Improving Power in Contrasting Linkage-Disequilibrium Patterns between Cases and Controls

Tao Wang, Xiaofeng Zhu, and Robert C. Elston

Genetic association studies offer an opportunity to find genetic variants underlying complex human diseases. The success of this approach depends on the linkage disequilibrium (LD) between markers and the disease variant(s) in a local region of the genome. Because, in the region with a disease mutation, the LD pattern among markers may differ between cases and controls, in some scenarios, it is useful to compare a measure of this LD, to map disease mutations. For example, using the composite correlation to characterize the LD among markers, Zaykin et al. recently suggested an "LD contrast" test and showed that it has high power under certain haplotype-driven disease models. Furthermore, it is likely that individual variants observed at different positions in a gene act jointly with each other to influence the phenotype, and the LD contrast test is also a useful method to detect such joint action. However, the LD among markers introduced by mutations and their joint action is usually confounded by background LD, which is measured at the population level, especially in a local region with disease mutations. Because the measures of LD that are usually used, such as the composite correlation, represent both effects, they may not be optimal for the purpose of detecting association when high background LD exists. Here, we describe a test that improves the LD contrast test by taking into account the background LD. Because the proposed test is developed in a regression framework, it is very flexible and can be extended to continuous traits and to incorporate covariates. Our simulation results demonstrate the validity and substantially higher power of the proposed method over current methods. Finally, we illustrate our new method by applying it to real data from the International Collaborative Study on Hypertension in Blacks.

Genetic association studies offer an opportunity to find genetic variants underlying complex human diseases.[1] Currently, with the availability of large-scale genotyping techniques, genomewide association studies are underway. Nevertheless, the success of this approach relies on the linkage disequilibrium (LD) pattern between genetic markers, which are typically SNPs, and the functional mutations in a local region of the genome. It has been shown that LD patterns are quite variable in the genome.[2–4]

Various statistical methods have been developed to map functional variants. The most direct approach is single-marker analysis, which involves testing each SNP in turn for association with the disease. However, this simple approach may be inefficient, because any single marker may have limited information to predict the functional variant. Methods that can jointly make use of multiple marker information are therefore very useful. Multiple-marker association analysis may depend directly on either haplotypes or genotypes. Lack of parsimony is a major limitation of the multiple-marker approach, in which a large number of degrees of freedom is often involved in the test statistic. It is likely that there is no single uniformly optimal approach to mapping complex-disease genes.

Another approach is to contrast LD patterns between cases and controls because, in a local region that harbors the disease variant, the extent of LD may be different between cases and controls. For example, following the work of Nielson et al.,[5] Zaykin et al.[6] recently suggested a new "LD contrast test" to compare the pairwise matrices of disequilibrium measures between cases and controls. The use of composite coefficients to characterize the LD pattern in a local region allows their method to be robust to Hardy-Weinberg disequilibrium (HWD), which is expected to occur in the region with the disease variant. The LD contrast test was also suggested to test gene-gene interaction.[7] The rationale behind this approach is that the joint effect of two variants would generate different LD patterns in cases and controls.

However, the LD between two SNP markers in a trait group, whether cases or controls, is the consequence of both selection on the basis of the disease variant and "background LD" due to various other factors. The use of the usual LD coefficients, which measure the whole correlation between two SNPs, may not be optimal for the purpose of detecting association, because of noise coming from the background LD. The most powerful measure for contrasting LD patterns must be able to discount appropriately the background LD. Therefore, it is desirable to find a new measure for capturing the local LD difference between cases and controls. Here, we propose a new test to overcome this problem. The proposed test is developed under a regression model and is flexible enough to incorporate covariates and continuous traits. Because the test depends directly on genotypes, it is also robust to

From the Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Received January 22, 2007; accepted for publication February 26, 2007; electronically published March 28, 2007.

Address for correspondence and reprints: Dr. Robert C. Elston, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106. E-mail: rce@darwin.case.edu

Am. J. Hum. Genet. 2007;80:911–920. 0002-9297/2007/8005-0010$15.00
DOI: 10.1086/516794

HWD. Our simulation results demonstrate the validity and substantially improved power of this new test over the LD contrast test.

Methods

To illustrate the idea, we first consider a comparison of the LD for two SNPs between cases and controls. There is a variety of measures available to characterize LD in cases or controls. Following the notation of Zaykin et al.,[6] the direct measure of the LD coefficient is given by

$$D_{AB} = P_{AB} - P_A P_B \ ,$$

where P_{AB} is the frequency of the haplotype with alleles A and B and where P_A and P_B are the frequencies of alleles A and B, respectively. It can be shown that D_{AB} is related to the Pearson correlation coefficient by

$$r = \frac{D_{AB}}{\sqrt{P_{A1}P_{A2}P_{B1}P_{B2}}} \ ,$$

where P_{A1}, P_{A2}, P_{B1}, and P_{B2} are the frequencies of the alleles 1 and 2 for markers A and B. A standardized measure, which is robust to allele frequency, is the coefficient D'_{AB},[8] which is given by

$$D'_{AB} = \frac{D_{AB}}{\max(D_{AB})} \ ,$$

where

$$\max(D_{AB}) = \begin{cases} \min[P_A P_B, (1-P_A)(1-P_B)] & \text{if } D_{AB} < 0 \\ \min[P_A(1-P_B), (1-P_A)P_B] & \text{if } D_{AB} > 0 \end{cases} \ .$$

This procedure restricts the range of D' between 0 and 1. Complications in estimating the LD coefficients arise when only genotype data are observed. The estimates of both D'_{AB} and D_{AB} rely on the estimate of the haplotype frequency, $\hat{P}_{AB}$, which requires an assumption with regard to Hardy-Weinberg equilibrium. Because, in the local region with disease variants, deviation from Hardy-Weinberg proportions is expected in both cases and controls, Zaykin et al.[6] suggested using composite LD measures, which are robust to HWD. Let $P_{A/B}$ be the joint frequency of alleles A and B in two different gametes, so that the composite LD coefficient is $\Delta_{AB} = P_{AB} + P_{A/B} - 2P_A P_B$.[9] We can see that composite LD coefficients do not distinguish between the two possible phases of the double heterozygotes but rather consider the deviation from random association. The composite correlation is given by

$$r_{AB} = \frac{\Delta_{AB}}{\sqrt{[P_A(1-P_A) + D_A][P_B(1-P_B) + D_B]}} \ ,$$

where D_A and D_B are the HWD coefficients at the two loci—for example, at marker A, $D_A = p_{11} - p_1^2$, in which p_{11} and p_1 are the frequencies of genotype 11 and allele 1. The results of Zaykin et al.[6] showed that tests based on composite correlations and composite LD coefficients have similar power.

