1,N,\sum_{i=1}^{m}\delta_i+\theta) \geq F^{-1}_{(m+1,N)}(1-\alpha)). \end{aligned} \tag{6}$$

This equation implies that if the non-centrality parameter δ_{m+1} is larger than θ, the test $F_{\mathrm{PCR},m+1}$ will have greater power than $F_{\mathrm{PCR},m}$. However, the power of $F_{\mathrm{PCR},m+1}$ will be smaller than $F_{\mathrm{PCR},m}$ when δ_{m+1} is smaller than θ. Fig.1 shows the numerical solutions for this equation under different scenarios of m and $\sum_{i=1}^{m}\delta_i$ with N=100. We can see that a larger $\sum_{i=1}^{m}\delta_i$ results in a relatively larger critical value θ, and for the same value of $\sum_{i=1}^{m}\delta_i$, the critical value is much smaller for a larger m because the difference between $F^{-1}_{(m,N)}(1-\alpha)$ and $F^{-1}_{(m+1,N)}(1-\alpha)$ becomes smaller as m increases. Thus, the power in some

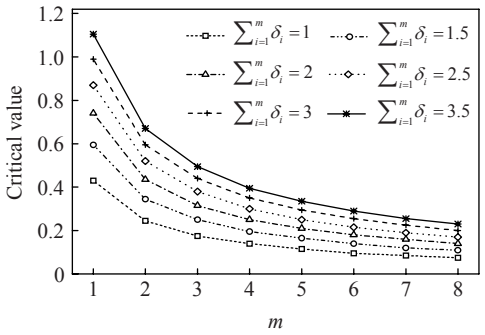

Fig.1 Numerical solutions for Eq.(6) under different scenarios of m and $\sum_{i=1}^{m}\delta_i$

sense may be sensitive to the number of PCs. This phenomenon will be demonstrated using the simulation data in section SIMULATIONS.

The last features of the power function concern the sample size n and the variance of random error σ^2. The power of $F_{\mathrm{PCR},m}$ is also related to these two parameters. With regard to sample size, one question may be how many samples are required to reach a certain level of power. The problem of sample size determination is relatively complicated and is beyond the scope of this article. Here we simply note that if one of the first m PCs does associate with the phenotype, the power of $F_{\mathrm{PCR},m}$ will be arbitrarily close to 1 as the sample size n becomes infinite. With regard to the variance σ^2, we note that when it decreases, the positive non-centrality parameters δ_i will increase, and hence the power increases.

A general weighted principal components test

In this subsection, we will provide a simple modification of the test statistic of PCR Eq.(4), which includes some existing test statistics as special cases. Let $\boldsymbol{w}=(w_1, w_2, \ldots, w_p)$ be a non-negative vector. We consider a weighted PCs test statistic as follows:

$$F_{\mathrm{w}}=\frac{\sum_{i=1}^{p} w_i SSR_i}{RSS/(n-1-rk)}.$$

As a result of lemma 1, under H_0 the distribution of $\sum_{i=1}^{p} w_i SSR_i \big/ \sigma^2$ follows a mixture χ_1^2 distribution, which is called a chi-bar squared distribution (Robertson *et al*., 1988). According to the results of Zhang and Lin (2003), this mixture distribution can be approximated by $s\chi_f^2$ with $s=\sum_{i=1}^{p} w_i^2 \big/ \sum_{i=1}^{p} w_i$ and $f=\left(\sum_{i=1}^{p} w_i\right)^2 \Big/ \sum_{i=1}^{p} w_i^2$. Thus, under H_0, F_{w}/sf can be approximated by an $F_{f,n-1-rk}$ distribution. Note that if all w_i=1, the test F_{w}/sf is equivalent to the traditional F test in a linear regression model. If $w_1=w_2=\ldots=w_m=1$ and $w_{m+1}=w_{m+2}=\ldots=w_p=0$, F_{w}/sf simply becomes the PCR test $F_{\mathrm{PCR},m}$. In addition, we could consider the weights $w_i=d_i^4/(d_i^2+\kappa)^2$, where κ is a non-negative constant. In this case the test statistic F_{w}/sf is the test based on the ridge regression (RR) model. To clarify this point, we note firstly that the ridge coefficients of the full model are $\hat{\boldsymbol{\beta}}_{\mathrm{r}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\kappa\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$. Using SVD we can write the ridge fitting vector as $\hat{\boldsymbol{y}}_{\mathrm{r}}=\sum_{i=1}^{p} d_i^2\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y}\big/(d_i^2+\kappa)$, then the regression sum of squares of RR is $\sum_{i=1}^{p} d_i^4\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{y}\big/(d_i^2+\kappa)^2$. So, the F-ratio test of RR can be regarded as a special case of F_{w} with $w_i=d_i^4/(d_i^2+\kappa)^2$.

