Published in final edited form as:

Ann Hum Genet. 2009 July ; 73(Pt 4): 456–464. doi:10.1111/j.1469-1809.2009.00527.x.

Multivariate Association Test Using Haplotype Trend Regression

Yu-Fang Pei^{1,2}, Lei Zhang^{1,2}, Jianfeng Liu², and Hong-Wen Deng^{1,2,3,*}

¹The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P. R. China.

²School of Medicine, University of Missouri-Kansas City, 2411 Holmes Street, Kansas City, MO 64108, USA

³Center of System Biomedical Sciences, Shanghai University of Science and Technology, Shanghai 200093, P. R. China.

Summary

Genetic association analyses with haplotypes may be more powerful than analyses with single markers, under certain conditions. Furthermore, simultaneously considering multiple correlated traits may make use of additional information that would not be considered when analyzing individual traits. In this study, we propose a haplotype based test of association for multivariate quantitative traits in unrelated samples. Specifically, we extend a population based haplotype trend regression (HTR) approach to multivariate scenarios. We mainly focused on bivariate HTR, and the simulation results showed that the proposed method had correct pre-specified type-I error rates. The power of the proposed method was largely influenced by the size and source of correlation between variables, being greatest when correlation of a specific gene was opposite in sign to the residual correlation.

Keywords

Multivariate haplotype trend regression; residual correlation

Introduction

Genetic association analyses can be performed by analyzing either individual SNPs or haplotypes, both of which are commonly implemented in real applications (Martin et al., 2001; Shifman et al., 2002; Guo et al., 2006). It is still debatable which of the two strategies is more powerful in detecting common genetic risk factors. It is likely that the relative superiority of either strategy will depend upon conditions (Bader, 2001; Wessel & Schork, 2006). For example, haplotype based association tests will have improved statistical power compared to tests using single SNPs when haplotypes are directly responsible for observed variations of traits or when a casual variant is strongly associated with a particular haplotype, but not with any individual SNP (Scheet & Stephens, 2006). In experimental studies, it has been shown that analyses with haplotypes successfully detected significant associations while detection of association with single-SNPs was not successful (Fallin et al., 2001; Shifman et al., 2002).

Various haplotype based association methods have been proposed (Zaykin et al., 2002; Zhang et al., 2002; Tan et al., 2005; Browning, 2006; Li et al., 2007). These methods generally focus on analyses of individual traits. In practice, suites of correlated phenotypes are usually characterized; joint consideration of such traits can provide additional information compared to information contained in individual traits. Theoretically, explicitly modeling correlations between phenotypes can provide greater power than that provided by the analysis of individual phenotypes. Earlier studies on joint genetic linkage analyses and association tests of multiple correlated traits have been shown to improve the statistical power to localize and evaluate the effects of genes that jointly influence complex traits (Williams et al., 1999; Soria et al., 2002; Lange et al., 2003; Liu et al., 2008a). However, there has been surprisingly little work done on haplotype based multivariate association analyses.

An issue with regard to haplotype based analyses is the uncertainty of individual haplotype phases. For SNP markers, n heterozygous sites will correspond to 2^n possible haplotypes. Haplotype phasing algorithms usually produce a probability vector for each individual to quantify this uncertainty, but most haplotype based methods take only the most likely haplotypes as the “real” phases, which can result in bias estimation of haplotype effects (Li et al., 2007). An alternative is to take haplotype phasing uncertainty into account and to model on the probability vector rather than on particular haplotypes. This strategy will make more use of the data and thus will have more power than the competition. Such a method, haplotype trend regression (HTR) analysis (Zaykin et al., 2002), has been widely used and has been shown to be efficient in genetic association analyses (Boutin et al., 2003; John et al., 2004; Lohmusaar et al., 2005; Li et al., 2006).

In this study, we propose a method for haplotype based association tests for multivariate quantitative traits. Specifically, we extend the HTR analysis proposed by Zaykin et al. (2002) to multivariate scenarios. The statistical properties of the method will be investigated by simulation studies.