Define the observed genotype value at the diallelic locus j for subject i as follows:

$$x_{ji} = \begin{Bmatrix} -1 & \text{if the genotype of individual } i \text{ is } 11 \\ 0 & \text{if the genotype of individual } i \text{ is } 12 \\ 1 & \text{if the genotype of individual } i \text{ is } 22 \end{Bmatrix} \ .$$

The composite correlation can then be estimated by

$$\hat{r}_{AB} = \frac{\hat{\sigma}_{x_A,x_B}}{\sqrt{\hat{\sigma}^2_{x_A}\hat{\sigma}^2_{x_B}}} \ ,$$

where, by denoting the sample means $\bar{x}_A$ and $\bar{x}_B$, $\hat{\sigma}_{x_A,x_B} = \sum(x_{Ai} - \bar{x}_A)(x_{Bi} - \bar{x}_B)/(n-1)$ is the estimated covariance between the genotype values for loci A and B and $\hat{\sigma}^2_{x_A} = \sum(x_{Ai} - \bar{x}_A)^2/(n-1)$ and $\hat{\sigma}^2_{x_B} = \sum(x_{Bi} - \bar{x}_B)^2/(n-1)$ are the estimated variances of the genotype values for loci A and B, respectively.[10] Now, consider the sample mean–corrected, standardized genotype values z_{Ai} and z_{Bi}. Then, we can estimate the composite correlation by

$$\hat{r}_{AB} = \frac{\sum z_{Ai}z_{Bi}}{n-1} \ .$$

The statistic to compare the LD between cases and controls can be given by

$$T_C = \hat{r}_Y - \hat{r}_N \ ,$$

where $\hat{r}_Y$ and $\hat{r}_N$ are the estimated composite correlations between marker A and B in the case and control groups, respectively. Under the null hypothesis that no disease variant exists, T_C is expected to be close to 0, so an unusually large or small T_C statistic indicates the possibility of a disease variant.

Now, the statistic T_C may be considered as equivalent to the regression coefficient in a regression model that has as dependent variable the cross-product of the standardized genotype values of two SNPs—that is,

$$E(z_{Ai}z_{Bi}) = \alpha + \beta t_i \ , \qquad (1)$$

where the predictor variable t_i is an indicator variable for cases and controls and α and β are unknown parameters to be estimated. The parameter β describes the relationship between the predictor variable—in our example here, the case-control classification—and the correlation between two markers (i.e., LD), so a large or small observed value of the standardized estimate of β suggests association between the predictor (i.e., the trait) and two markers. The efficiency of the regression model (1) can be seen by considering the proportion of the dependent variable's variance that is explained by the regression—that is, R^2. It is useful to rewrite the dependent variable of model (1) as $z_{Ai}z_{Bi} = [(z_{Ai} + z_{Bi})^2 - (z_{Ai} - z_{Bi})^2]/4$, which shows that the dependent variable of regression model (1) is the special linear combination of the squared sum and squared difference of genotype values at two loci with equal weights. As discussed in the different context of Haseman-Elston regression to detect linkage,[11–13] the test based on regression model (1) may not be efficient, because the background correlation is not taken into account. The efficiency of regression model (1) depends on the variance of the dependent variable explained by the trait status t_i. Let us consider a class

of dependent variables defined as all linear combinations of the squared sum and squared difference $-w(z_{Ai} - z_{Bi})^2 + (1 - w)(z_{Ai} + z_{Bi})^2$, where w is a weight, between 0 and 1. The most efficient such dependent variable should have the least overall variance. Let the variance of $(z_{Ai} + z_{Bi})^2$ be σ_1^2, the variance of $-(z_{Ai} - z_{Bi})^2$ be σ_2^2, and the covariance between $(z_{Ai} + z_{Bi})^2$ and $-(z_{Ai} - z_{Bi})^2$ be σ_{12}. The optimal weight can be found by solving

$$\frac{\partial \operatorname{Var}[-w(z_{Ai} - z_{Bi})^2 + (w - 1)(z_{Ai} + z_{Bi})^2]}{\partial w} = 0 ,$$

from which we find

$$w = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} .$$

It can be seen that it is optimal for the squared sum and squared difference of genotype values at two loci to have equal weights when $\sigma_1^2 = \sigma_2^2$. However, this is not expected to be the case, because of background correlation among the multiple markers in a local region. So, the composite correlation–based measure may lead to a severe loss in power.

The genotype value follows a multinomial distribution that is in the exponential family and therefore has the form $e^{\psi(x_{ji},\theta_{ji},\phi)}$, where j ($j = A$ or B) denotes the marker and $\psi(x_{ji},\theta_{ji},\phi) = [x_{ji} - b(\theta_{ji})]/d(\phi) + c(x_{ji},\phi)$. Let random effects be denoted by bold letters. To improve the power of regression model (1), we consider modeling the genotype data as follows:

$$E(x_{ji}|\mathbf{u}_{ji},\mathbf{d}_{ji}) = h^{-1}(\mu_j + \mathbf{u}_{ji} + \mathbf{d}_{ji}) ,$$

where h is a link function, x_{ji} represents the genotype value of marker j for the ith ($i = 1,\ldots,n$) subject, μ_j is the fixed overall intercept of the genotype value, $\mathbf{u}_{ji}$ is the marker intercept specific to individual i, and $\mathbf{d}_{ji}$ is the random effect of the trait on the specific genotype value. In this model, the genotype value of a marker for a subject is determined by overall mean μ_j, which may be looked at as the marginal allele frequency, the subject-specific effects $\mathbf{u}_{ji}$, which lead to the LD observed in a general population, and the effect $\mathbf{d}_{ji}$ of the trait selection, which introduces the additional LD of interest. Under this model, the background LD is modeled by $\mathbf{u}_{ji}$, and $\mathbf{d}_i = (\mathbf{d}_{Ai},\mathbf{d}_{Bi})$ is a random vector with mean 0 and covariance matrix modeled as