All these tests are the linear weighted combinations of the statistic $SSR_i(n-1-rk)/RSS$, which tests whether the ith PC is associated with the phenotype. While the PCR test $F_{\mathrm{PCR},m}$ puts the same weights on the first m PCs, the weights of the RR test are proportional to the variances of PCs. We would expect the more powerful methods to be those in which heavier weights are put on the PCs with the most association strength. Consequently, the PCR test and the RR test often have higher power than the F test when large variance PCs account for more of the variation in the trait than the small variance PCs.

Alternative methods

Clearly, using the 80%~90% rule to choose the number of PCs is often not optimal. To avoid selecting a specific m, in this subsection we introduce several alternative tests based on the PCs. The first strategy is to combine p-values from multiple tests by $T_{\min p}=\min_{1\le i\le p} p_i$, where p_i is the p-value obtained from the test $F_{\mathrm{PCR},i}$. A small value of $T_{\min p}$ suggests a rejection of H_0. Indeed, we simply select a test such that the corresponding p-value reaches the minimum among all p tests $\{F_{\mathrm{PCR},i}, i=1, 2, \ldots, p\}$. Naturally, $T_{\min p}$ is not the correct p-value for such an approach. In this case we can use a standard permutation procedure to find the p-value by permuting the phenotype value $\boldsymbol{y}$ across all individuals. There is another well known method for combining strength across multiple tests. Fisher (1932) proposed the following test statistic $T_{\mathrm{Fisher}}=-2\sum_{i=1}^{p}\log(p_i)$. Because the asymptotical distribution of T_{Fisher} is unknown, one can use a permutation procedure to obtain the p-value.

The second alternative strategy is to combine information across all PCs rather than all PCR tests. Let $\hat{\sigma}=\sqrt{RSS/(n-1-rk)}$ be an estimator of σ and

$T_i = \boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{y} / \hat{\sigma}$ be an estimator of the normalized regression coefficient of α_i in model Eq.(3), i=1, 2, …, p. Following Xu *et al*.(2003), we could construct a test statistic $W(c) = \sum_{i=1}^{p} \max(|T_i|, c) \cdot |T_i|$, where c is a non-negative constant. Next let us see how this test statistic works. Under the assumption of lemma 1, statistic T_i follows a non-central t distribution with DF $n-1-rk$ and non-centrality parameter $d_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{\beta}/\sigma$. When n is relatively large, T_i could be approximated by a normal distribution with mean $d_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{\beta}/\sigma$ and variance 1. Thus, when c=0, the test statistic $W(c)$ equals the sum of T_i^2, so $W(c)$ is simply equivalent to a classical F test in multiple linear regression. When c=2, under the null hypothesis, the distribution of T_i is close to the standard normal distribution, and therefore $W(c)$ is approximately a linear combination of $|T_i|$. However, under the alternative hypothesis, $W(c)$ could give a larger weight to $|T_i|$ if the absolute value of the non-centrality parameter $d_i \boldsymbol{v}_i^{\mathrm{T}} \boldsymbol{\beta}/\sigma$ of T_i is relatively large. Since there is no single c to make the test optimal, Xu *et al*.(2003) suggested a computation-intensive procedure for testing H_0 against H_1, which chooses c from an interval [0,4]. First, let $w(c)$ be the observed value of $W(c)$, then its p-value $p(c)=P(W(c)\geq w(c))$ can be calculated using Monte Carlo methods. Second, we can define test statistic $W = \min_{0\leq c\leq 4} p(c)$. As for test $T_{\min p}$, we use a Monte Carlo method to compute the correct p-value of such a test statistic.

Our simulation studies show that the last few PCs may be less likely to associate with the phenotype. To save the DF and improve test power, we could drop the last few PCs. In practice, we could choose the first q PCs such that they account for 95%~99% of the variation in $\boldsymbol{X}$ and then apply our alternative methods based on these q PCs. We denote these test statistics by $T^{\mathrm{d}}_{\min p}$, $T^{\mathrm{d}}_{\mathrm{Fisher}}$ and W^{d}.