Methods

Model

For simplicity of illustration, we mainly focused on two correlated quantitative traits. Assume a total of N individuals are recorded for two quantitative traits. Let $y_i = (y_{1i}, y_{2i})$, and X_i be the trait values and the design matrix for covariates, respectively, for the i th individual ($i = 1, \dots, N$), and let $\mathbf{Y} = (y_1, \dots, y_N)$ ($2 \times N$) be the trait values for all individuals. Assume there are M SNP markers, which produce the number of $m = 2^M$ possible haplotypes. Genotype data are available for all individuals at all markers. Let $h_i = (h_{1i}, h_{2i}, \dots, h_{mi})'$ be the probability vector for the i th individual possessing these haplotypes, so that $\sum_{l=1}^m h_{li} = 1$. The phenotypes for the i th individual can be modeled as

$$y_i = \mu + X_i \beta_x + \beta_h h_i + \varepsilon_i$$

where $\mu = (\mu_1, \mu_2)'$ are the grand means; β_x are covariate effects; β_h ($2 \times m$) are haplotype effects; and $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i})'$ are random residuals. Following convention, assume that ε_i ($i = 1, \dots, N$) are independently and identically distributed from a bivariate normal distribution with zero means and variance-covariance matrix Σ , so that,

$$E(Y) = \mu + X \beta_x + \beta_h H$$

and $\text{var}(Y) = \Sigma$.

Tests of Association

The set of parameters in the above model is $\theta = [\mu, \beta_X, \beta_h, \Sigma]$. In the null hypothesis of no association with haplotypes, the regression coefficients β_h are restricted to zeros, whereas in the alternative hypothesis β_h are estimated. With the multivariate normality distributions of phenotypes, the likelihood of data is expressible as

$$L(Y, \theta) = (2\pi)^{-N} |\Sigma|^{-N/2} \prod_{i=1}^N \exp \left(-\frac{1}{2} y_i^{*'} \Sigma^{-1} y_i^* \right),$$

where

$$y_i^* = y_i - \mu - \beta_X X_i - \beta_h h_i.$$

. We test the alternative hypothesis by comparing the likelihood under the constraint $\beta_h = 0$ with that under estimated β_h . Define the likelihood ratio as

$$\Lambda = |\hat{\Sigma}| / |\hat{\Sigma}_0|,$$

where $\hat{\Sigma}_0$ and $\hat{\Sigma}$ are the maximal likelihood estimates of Σ under null and alternative hypotheses, respectively. The test statistic is then defined as

$$F = \frac{r s - p q / 2 + 1}{p q} \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}},$$

where q is the rank of $(H'H)$, p is the rank of $\hat{\Sigma}_0$ and $r = v - (p - q + 1)/2$, and

$$s = \begin{cases} \sqrt{\frac{p^2 q^2 - 4}{p^2 + q^2 - 5}} & \text{if } p^2 + q^2 - 5 > 0 \\ 1 & \text{otherwise} \end{cases}.$$

v is the error degree of freedom. Under the null hypothesis, F approximately follows an F -distribution with pq and $rs - pq/2 + 1$ numerator and denominator degrees of freedom.

Data Simulation

Haplotypes representing a 300Kb chromosome region and containing 300 SNPs were generated using the program *ms* (Hudson, 2002) in a coalescent model. Specifically, the effective population size was set to 10,000 and the uniform recombination rate was set to 1.0×10^{-8} per site per generation. From the resulting generated haplotypes, we simulated a sample of 1000 individuals each of which consisted of two randomly selected haplotypes so that the sample size of haplotypes in follow-up analyses would be 2000. To mimic a typical genetic association study, we first excluded rare SNPs ($MAF < 0.03$), and then selected tagging SNPs based on pairwise r^2 value of 0.8 using the software Haploview (Barrett et al., 2005). Finally, 81 SNPs were included for subsequent analyses.