$$R = \begin{pmatrix} 1 & \varphi(t_i)\delta \\ \varphi(t_i)\delta & 1 \end{pmatrix} \sigma^2 ,$$

where $\varphi(t_i)$ is a function of the trait values. Let $\bar{t}$ be the sample mean of the trait values, and define $\varphi(t_i) = t_i - \bar{t}$. Under this model, whether the correlation between markers (i.e., LD) is related to the trait value can be examined by testing H_0:$\delta = 0$. We consider a canonical link function. The score statistic, which is the first derivative with respect to δ evaluated at the null hypothesis, is (see appendix A)

$$U = \sum_i [x_{Ai} - E(x_{Ai})][(x_{Bi} - E(x_{Bi})]\varphi(t_i) ,$$

where x_{Ai} and x_{Bi} are the genotype values of two markers for individual i. Let $\mu_{ji} = \mu_j + \mathbf{u}_{ji}$. The mean $E(x_{ji}) = b'(\mu_{ji})$, in which $j = A$ or B, depends on μ_{ji}, which is unknown. So, we approximate the score statistic by using an estimate of the mean of x_{ji},

$$U = \sum_i [x_{Ai} - \hat{E}(x_{Ai})][(x_{Bi} - \hat{E}(x_{Bi})]\varphi(t_i) . \qquad (2)$$

We note here that, when the genotype values are treated as constants, $E(\varphi_i) = 0$, and, therefore, $E(U) = 0$, so the validity of this statistic is not affected by the estimation of $E(x_{ji})$, regardless of the value of $\hat{E}(x_{ji})$. However, the value of $\hat{E}(x_{ji})$ influences the power. There are several estimates available. One possibility is the sample mean of the genotype values over all subjects for each marker obtained by ignoring the background LD. For a case-control study with standardized genotype values, the statistic (2) is then equivalent to the composite correlation–based LD contrast test statistic. When the background LD is strong, the variation among individuals may be greater than the variation among markers within an individual, and, therefore, this test is not optimal in terms of power. To take into account the effect $\mathbf{u}_i$, we use a linear mixed model to estimate $\hat{E}(x_{ji})$. Letting $\mathbf{I}$ be a vector of indicators for markers, which can also include appropriate covariates that we wish to adjust for, and β^{T} be the corresponding row vector of regression coefficients, the model $x_{ji} = \beta^{\mathrm{T}}\mathbf{I} + \mathbf{u}_i + \varepsilon_{ji}$ can be conveniently fitted using the *lme* function in the R package, which gives the best linear unbiased predictor (BLUP) of μ_{ji}. Because the BLUP takes both types of information into account—the information across subjects and the information across markers—we expect it to improve the power of statistic (2). If there is concern over the sensitivity of the asymptotic distribution of this statistic, a simple permutation procedure that randomly shuffles the disease status can be adopted to determine the P value of the above statistic.

The new score statistic (2) is closely related to the composite correlation–based LD contrast statistic. It corresponds to testing $\beta_1 = 0$ in the regression model $E[(x_{Ai} - E(x_{Ai})][(x_{Bi} - E(x_{Bi})] = \alpha_1 + \beta_1 t_i$, and so the parameter δ can be expressed as the regression parameter β_1. The statistic T_{C} is equivalent to testing the regression parameter $\beta = 0$ in regression model (1). The dependent variables of both these regression models describe the correlation between two markers (the LD). Hence, both statistics detect association by testing whether the trait is related to the correlation between markers. However, with the aim of improving the power, the proposed statistic uses a different measure to describe this correlation, rather than using the conventional composite correlation. In the new statistic, the genotype values are centered by the individual specific means $E(x_{Ai})$ and $E(x_{Bi})$, which absorb background LD. So, the new test is in fact a test to compare "background-corrected LD" between cases and controls.

The comparison of the LD measure between cases and controls for only two SNPs can be directly extended to all pairwise LD statistics for a set of SNPs in a local region. Zaykin et al.[6] showed, in their simulation, that the most powerful statistic is based on the overall difference of composite correlations between cases and controls, which is given by

$$T = \operatorname{trace}[(\Lambda_{\mathrm{Y}} - \Lambda_{\mathrm{N}})^T(\Lambda_{\mathrm{Y}} - \Lambda_{\mathrm{N}})] , \qquad (3)$$

where Λ_{Y} and Λ_{N} are matrices of the composite correlations for cases and controls, respectively. Here, we define Λ_{Y} and Λ_{N} such that each element is the corresponding sum of pairwise mean-corrected cross-products, by use of BLUPs of the means.

Results

Proof of Concept

As an initial proof of concept, we first provide evidence to show that discounting the background LD leads to increased efficiency of the test, in contrasting the LD patterns between cases and controls in the simplest case of only two SNP markers. We consider a situation in which two mutations independently occurred at a third (untyped trait) locus on haplotypes 00 and 11. In this case, the LD contrast test should have superior power over a single-marker analysis, because of the weak marginal effect of each marker, and also over a haplotype-based analysis, because of fewer degrees of freedom.[5] First, four haplotype frequencies were determined by the allele frequencies and D', and a pair of haplotypes randomly sampled from the corresponding multinomial distribution for each subject. The disease status is defined by a model similar to that used by Zaykin et al.[6] For a dichotomous trait, the model assumes that the trait is due to an underlying continuous liability (y) to which the trait-locus effects (g) and random environmental effects (e) contribute additively and independently: $y = g_1 + g_2 + e$, where g_1 and g_2 are the trait-locus effects of the two haplotypes on an individual's y value. The trait-locus effects are set to be 2.5 and 0.48 for haplotypes 00 (or 11) and 01 (or 10), respectively. Affection status is defined by a threshold Z, such that all individuals with $y > Z$ are classified as cases. The random effect e is sampled from $N(0,7.5^2)$. The prevalence is set to be 0.02. We consider sampling 100 cases and 100 controls. For each model, we simulate 1,000 data sets, and the permutation test is based on 1,000 replicates of each data set.