SIMULATIONS

We used simulations to test the arguments that have been raised in the previous section. We based our simulations on the Center d′Etude du Polymorphisme Humain (CEU) genotypes from build 35 of the International HapMap Project. We downloaded the genotype data of gene CHI3L2 from http://www.hapmap.org. There are 49 SNPs within CHI3L2 in the CEU sample that contains 90 individuals. Markers having minor allele frequencies of <0.05 were removed from our simulation study as in other studies (Roeder *et al*., 2005; Tzeng *et al*., 2006; Wang and Elston, 2007). The majority of the SNPs are in strong LD (see, for example, Kwee *et al*.(2008) for the LD plot of SNPs within CHI3L2). We observed that the genotype data of the remaining 30 SNPs contained some missing values. To compute the PCs of the genotype data, we imputed missing genotypes using a simple procedure used by Wang and Abbott (2008). After basic computation, we found that the first three PCs had captured 89% of the variation in the original markers (95% for the first five).

Type I error rates

We then examined whether the PCR test and a general weighted PCs test (the RR test) have appropriate size. To achieve these objectives, we considered simulations under the null hypothesis that no SNPs within the region have association with the outcome. For simplicity, we first defined the DF of these two tests. Here the DF of PCR test can be defined as m and the DF of the RR test can be defined as $df(\kappa) = \sum_{i=1}^{p} d_i^2 / (d_i^2 + \kappa)$. After calculating d_i^2 from SVD of $\boldsymbol{X}$, we can select a specific κ such that the DF varies from 1 to p. We generated the phenotype from a standard normal distribution for 90 individuals and repeated 5000 times. The empirical type I error rates for two tests with DF from one to eight are reported in Fig.2. We can see that both tests had appropriate size regardless of the choice of the DF. This result is

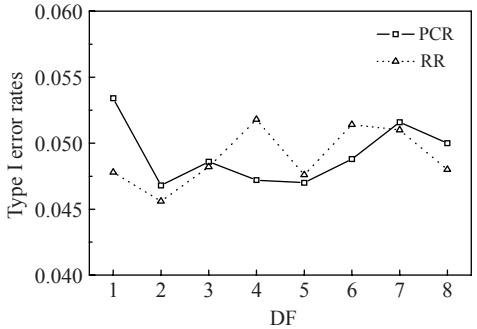

Fig.2 Type I error rates of the PCR and RR tests with degrees of freedom (DF) from one to eight at a significance level of 5%

important especially for the RR test, since we applied a scaled F distribution to approximate the complicated mixture distribution of the RR test statistic in our simulation study. Therefore, the approximated critical values at level α=0.05 for the RR test statistic can be used in the power analysis.

Power comparisons

To assess the power performance of test statistics, we considered the true models where one locus within the region is selected to serve as the disease susceptibility locus (DSL). Let s denote the disease locus, now the phenotypic value for individual i could be generated with the following model: $Y_i=G_{i,s}+\varepsilon_i$. Here $G_{i,s}=\begin{cases}-a, & x_{i,s}=0\\ d, & x_{i,s}=1\\ a, & x_{i,s}=2\end{cases}$, where a and d are called additive effect and dominance effect in genetics, respectively. We generate ε_i under a standard Gaussian distribution. For the analysis to be described, we assumed an additive genetic model (d=0) and chose parameter a in each scenario such that the DSL explained 5% of the trait variation. For each simulation design, we ran 5000 simulations.

Four different choices for the DSL were considered (SNP 4, SNP 10, SNP 18 and SNP 22, in Table 1). We also present the non-centrality parameters of PCs under four different disease models (Fig.3). For each setting, mostly positive non-centrality parameters were concentrated on the first few PCs. Thus, SNPs, which are not correlated with the others, will not contribute much to the first few PCs. These non-centrality parameters do not often follow a descending order. When the DSL was at SNP 10, for instance, the non-centrality parameters corresponding to the first three PCs were much smaller than that corresponding to the fourth one.