To simulate phenotypes, we selected a haplotype region containing five adjacent SNPs in the middle of the sequence (from 36th to 40th) as susceptible haplotypes to each of two traits. Specifically, for each trait, the phenotype value for individual i was generated from the following expression

$$y_{li} = \sum_{k=1}^5 \lambda_{lk} z_{ki} + \delta_{li} (l=1, 2),$$

where z_{ik} was the 0–1–2 genotype score at the k th causal SNP, λ_{lk} was determined by the allelic frequency and by the magnitude of the effect of the SNP. For a particular SNP z , the genotype ‘11’, ‘12’ and ‘22’ were coded as 0, 1 and 2 respectively, according to additive genetic model. Assuming the frequency of allele ‘1’ in the sample was p , the phenotypic variance caused by the SNP would be $\text{var}(\lambda z) = 2\lambda^2 p(1-p)$. When estimating power, we set total phenotypic variance to 100.0, of which 0.5% was assumed to be explained by individual SNP. Thus,

$$\lambda = \sqrt{\frac{0.5\% \times 100}{2p(1-p)}} = \frac{1}{2\sqrt{p(1-p)}}.$$

Since allelic frequency varied in different samplings, the value of λ also varied accordingly.

In our simulations, we set all λ_{lk} as positive values, which means they were always in the same direction for each phenotype and for each causal SNP, so the causal SNP induced phenotypic correlation was positive. For simplicity of illustration, we here set population means to zero, and omitted the effects of covariates, though incorporation of them was direct and easy. In addition, the two residual effects: δ_{1i} and δ_{2i} , were assumed to have a bivariate normal distribution with correlation coefficient ρ_e ($-1 \leq \rho_e \leq 1$). In this study, we considered seven values of residual correlation coefficients: $\rho_e = 0.9, 0.6, 0.3, 0, -0.3, -0.6$ and -0.9 , respectively.

After phenotypes were simulated, causal SNPs were deleted from the sequence and only genotypes were assumed known for each individual. Association tests were performed on haplotypes with a ‘sliding window’ strategy with fixed window sizes. In each window, probability vectors of haplotypes were inferred using Expectation-Maximization (EM) algorithm (Zaykin et al., 2002). We took the minimal p -values over all windows and adjusted them by Bonferroni correction to form the final p -values over the sequences. Here we studied the performance of the proposed method with window size 1, 3 and 5, respectively. Powers were obtained based on 5,000 replicates in which each causal SNP’s contribution to the total phenotypic variance was set at 0.5%. Type-I error rates were estimated by setting the effects of all causal sites to zero, and were obtained based on 10,000 replicates. For comparison, we also studied the performance of the univariate haplotype trend regression method proposed by Zaykin et al. (2002) and principle component (PC) based univariate haplotype trend regression test. For the univariate haplotype trend regression method, both phenotypes were tested in turn, and then the Bonferroni correction was applied. For PC based univariate haplotype trend regression tests, we took the first PC constructed from the bivariate phenotypes and then applied the univariate test to that. The bivariate, univariate and PC based univariate tests were denoted as BHTR, UHTR and PCHTR, respectively, throughout the study.

In our simulations we assumed the residuals followed a normal distribution. To investigate the importance of the normality assumption, we used the following procedure to generate non-normal correlated residuals with a controlled correlation. Assuming Z_0, X_0, Y_0 were drawn from Gamma distributions,

$$\begin{aligned}\delta_{1i} &= X_0 + Z_0 \\ \delta_{2i} &= Y_0 + Z_0\end{aligned}$$

would make δ_{1i} , δ_{2i} positively correlated with

$$\begin{aligned}\text{cor}(\delta_{1i}, \delta_{2i}) &= \text{var}(Z_0) / \sqrt{(\text{var}(Z_0) + \text{var}(X_0)) \times (\text{var}(Z_0) + \text{var}(Y_0))} \\ \delta_{1i} &= X_0 - Z_0 \\ \delta_{2i} &= Y_0 + Z_0\end{aligned}$$

would make δ_{1i} , δ_{2i} negatively correlated with

$$\begin{aligned}\text{cor}(\delta_{1i}, \delta_{2i}) &= \\ &- \text{var}(Z_0) / \sqrt{(\text{var}(Z_0) + \text{var}(X_0)) \times (\text{var}(Z_0) + \text{var}(Y_0))}\end{aligned}$$

Specifically, the shape parameter k for all three Gamma distributions was set to 1.0 so that all variables deviated from normal distributions substantially. The scale parameter θ for each distribution was set to control variance v by that $v = k\theta^2$. Correlations between phenotypes were also controlled in a similar manner.