We first consider the influence of the background LD. We vary D' from 0 to 0.8. To avoid any influence of the allele frequency, we choose the allele frequencies of both SNPs to be 0.5. Figure 1 shows the empirical power and type I error rate of three different tests for various values of background LD, including a test statistic in which $\hat{E}(x_{ji})$ is estimated by the average genotype values of the two markers, $\hat{E}(x_{ji}) = (x_{Ai} + x_{Bi})/2$ ("D" in fig. 1). This statistic is equivalent to using the squared Euclidean distance between the genotype values of the markers as the dependent variable in regression model (1). Intuitively, when the background LD between two markers is high and the two markers have similar allele frequencies, this test is favored because the variability between markers is less than that among subjects. Otherwise, the correlation-based test is favored. The test proposed in this article that uses a mixed-correlation model to estimate the correlation of two markers is denoted "Mc" in the figures and should have high power when there is any background LD. We can see that all tests maintain good control of type I error rate at the 5% significance level (fig. 1, right). Figure 1 (left) shows that the power of the correlation-based test decreases with an increase in the background LD. As expected, we observe that the power profiles of the correlation- and distance-based tests cross each other, and the test that we propose is uniformly more powerful. Here, we show the results of the distance-based statistic only to illustrate that the background LD has a different impact on the various tests. When the markers have quite different allele frequencies, it is clear that the use of the average of different markers to predict μ_{ji} is not suitable. We also evaluated the power under different prevalences, finding similar results. In figure 2, we assume that the quantitative phenotype values y—for example, for blood pressure—of cases and controls can be observed. Because the proposed statistic can make use of this quantitative information, it can further improve the power of the LD contrast test.

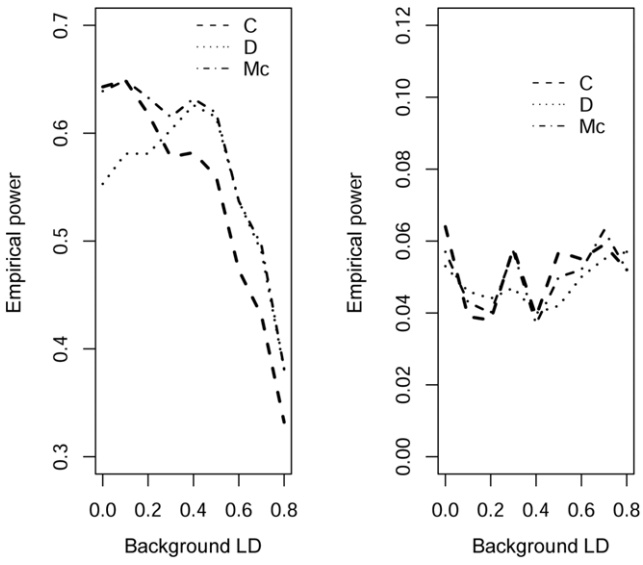

Figure 1. Comparison of the empirical power for 100 controls and 100 cases between the composite correlation–based LD contrast test (C), the proposed test (Mc), and the analogous statistic (D) in which the mean is estimated by the average genotype values of the two markers, at different background LD. The marker-allele frequencies of the two SNPs are both 0.5. The background LD is measured as D'.

Power

We further compare the power of tests for a set of markers with m SNPs in a local region or candidate gene. We also consider a single-marker analysis in the simulations, in which we fit the regression one marker at a time and the minimal P value (T_P) is evaluated on the basis of a permutation procedure by shuffling the trait values to maintain the dependence among the markers. For multiple-marker analysis, we consider Hotelling's T_H, which jointly tests the marginal effects of multiple markers while accounting for the correlations among them.

The haplotypes for $m = 4$ and 10 correlated SNP markers are simulated on the basis of a multivariate normal distribution with pairwise correlations ρ. Each allele of a haplotype is generated by dichotomizing the marginal

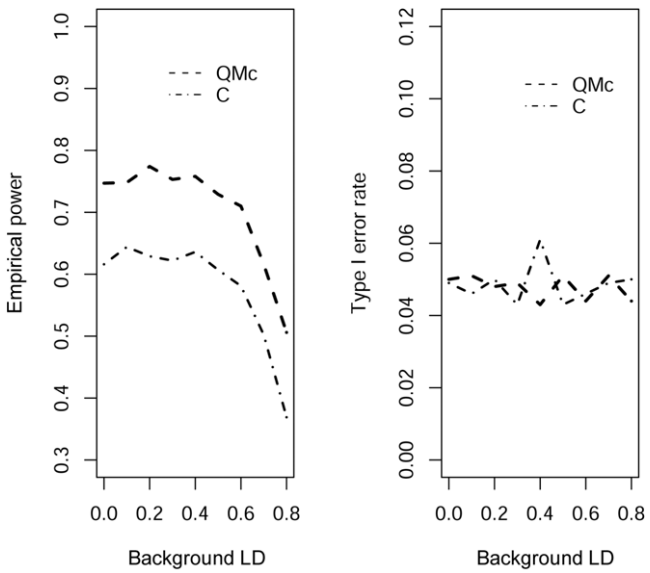

Figure 2. Comparison of the empirical power and type I error rate for 100 controls and 100 cases between the composite correlation–based LD contrast test (C) and the proposed test (QMc) at different marker-allele frequencies when continuous phenotypes are observed. The background LD is D'.

normal distribution, and the cutoff is determined by an allele frequency that either is set to be between 0.1 and 0.5 or is randomly sampled from a uniform distribution between 0.1 and 0.5. The disease status is simulated as before, in which larger effects tend to be defined by haplotypes that are most different. In the power comparison, we consider three scenarios—that is, ρ_{ij} is set to be a constant, is $0.8^{\log(1+|i-j|)}$, or is randomly sampled from a uniform distribution between 0.6 and 0.9. The scenario in which ρ_{ij} equals a constant corresponds to an average pairwise correlation of 0.25 between SNPs, with each marker providing similar information about the disease locus. The scenario $\rho_{ij} = 0.8^{\log(1+|i-j|)}$ is similar to an LD pattern in which LD is primarily a function of marker distance. However, because of population phenomena such as genetic drift, mutation, nonrandom mating, and so forth, the actual LD pattern is more complicated; to simulate this last scenario, we sample ρ values from a uniform distribution between 0.6 and 0.9.

Here, we only show the results for $m = 4$ markers because the results are quite similar for $m = 10$ markers. The type I error rates for the four tests are all close to the nominal 0.05 level (data not shown). As seen in figure 3, for the case of multiple markers, we find results similar to those for the case of two markers. The proposed test usually performs better than the correlation-based test when background correlation exists among the SNPs, and the gain in power increases with increasing background LD. Figures 4 and 5 further show that the proposed test has uniformly the best performance compared with the other three tests in our simulations with two different LD patterns: LD as a function of marker distance and LD that does not necessarily follow the distances between pairs of markers. Because the LD contrast test detects different association information from that detected by T_H and T_P, as discussed by Zaykin et al.,[6] and the marginal association information for each marker is small in our simulation, it is not surprising that we found much higher power for our statistic (figs. 4 and 5).