Table 1 Details for the four SNPs with CHI3L2 used as the disease susceptibility locus

DSL	Position	MAF	Max R^2	Geno (%)
SNP 4	111485642	0.12	1.00	100
SNP 10	111490367	0.08	1.00	100
SNP 18	111493928	0.25	0.92	97.8
SNP 22	111496180	0.32	1.00	100

DSL: disease susceptibility locus; MAF: minor allele frequency; Max R^2: maximum pairwise R^2 with other SNPs; Geno (%): genotype proportion across all samples

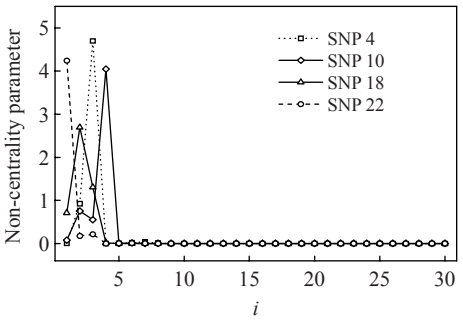

Fig.3 Non-centrality parameters of all PCs under four different choices of the disease susceptibility locus

We next evaluated the effects of the DF on the power of the PCR test and the RR test. The results for the four different disease location scenarios are summarized in Fig.4. The power of the PCR test was more sensitive to the DF when the DSL was at SNP 4 or SNP 10. Although the first three PCs explained 89% variation of $\boldsymbol{X}$, these three PCs contributed little to the variation of the trait in the case of the DSL at SNP 10. So we would be taking a large risk to use the 80%~90% rule to select the number of PCs. It is not surprising that the RR test is more powerful than the PCR test when large variance PCs account for more of the variation in $\boldsymbol{y}$ than do small variance PCs (for example, when the DSL was at SNP 22). However, the power pattern is different when those PCs with large variances do not explain more of the variation in $\boldsymbol{y}$. In the scenario of SNP 18, for example, only the first three PCs contributed to the variation in the phenotype, and of those the second was the most important. We can see that when the DF is 1, the RR test is more powerful than the PCR test because the RR test gives some weight to the second PC while the PCR test uses only the first. If the DF is 2 or 3, the PCR test has a higher power than the RR test since the PCR test gives more weight to the second PC. The power of the PCR test is not greater than that of the RR test when the DF increases from 4 to 8, because the RR test gives relatively small weight to 'noise' PCs that arrange behind the first three PCs. In general, the RR test is more robust than the PCR test against the DF since the RR shrinks smoothly whereas the PCR test makes use of PCs in discrete steps.

We now compare three alternative methods ($T^{\mathrm{d}}_{\min p}$, $T^{\mathrm{d}}_{\mathrm{Fisher}}$ and W^{d}) with the traditional single-SNP

test and the PCR test (T_W) proposed by Wang and Abbott (2008). Here the single-SNP test is F_{max}= max{F_1, F_2, ..., F_p}, where F_i is the F-statistic corresponding to test association of each marker with the trait. We apply the first five PCs to construct three test statistics ($T^d_{\min\ p}$, T^d_{Fisher} and W^d). We investigated the performance of each test method when the DSL was typed and when it was not typed. For each setting and testing method, we used 1000 permutations of the data to evaluate significance, and the power was estimated as the proportion of the p-values less than α=0.05. In this simulation, all the type I error rates tended to be closer to the nominal level (data not shown). Fig.5 shows power results for simulation under four different disease models. Since the four functional loci were in strong LD, we see that the power of each method was quite similar, regardless of whether the DSL was typed. It is not surprising that the performances of $T^d_{\min\ p}$ and T^d_{Fisher} were similar under each disease model since the two tests were based on the same PCR tests. However, W^d and F_{max} also had similar power performance and were relatively robust against the four different choices of the disease locus. Although our simulations were generated as single-locus models, we observe that the single-SNP test did not consistently outperform the other three tests. Thus, there is no single consistently optimal method for detecting genetic variants.