To evaluate the performance of the proposed approach when applied to data sets with more than two phenotypes and to investigate the influence of misspecification of variance-covariance structure, we added simulation analyses on trivariate phenotypes under a specific parameter combination (residual correlation coefficient equaled -0.3 and window size equaled 3). Specifically, we simulated trivariate phenotypes with autoregressive, exchangeable and independent variance-covariance structures, respectively. Then we specified the correlation matrix as independent, exchangeable, autoregressive, and unstructured to detect the association. Considering the computational time, we chose to simulate only 100 individuals and only tested the 3 adjacent SNPs close to the causal SNPs.

Results

Type-I Error Rates

Table 1 lists the type-I error rates of UHTR, PCHTR and BHTR under various window sizes and levels of residual correlations between phenotypes. All three tests had error rates that were close to the target level of 5% in all the settings.

Power Estimates

Table 2 lists power estimates of the three methods under the scenario that the residuals followed normal distribution and both phenotypes were associated with the causal locus. The powers of UHTR, PCHTR and BHTR became higher as window size increased within the range (1–5) investigated. For example, when the residual correlation coefficient was set to 0, the power increased from 38.8% to 73.3% for UHTR, from 41.9% to 67.6% for PCHTR and from 51.7% to 76.4% for BHTR, as window size increased from 1 to 5. For BHTR, its power increased as the residual correlation decreased from 0.9 to -0.9 . In additional simulations where major-gene correlation between phenotypes was constrained to -1.0 , we observed an opposite trend in that the power increased as ρ_e increased from -0.9 to $+0.9$ (data not shown). PCHTR had improved power over UHTR and, in many cases, over BHTR when the residual correlation was positive. In contrast, when the residual correlation coefficient was negative, PCHTR had largely

decreased power that actually approached type-I error rates. For example, in the case where the residual correlation coefficient was equal to -0.3 and window size was set to 3, PCHTR had a power of only 3.9%, while values for UHTR and BHTR were 53.8% and 86.1%, respectively. Among those methods, when the residual correlation was the same in sign as the correlation at the causal locus, the power from highest to lowest was: PCHTR, UHTR and BHTR. In contrast, when the residual correlation was opposite in sign to the correlation at causal locus, BHTR had the highest power, followed by UHTR and PCHTR.

Table 3 lists the power estimates of the three methods under the scenario that only one phenotype was associated with the causal locus. The power of BHTR had a trend of increasing with a rising absolute value of residual correlation from 0 to 0.9, regardless of its direction, while the power of PCHTR had a trend of decreasing with a rising absolute value of residual correlation from 0 to 0.9, regardless of its direction. PCHTR had the lowest power under all situations. When the residual correlation coefficient was close to 0.0, BHTR had approximately equal powers to UHTR. Otherwise, BHTR instead had improved power, with an increase in power improvement as the absolute value of residual correlation coefficient moved towards 1.0.

Table 4 lists the power estimates under the scenario that the distribution of phenotypes departed from a normal distribution. The results showed that non-normality distributed phenotypes did cause some loss of power, illustrating the importance of the normality assumption of phenotypes.

Table 5 lists the power estimates for trivariate scenarios. The model when correctly specifying variance/covariance structure had the highest power; misspecification of the model can cause some loss of power. Without specifying any form of variance-covariance matrix, unstructured models had power estimates that approximated those attained under correct models for all simulations.