We further performed simulations based on the real LD pattern at the angiotension I–converting enzyme (ACE) locus (MIM 106180). The genotype data of 13 SNPs in the ACE locus for 310 independent subjects were selected from the Nigeria data set of the International Collaborative Study on Hypertension in Blacks.[14] The LD pattern for the 13 SNPs at the ACE locus is given in figure 6. The squared correlation coefficient (r^2) among SNPs is between 0 and 0.93, although most correlations among the SNPs

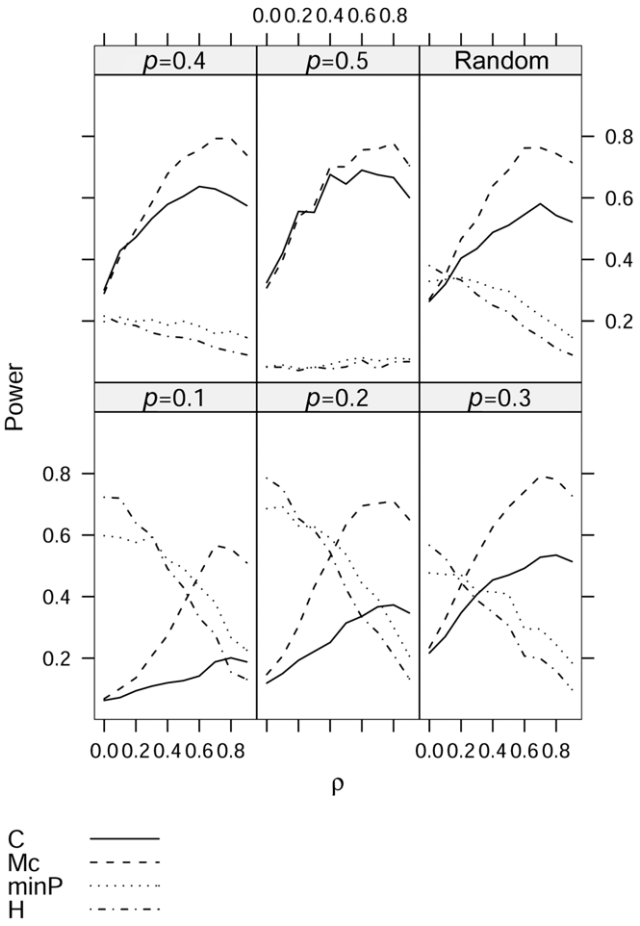

Figure 3. Comparison of the empirical power for 100 controls and 100 cases with four markers at different background LD (ρ) and allele frequencies (p) between the composite correlation–based LD contrast test (C), the proposed test (Mc), the minimum P value in single-marker analysis (minP), and Hotelling's test of the marginal effects of multiple markers (H).

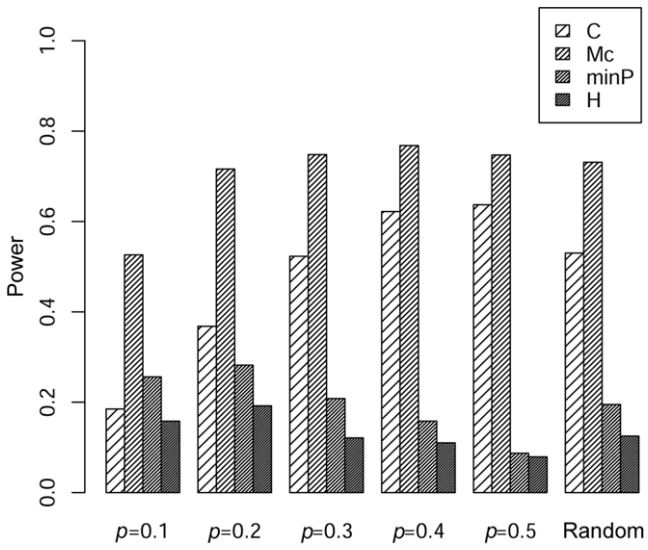

Figure 4. Comparison of the empirical power for 100 controls and 100 cases with four markers between the composite correlation–based LD contrast test (C), the proposed test (Mc), the minimum *P* value in single-marker analysis (minP), and Hotelling's test of the marginal effects of multiple markers (H). The LD pattern is simulated to be a function of marker distance. The marker-allele frequency (*p*) is set to be 0.1–0.5 or is randomly sampled from a uniform distribution between 0.1 and 0.5.

are small. An underlying continuous liability value is similarly simulated as before, and a balanced case-control data set is determined by dichotomizing this continuous variable. The empirical power for the real LD pattern for the tests can been seen in figure 7, which shows that the proposed test is most powerful in the case of this real LD pattern.

Application to Data on ACE Levels

The rennin-angiotensin system (RAS) is known to have a key role in blood-pressure regulation. ACE is a key component of the RAS because it catalyzes the conversion of angiotensin I to angiotensin II, a potent vasoconstrictor that leads to the constriction of blood vessels and the retention of salt and water. The ACE gene polymorphism has been extensively studied, although a causative effect of the ACE gene on hypertension is still not established.[14–16] Bouzekri et al.[17] described the association between 13 variants in the ACE gene at an average distance of 2 kb apart and the ACE plasma level in three population samples, from Nigeria, Jamaica, and an African American community in the United States. Their results suggest that there is more than one functional variant affecting the ACE plasma level. However, whether these variants affect the ACE plasma level interactively is unclear. To illustrate the application of our method, we tested whether the LD patterns of these 13 SNPs are different between subjects with higher and lower ACE plasma levels. We compare the new statistic (T_{Mc}) with two other LD contrast test statistics, in which the composite LD correlation (T_C) and the standardized composite LD coefficient ($T_{\Delta'}$) are used to describe the LD pattern.