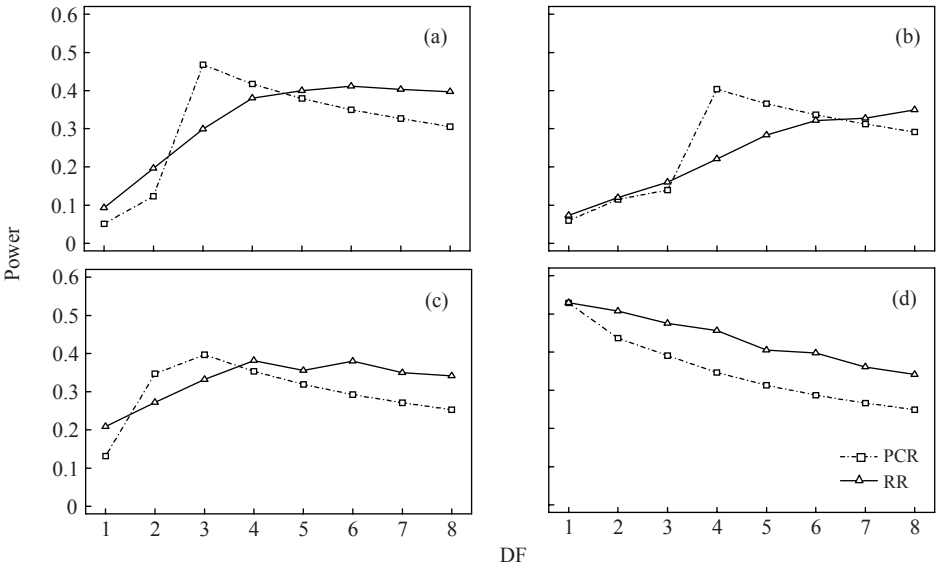

Fig.4 Empirical power of the principal component regression (PCR) test and ridge regression (RR) test, varying the degrees of freedom (DF) from one to eight, under four different choices of the disease susceptibility locus: (a) SNP 4, (b) SNP 10, (c) SNP 18, and (d) SNP 22

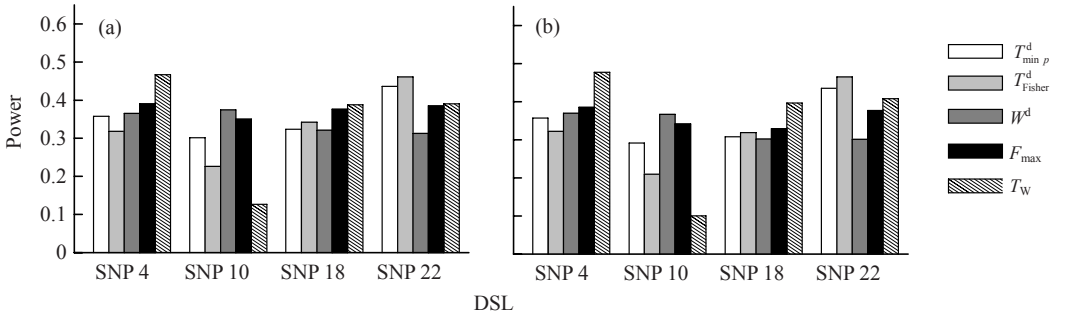

Fig.5 Empirical power of test statistics under four different choices of the disease susceptibility locus (DSL). (a) The disease susceptibility locus is typed; (b) The disease susceptibility locus is not typed

DISCUSSION

Recently, the development of large-scale genotyping techniques has paved the way for using association studies to detect genetic variants of complex traits. However, the high density of biomarkers (i.e., SNPs) not only causes collinearity among regressors but also introduces many DF to statistical models. Consequently, the power for detecting association may be reduced in some cases. Methods are needed, which can use the information from multiple correlated markers but with smaller DF. The PCR approach is just one approach that satisfies this need. When the candidate gene has a relatively high level of LD, the first few PCs can be used to efficiently summarize the major variances of those markers. The smallest variance PCs could represent various groups, such as SNPs that are not correlated with others or rare SNPs (Gauderman *et al*., 2007). Moreover, this approach does not lead to any spurious results because the computation of PCs uses only the information from the original markers.

In this article we have explored the theoretical properties of the test statistics based on the PCs of $\boldsymbol{X}$. A key result of this work is that we have derived the power function of the PCR test. Doing this allows us to clarify the relationship between the power of each test and the DF. We also point out that applying the common rule to select the number of PCs may be risky because the PCs use only the information of $\boldsymbol{X}$. In addition, we have extended the PCR test to a general weighted PCs test which could provide a unified framework to understand the performance of some related test statistics. Lastly, we introduced several alternative strategies to handle the difficulty of choosing the number of PCs. These methods essentially are data-driven adaptive procedures. In summary, our results could be seen as meaningful and important supplements for the PCR test.