As we focused on BHTR, we also analyzed the results from a randomly selected replicate of simulation data, when setting the residual correlation coefficient to -0.6 , as displayed in Figure 1. Both BHTR and UHTR correctly detected the association signal. However, the $-\log(p\text{-value})$ from BHTR at the causal sites were higher than those from UHTR, especially as window size increased.

Discussion

In this study, we propose an approach to test the association between haplotypes and multivariate quantitative traits. This approach has the potential to improve on statistical power by joint analysis of suites of correlated phenotypes, and is appealing to real studies in which multiple correlated traits are collected. Examples of such studies include: 1) longitudinal studies, e.g., Framingham heart study (FHS), where each subject is measured for a single trait more than one time at distinct time points; 2) investigation of genetic etiologies underlying susceptibility to common disorders. Genetic mapping for common disorders, such as diabetes (Wiltshire et al., 2002), asthma (Cookson, 2002), cardiovascular disease (Mitchell et al., 1996), and developmental dyslexia (Fisher & Defries, 2002), often depend on the use of a number of related indices of severity and no single measure fully reflects the complex etiology of these diseases. In such cases, univariate analyses may have some drawbacks, such as, how best to adjust for the multiple testing of correlated measures in respective univariate analyses and how to interpret and integrate data from univariate analyses of different trait measures. Previous linkage studies have demonstrated that the results of multivariate linkage analyses for these complex traits are substantially clearer than those of univariate analyses of the same data, helping to resolve the above issues (Marlow et al., 2003). In the vast majority of previous

association studies involving multiple correlated trait measures, each measure had been analyzed independently (Guo et al., 2006; Liu et al., 2008b). Our proposed method provides novel and powerful analyses to appropriately analyze such correlated complex traits in an integrated manner.

Another aspect that we considered to improve statistical power was to use haplotypes rather than single markers, since the former are shown to outperform the latter under certain conditions (Balding, 2006; Waldron et al., 2006). Notably, we take haplotype phase uncertainty into account in our model, which should make better use of data.

When the causal SNP induced correlation was 1.0, the power of BHTR increased as residual correlation decreased, and it exceeded the power of UHTR in several major settings investigated. Our simulation results are consistent with that of previous bivariate linkage and association analyses: bivariate analyses become more powerful when the gap between major-gene correlation and residual correlation widens, e.g. the power of bivariate tests is greater than that of univariate tests when the genetic and residual induced phenotypic correlations are in opposite directions (Amos et al., 2001; Evans, 2002; Evans & Duffy, 2004; Liu et al., 2008a). Evans' study gave theoretical interpretations on why the sign of correlation between variables affects the power of bivariate analyses (Evans, 2002). The rationale lies in that the opposite sign between causal site-induced and residual correlation can result in a corresponding positive term involved in the expression of the non-centrality parameter (NCP), increasing the magnitude of the NCP and the power; otherwise, a negative term of the NCP arises, decreasing the NCP and the power. Although the above interpretation is originally for bivariate linkage analyses, from a statistical perspective, it should have a similar effect in association studies (Evans, 2002; Liu et al., 2008a). When only one phenotype was associated with the causal site, the major-gene correlation was essentially equal to 0.0, and consequently, bivariate approaches had increased power as residual correlation moved from 0.0 to either +0.9 or -0.9.

In real datasets, as causal loci usually contribute a small proportion to the total phenotypic variance, residual correlations approximate phenotypic correlations between traits. For genes influencing both phenotypes, bivariate approaches always seem to be valuable unless the phenotypes were highly correlated in the same direction as major-gene correlations, corresponding to the value of ρ close to either +1.0 or -1.0 depending on the unknown major-gene effects. For genes influencing only one phenotype, bivariate approaches were still appealing unless the phenotypes were weakly correlated, corresponding to the value of ρ close to 0.0.