The data consist of 2,776 family members from Nigeria and Jamaica and an African American community. Our analysis is restricted to independent subjects with nonmissing genetic data from these families, by sampling one subject from each family. As a result, our analysis is based on 310 subjects from Nigeria, 116 subjects from Jamaica, and 252 subjects from the African American community. We further created a balanced case-control data set by equally dichotomizing the ACE level for each population. The *P* values for all tests were obtained using a permutation procedure with 500,000 replicates.

Table 1 presents the *P* values of the test statistics T_{Mc}, T_C, and $T_{\Delta'}$ across the three population samples. In general, we consistently observed T_{Mc} and T_C to have more power than $T_{\Delta'}$ in the three samples, whereas T_{Mc} also tends to show slightly stronger evidence of association than does T_C, which is consistent with what we observed in the simulation studies.

Discussion

In the present study, we extend the LD contrast test under the framework of a generalized linear model. There are various analytic methods developed for a genetic association study. The LD contrast test relies on the difference

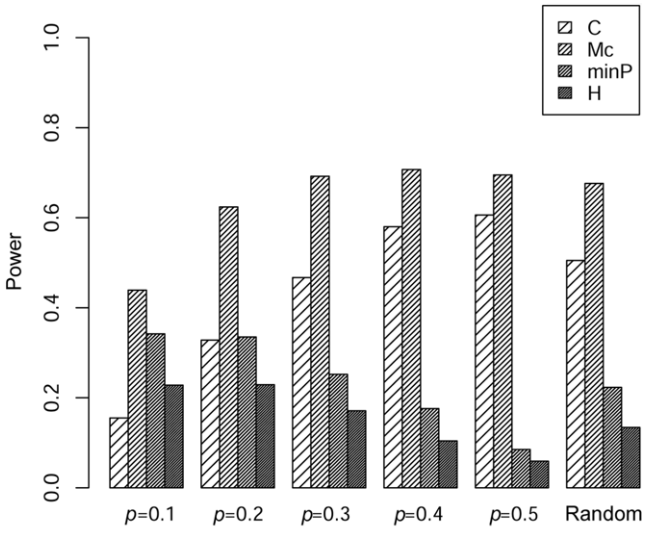

Figure 5. Comparison of the empirical power for 100 controls and 100 cases with four markers between the composite correlation–based LD contrast test (C), the proposed test (Mc), the minimum *P* value in single-marker analysis (minP), and Hotelling's test of the marginal effects of multiple markers (H). The ρ values are sampled from a uniform distribution between 0.6 and 0.9. The marker-allele frequency is set to be 0.1–0.5 or is randomly sampled from a uniform distribution between 0.1 and 0.5.

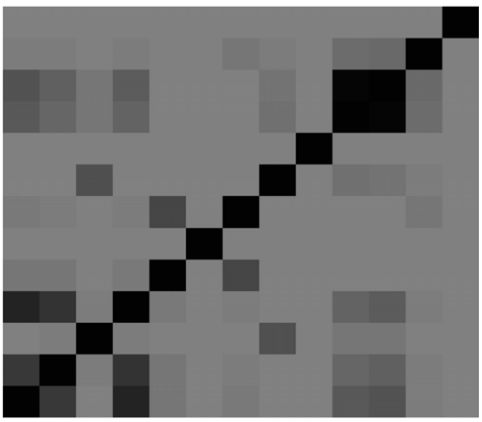

Figure 6. The LD-pattern plot for 13 SNPs of the ACE locus. The color scale from white (lower values) to black (higher values) corresponds to an increase in the absolute values of correlation.

of pairwise LD among markers, rather than on the change of the marginal allele frequencies. So, the LD contrast test and the single-marker or multiple-marker genotype score tests, such as Hotelling's test, tend to detect different information available in the data. The genotype score–based tests are likely to fail in models in which there are no substantial marginal SNP effects. An example is seen when susceptibility haplotypes tend to be "yin-yang" haplotypes.[18] There has been a report of an exceptional abundance of this particular haplotype pattern, in which two high-frequency haplotypes have different alleles at every SNP site (thus the name "yin-yang haplotypes"). The LD contrast is expected to have high power in this case. Haplotypes provide more information than do the allele frequencies and the pairwise LD. However, the haplotype-based tests often involve a large number of degrees of freedom. Because the LD extending more than two loci decays rapidly, it is reasonable to consider the allele frequencies and pairwise LD, rather than whole haplotypes, when the number of haplotypes is too large.

Currently, the LD contrast test depends on conventional LD measures, such as the composite correlation, to test whether there is a significant difference in these measures between cases and controls. One problem with the current method is that the LD introduced by trait selection is confounded by the background LD. Often, background LD is far greater than the trait-related LD in a local region of the genome. We show by simulations that the method proposed in this article can improve on the previous method by taking into account the background LD. However, the new test does not replace the LD contrast tests based on conventional LD measures. In practice, an investigator may be specifically interested in whether there is a significant difference in a conventional LD measure between cases and controls. In this case, the test with the corresponding LD measure, accompanying the graphical LD plots, is useful.

Our simulation studies suggest that the proposed test usually performs better than the correlation-based test when background correlation exists among SNPs. This can be further observed in the application of the method to ACE data in three population samples. The proposed method tends to detect the joint effects of SNPs. Therefore, it is understandable that we did not observe small *P* values in the original report, which focused on detecting the marginal effects.[17] Our analysis here was restricted to independent subjects with nonmissing genetic data sampled from families. The sample sizes were thus much smaller than those used by Bouzekri et al.[17] Both statistics T_{Mc} and T_{C} consistently suggested an association between the ACE polymorphism and plasma ACE level. However, we found that the *P* values were above the 5% significance level for all the three test statistics in the African American sample ($P = .073$, $P = .071$, and $P = .074$ for T_{Mc}, T_{C}, and $T_{\Delta'}$, respectively). The less significant results obtained from the African American sample might be because of its larger proportion of European ancestry, resulting in different background LD among the SNPs and therefore affecting the power.