We have discussed one of the main issues in principal component analysis that is determining the number of PCs. Another issue is how to specify their biological interpretation because the PCs are a linear transformation of the original variables. However, we are interested mainly in testing the association between the markers and the outcome. So in the context of global association tests (Goeman *et al*., 2004), the biological meaning of PCs is less important. After identifying a significantly global association, other standard approaches, such as single-locus methods, could be applied to further analyze particular markers.

In this article we have derived the exact distribution of the PCR test under the assumption that the trait data follow normality. If this normality assumption is violated, the PCR test will not follow an F distribution. Nevertheless, some nonparametric methods, such as permutation and bootstrap, could be used to approximate the distribution of the test statistic, and then the significance of the test could be assessed. In addition to quantitative traits, there are cases in practice where the data are discrete or categorical. Although logistic principal component regression has been proposed to analyze association (Gauderman *et al.*, 2007), further research work is still needed to examine the theoretical performance of PCR within the generalized linear model framework (McCullagh and Nelder, 1983).

We have compared the relative performance of the tests described in a simulation study based on real LD patterns within 90 CEU individuals. However, these individuals were related. Therefore, we have re-analyzed our simulation based on the genotype data from the 90 CHB+JPT individuals (results not shown). The performance of each test coincided with that described in section SIMULATIONS. Although no method can uniformly outperform the others, our work provides a comprehensive examination of the PCR test for the benefit of other researchers.

References

Chapman, J.M., Cooper, J.D., Todd, J.A., Clayton, D.G., 2003. Detecting disease associations due to linkage equilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.*, **56**(1-3):18-31. [doi:10.1159/000073729]

Clayton, D., Chapman, J., Cooper, J., 2004. Use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol*., **27**(4):415-428. [doi:10.1002/gepi.20032]

Excoffier, L., Slatkin, M., 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**(5):921-927.

Fisher, R.A., 1932. Statistical Methods for Research Workers, 4th Ed. Oliver and Boyd, London, p.99-101.

Gauderman, W.J., Murcray, C., Gilliland, F., Conti, D.V., 2007. Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.*, **31**(5):383-395. [doi:10.1002/gepi.20219]

Goeman, J.J., van de Geer, S.A., Kort, F., van Houwelingen, H.C., 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**(1):93-99. [doi:10.1093/bioinformatics/btg382]

Jolliffe, I.T., 2002. Principal Component Analysis. Springer, New York, p.167-190.

Kwee, L.C., Liu, D.W., Lin, X.H., Ghosh, D., Epstein, M.P., 2008. A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet*., **82**(2): 386-397. [doi:10.1016/j.ajhg.2007.10.010]

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Chapman and Hall, London.

Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science*, **273**(5281): 1516-1517. [doi:10.1126/science.273.5281.1516]

Robertson, T., Wright, F.T., Dykstra, R.L., 1988. Order Restricted Statistical Inference. Wiley, New York, p.59-86.

Roeder, K., Bacanu, S.A., Sonpar, V., Zhang, X.H., Devlin, B., 2005. Analysis of single-locus tests to detect gene/disease associations. *Genet. Epidemiol*., **28**(3):207-219. [doi:10.1002/gepi.20050]

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., Poland, G.A., 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet*., **70**(2):425-434. [doi:10.1086/338688]

Tzeng, J.Y., Wang, C.H., Kao, J.T., Hsiao, C.K., 2006. Regression-based association analysis with clustered haplotypes through use of genotypes. *Am. J. Hum. Genet.*, **78**(2):231-242. [doi:10.1086/500025]

Wang, K., Abbott, D., 2008. A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol*., **32**(2):108-118. [doi:10.1002/gepi.20266]

Wang, T., Elston, R.C., 2007. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.*, **80**(2):353-360. [doi:10.1086/511312]

Xiong, M.M., Zhao, J.Y., Boerwinkle, E., 2002. Generalized T^2 test for genome association studies. *Am. J. Hum. Genet.*, **70**(5):1257-1268. [doi:10.1086/340392]

Xu, X., Tian, L., Wei, L.J., 2003. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics*, **4**(2):223-229. [doi:10.1093/biostatistics/4.2.223]

Zhang, D.W., Lin, X.H., 2003. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**(1): 57-74. [doi:10.1093/biostatistics/4.1.57]