Another feature of the proposed method is its flexibility. Our method can be used for both case-control and quantitative-trait data. When quantitative traits are observed, such as blood pressure or blood glucose level, the quantitative information of cases and controls can further im-

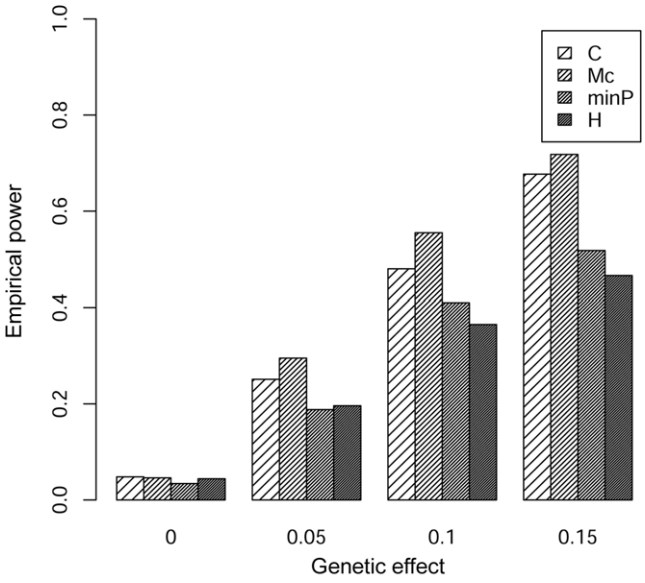

Figure 7. Comparison of the empirical power for 155 controls and 155 cases at different genetic effects between the composite correlation–based LD contrast test (C), the proposed test (Mc), the minimum *P* value in single-marker analysis (minP), and Hotelling's test of the marginal effects of multiple markers (H). The LD pattern is from the real ACE genotype data of 310 subjects from Nigeria. The genetic effect (*X*-axis) is defined as the ratio of the genetic variance to the noise variance.

Table 1. Results of Different LD Contrast Tests for the ACE Level and 13 Variants in the ACE Locus in Samples from Nigeria, Jamaica, and the United States

Population	Sample Size	*P*		
		T_{MC}	T_C	$T_{\Delta'}$
Nigerian	310	.0093	.0689	.6885
Jamaican	116	.0020	.0021	.1902
African American	252	.0738	.0711	.0742

prove the power of the method. Because the score statistic is derived under the framework of a generalized linear model, appropriate covariates, which can effectively control stratification effects that could otherwise invalidate a permutation test,[19] can also be incorporated without difficulty into both the fixed and the random effects. To model pairwise LD, the genotype values of multiple markers are defined as correlated responses, and the trait value is defined as the predictor variable in the generalized linear model. The treatment of multiple markers in a cluster as a dependent variable has already been applied to association studies. Liang et al.[20] applied generalized estimating equations for cluster genotype data. As an alternative, we use a mixed model to be able to separately model background LD and trait-related LD.

We have shown that the statistic T_C, by directly comparing two correlation coefficients between cases and controls, is inefficient for detecting association when background LD is not negligible. It is well known that although valid statistics can be obtained that do not depend on correctly modeling the correlation structure, inappropriate specifications can result in a loss of efficiency. Direct comparison of two correlation coefficients is not optimal in terms of power, in that it does not take into account the correlation structure. The correlation can be looked at as a measure of the similarity between two variables. This argument is related to the more general statistical question of how to measure this similarity efficiently. A similar discussion can also be found among genetic linkage analyses, in which the similarity of trait and markers among family members is of interest.

Recently, Zhao et al.[7] proposed to use the pairwise LD contrast to detect interaction between two loci. With the assumption that the two loci are unlinked, they showed that interaction between two loci indeed generates LD in the disease population and that the LD level generated by interaction depends on the magnitude of the interaction between the two loci. The method proposed in this article is a good complement to their method. Our method can improve the power because of the advantage that the BLUP has in making use of information both across subjects and across SNPs in each region. The proposed method should be especially useful when the LD contrast test is used to detect interaction among variants in LD, such as different variants in a candidate gene.[21]

The proposed method, just like other LD contrast tests, has limitations, as discussed by Zaykin et al.[6] As found in our simulations, the LD contrast test will fail when the allele frequency is low. In this case, the primary association information exists in the difference between the marginal allele frequencies. A summary measure to capture both marginal effect and pairwise effect is desirable. We have found, in our study, that the pairwise Euclidean distance between genotype values among markers could be more powerful than other tests by simultaneously using both sources of information (data not shown). However, the gain in power of this measure depends on the unknown trait model. Further studies are needed to find a test that generally performs well under various reasonable models. In summary, we have improved the LD contrast test by taking into account the background LD. The new test is feasible for handling continuous traits and covariates.

Acknowledgments

We thank Dr. R. S. Cooper for permitting us to access the ACE data. This work was supported in part by a US Public Health Service Resource Grant from the National Center for Research Resources (RR03655), a Research Grant from the National Institute of General Medical Sciences (GM28356), a Cancer Center Support Grant from the National Cancer Institute (P30CAD43703), and a research grant from the National Human Genome Research Institute (HG003054).

Appendix A

Derivation of the Score Statistic

We derive the score statistic for testing H_0:$\delta = 0$. The conditional likelihood function for subject i is

$$L_i(\delta) = \prod_j f(x_{ji}|\mathbf{d}_{ji},\mathbf{u}_{ji}) ,$$

where j denotes the SNP in a cluster genotyped. For convenience of notation, we write $f(x_{ji}|\mathbf{d}_{ji},\mathbf{u}_{ji})$ as $f(\mathbf{d}_j)$, where $\mathbf{d}_j$ is the trait related random effects for subject i. For simplicity, we omit the subject index i.The Taylor series expansion of L about $\mathbf{d}_j = 0$ yields

$$L_i \approx \prod_j f_j(0) + \sum_j \mathbf{d}_j \frac{\partial f_j(0)}{\partial \mathbf{d}_j} \prod_{k\neq j} f_k(0) + \frac{1}{2}\left[\sum_j \mathbf{d}_j^2 \frac{\partial^2 f_j(0)}{\partial \mathbf{d}_j^2} \prod_{k\neq j} f_k(0) + \sum \sum_{j\neq k} \mathbf{d}_j\mathbf{d}_k \frac{\partial f_j(0)}{\partial \mathbf{d}_j} \times \frac{\partial f_k(0)}{\partial \mathbf{d}_k} \prod_{l\neq j,k} f_l(0)\right] ,$$

where *j, k,* and *l* denote different markers. Because the $\mathbf{d}_j$ are not observed, we use the marginal likelihood by taking the expectation over $\mathbf{d}_j$:

$$L_i \approx \prod_j f_j(0) + \frac{1}{2}\left[\sum_j \sigma^2 \frac{\partial^2 f_j(0)}{\partial \mathbf{d}_j^2} \prod_{k \neq j} f_k(0) + \sum \sum_{j \neq k} \sigma^2 \varphi(t_i) \delta \frac{\partial f_j(0)}{\partial \mathbf{d}_j} \times \frac{\partial f_k(0)}{\partial \mathbf{d}_k} \prod_{l \neq j,k} f_l(0)\right] .$$

Let $l_j(\mathbf{d}_j) = \log[f_j(\mathbf{d}_j)]$. We have

$$\frac{\partial f_j(\mathbf{d}_j)}{\partial \mathbf{d}_j} = f_j(\mathbf{d}_j)\left[\frac{\partial l_j(\mathbf{d}_j)}{\partial \mathbf{d}_j}\right]$$

$$\frac{\partial^2 f_j(\mathbf{d}_j)}{\partial \mathbf{d}_j^2} = f_j(\mathbf{d}_j)\left[\frac{\partial^2 l_j(\mathbf{d}_j)}{\partial \mathbf{d}_j^2} + \left(\frac{\partial l_j(\mathbf{d}_j)}{\partial \mathbf{d}_j}\right)^2\right] .$$

Then,

$$L_i \approx \prod_j f_j(0)\left(1 + \frac{1}{2}\sigma^2\left\{\sum_j \left[\frac{\partial^2 l_j(\mathbf{d}_j)}{\partial \mathbf{d}_j^2} + \left(\frac{\partial l_j(\mathbf{d}_j)}{\partial \mathbf{d}_j}\right)^2\right] + \left[\sum \sum_{j \neq k} \varphi(t_i)\delta \frac{\partial l_j(0)}{\partial \mathbf{d}_j} \times \frac{\partial l_k(0)}{\partial \mathbf{d}_k}\right]\right\}\right) .$$

Assuming σ^2 is known, the score statistic is the first derivative with respect to δ evaluated at the null hypothesis that there is no correlation introduced by trait values

$$\frac{\partial \log L_i}{\partial \delta} = \frac{1}{L} \times \frac{\partial L}{\partial \delta} .$$

The likelihood function under the null hypothesis without the trait related random effect is $L = \prod_j f_j(0)$, and then

$$\frac{\partial logL_i}{\partial \delta} = \frac{\partial \sum \sum_{j \neq k} \varphi(t_i)\delta \frac{\partial l_j(0)}{\partial \mathbf{d}_j} \times \frac{\partial l_k(0)}{\partial \mathbf{d}_k}}{\partial \delta} .$$

If x_{ji} follows an exponential family distribution with a canonical link function, we have

$$\frac{\partial l_j(0)}{\partial \mathbf{d}_j} = [x_{ji} - E(x_{ji})]/a(\phi) .$$

Then,

$$U_i = \frac{\partial logL}{\partial \delta} \propto \sum \sum_{j \neq k} \varphi(t_i)[x_{ji} - E(x_{ji})][x_{ki} - E(x_{ki})] .$$

For two SNP markers, the score statistic is then simply given by

$$U = \sum_i [x_{Ai} - E(x_{Ai})][x_{Bi} - E(x_{Bi})]\varphi(t_i) .$$

Without considering the background correlation induced by $\mathbf{u}_i$, $E(x_{ji})$ can be estimated by the sample mean. However, it is often not appropriate to simply omit the background correlation due to various factors, especially in a local region. In this case, we suggest estimating $E(x_{ji})$ by its BLUP.

Web Resource

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for ACE locus)

References

1. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
2. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High resolution haplotype structure in the human genome.

Nat Genet 29:229–232
3. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, et al (2001) Linkage disequilibrium in the human genome. Nature 411: 199–204
4. Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610–622
5. Nielsen DM, Ehm MG, Zaykin DV, Weir BS (2004) Effect of two and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics 168:1029–1040
6. Zaykin DV, Meng Z, Ehm MG (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 78:737–746
7. Zhao J, Jin L, Xiong M (2006) Test for interaction between two unlinked loci. Am J Hum Genet 79:831–845
8. Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67
9. Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. Heredity 42: 105–111
10. Zaykin DV (2004) Bounds and normalization of the composite linkage disequilibrium coefficient. Genet Epidemiol 27:252–257
11. Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19:1–17
12. Xu X, Weiss S, Wei LJ (2000) A unified Haseman-Elston method for testing linkage with quantitative traits. Am J Hum Genet 67:1025–1028
13. Wang T, Elston RC (2004) A modified revisited Haseman-Elston method to further improve power. Hum Hered 57:109–116
14. Zhu X, McKenzie CA, Forrester T, Nickerson DA, Broeckel U, Schunkert H, Doering A, Jacob HJ, Cooper RS, Rieder MJ (2000) Localization of a small genomic region associated with elevated ACE. Am J Hum Genet 67:1144–1153
15. Zhu X, Bouzekri N, Southam L, Cooper RS, Adeyemo A, McKenzie CA, Luke A, Chen G, Elston RC, Ward R (2001) Linkage and association analysis of angiotensin I-converting enzyme (ACE)-gene polymorphisms with ACE concentration and blood pressure. Am J Hum Genet 68:1139–1148
16. Cox R, Bouzekri N, Martin S, Southam L, Hugill A, Golamaully M, Cooper R, Adeyemo A, Soubrier F, Ward R, et al (2002) Angiotensin-1-converting enzyme (ACE) plasma concentration is influenced by multiple ACE-linked quantitative trait nucleotides. Hum Mol Genet 11:2969–2977
17. Bouzekri N, Zhu X, Jiang Y, McKenzie CA, Luke A, Forrester T, Adeyemo A, Kan D, Farrall M, Anderson S, et al (2004) Angiotensin I-converting enzyme polymorphisms, ACE level and blood pressure among Nigerians, Jamaicans and African-Americans. Eur J Hum Genet 12:460–468
18. Zhang J, Rowe WL, Clark AG, Buetow KH (2003) Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am J Hum Genet 73:1073–1081
19. Chen HS, Zhu X, Zhao H, Zhang S (2003) Qualitative semiparametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 67:250–264
20. Liang KY, Hsu FC, Beaty TH, Barnes KC (2001) Multipoint linkage-disequilibrium mapping approach based on the case-parent trio design. Am J Hum Genet 68:937–950
21. Tao H, Cox DR, Frazer KA (2006) Allele-specific KRT1 expression is a complex trait. PLoS Genet 2:e93