For PCHTR, when the residual correlation was positive, its power was higher than that of UHTR. In contrast, when the residual correlation coefficient was negative, the first PC contains no information regarding locus effects, consequently it had largely decreased power that actually approached type-I error rates (refer to Appendix I for detailed explanation). When only one phenotype was associated with the causal site, the power of PCHTR had a trend of decreasing with a rising absolute value of residual correlation from 0 to 0.9, regardless of its direction. This may be because the contribution of the first PC increased with a rising absolute value of residual correlation from 0 to 0.9, which results in a decrease of the proportion of the total variance attributable to the causal SNP. In conclusion, in most real applications where no prior knowledge on the relative direction between major-gene correlation and residual correlation is available, we suggest a cautious approach when using principal components to analyze data, since this may result in a great loss of power.

A very real topic of concern when parametrically modeling a multivariate distribution is misspecification of the variance/ covariance matrix. Four kinds of variance-covariance structures are commonly modeled when analyzing multivariate phenotypes: autoregressive,

exchangeable, independent and unstructured. Which kind of structure is preferable depends on the context of the data set. For example, for repeatedly measured phenotypes, autoregressive structure is a reasonable approximation. For data sets without prior knowledge on the relationship between phenotypes, an unstructured approach can be used in which the variance/covariance matrix is estimated empirically from the sample. In this study, we evaluated the performance of the proposed approach when applied to data sets with three phenotypes in order to investigate the influence of misspecification of variance/covariance structure. The simulation results showed that the model in which variance/covariance structure was correctly specified had the highest power; misspecification of the model could cause some loss of power. Without specifying any form of variance-covariance matrix, an unstructured model had power estimates that approximated those attained under correct models for all simulations.

As it is still a common issue to reach the global maximum of the likelihood when the number of parameters is large, we suggest our proposed approach be applied to data sets with a moderate number of phenotypes and window sizes. One potential improvement of the method is grouping of 'similar' haplotype categories to reduce the number of parameters that need to be estimated. Several haplotype clustering methods have been proposed (Waldron et al., 2006; Browning & Browning, 2007) and further studies are needed to incorporate these methods into BHTR to improve power.

Acknowledgements

The study was partially supported by Xi'an Jiaotong University. The investigators of this work also benefited from grants from the Ministry of Education of China, NIH (R01 AR050496, R21 AG 027110, R01 AG026564, R21 AA015973 and P50 AR055081) and National Science Foundation of China. Computing support was partially provided by the High Performance Computing Cluster Center at Xi'an Jiaotong University. We also wish to acknowledge Dr. Christopher J. Papasian for his editorial assistance during the preparation of this manuscript.

Reference

- Amos C, De Andrade M, Zhu D. Comparison of multivariate tests for genetic linkage. *Hum Hered* 2001;51:133–144. [PubMed: 11173964]
- Bader JS. The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* 2001;2:11–24. [PubMed: 11258193]
- Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7:781–791. [PubMed: 16983374]
- Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21:263–265. [PubMed: 15297300]
- Boutin P, Dina C, Vasseur F, Dubois S, Corset L, Seron K, Bekris L, Cabellon J, Neve B, Vasseur-Delannoy V, Chikri M, Charles MA, Clement K, Lernmark A, Froguel P. GAD2 on chromosome 10p12 is a candidate gene for human obesity. *PLoS Biol* 2003;1:E68. [PubMed: 14691540]
- Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol* 2007;31:365–375. [PubMed: 17326099]
- Browning SR. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 2006;78:903–913. [PubMed: 16685642]
- Cookson WO. Asthma genetics. *Chest* 2002;121:7S–13S. [PubMed: 11893653]
- Evans DM. The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. *Am J Hum Genet* 2002;70:1599–1602. [PubMed: 11992269]
- Evans DM, Duffy DL. A simulation study concerning the effect of varying the residual phenotypic correlation on the power of bivariate quantitative trait loci linkage analysis. *Behav Genet* 2004;34:135–141. [PubMed: 14755178]

- Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* 2001;11:143–151. [PubMed: 11156623]
- Fisher SE, Defries JC. Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nat Rev Neurosci* 2002;3:767–780. [PubMed: 12360321]
- Guo Y, Xiong DH, Yang TL, Guo YF, Recker RR, Deng HW. Polymorphisms of estrogen-biosynthesis genes CYP17 and CYP19 may influence age at menarche: a genetic association study in Caucasian females. *Hum Mol Genet* 2006;15:2401–2408. [PubMed: 16782804]
- Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002;18:337–338. [PubMed: 11847089]
- John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, Eyre S, Jones KW, Ollier W, Silman A, Gibson N, Worthington J, Kennedy GC. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am J Hum Genet* 2004;75:54–64. [PubMed: 15154113]
- Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 2003;4:195–206. [PubMed: 12925516]
- Li M, Atmaca-Sonmez P, Othman M, Branham KE, Khanna R, Wade MS, Li Y, Liang L, Zarepari S, Swaroop A, Abecasis GR. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat Genet* 2006;38:1049–1054. [PubMed: 16936733]
- Li Y, Sung WK, Liu JJ. Association mapping via regularized regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. *Am J Hum Genet* 2007;80:705–715. [PubMed: 17357076]
- Liu J, Pei Y, Papasian CJ, Deng HW. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet Epidemiol* 2008a; 33:217–227. [PubMed: 18924135]
- Liu YJ, Liu XG, Wang L, Dina C, Yan H, Liu JF, Levy S, Papasian CJ, Drees BM, Hamilton JJ, Meyre D, Delplanque J, Pei YF, Zhang L, Recker RR, Froguel P, Deng HW. Genome-wide association scans identified CTNBL1 as a novel gene for obesity. *Hum Mol Genet* 2008b;17:1803–1813. [PubMed: 18325910]
- Lohmusaar E, Gschwendtner A, Mueller JC, Org T, Wichmann E, Hamann G, Meitinger T, Dichgans M. ALOX5AP gene and the PDE4D gene in a central European population of stroke patients. *Stroke* 2005;36:731–736. [PubMed: 15731479]
- Marlow AJ, Fisher SE, Francks C, Macphie IL, Cherny SS, Richardson AJ, Talcott JB, Stein JF, Monaco AP, Cardon LR. Use of multivariate linkage analysis for dissection of a complex cognitive trait. *Am J Hum Genet* 2003;72:561–570. [PubMed: 12587094]
- Martin ER, Scott WK, Nance MA, Watts RL, Hubble JP, Koller WC, Lyons K, Pahwa R, Stern MB, Colcher A, Hiner BC, Jankovic J, Ondo WG, Allen FH Jr. Goetz CG, Small GW, Masterman D, Mastaglia F, Laing NG, Stajich JM, Ribble RC, Booze MW, Rogala A, Hauser MA, Zhang F, Gibson RA, Middleton LT, Roses AD, Haines JL, Scott BL, Pericak-Vance MA, Vance JM. Association of single-nucleotide polymorphisms of the tau gene with late-onset Parkinson disease. *JAMA* 2001;286:2245–2250. [PubMed: 11710889]
- Mitchell BD, Kammerer CM, Blangero J, Mahaney MC, Rainwater DL, Dyke B, Hixson JE, Henkel RD, Sharp RM, Comuzzie AG, Vandeberg JL, Stern MP, Maccluer JW. The San Antonio Family Heart Study. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. *Circulation* 1996;94:2159–2170. [PubMed: 8901667]
- Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006;78:629–644. [PubMed: 16532393]
- Shifman S, Bronstein M, Sternfeld M, Pisante-Shalom A, Lev-Lehman E, Weizman A, Reznik I, Spivak B, Grisaru N, Karp L, Schiffer R, Kotler M, Strous RD, Swartz-Vanetik M, Knobler HY, Shinar E, Beckmann JS, Yakir B, Risch N, Zak NB, Darvasi A. A highly significant association between a COMT haplotype and schizophrenia. *Am J Hum Genet* 2002;71:1296–1302. [PubMed: 12402217]

- Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A, Martinez-Marchan E, Mateo J, Borrell M, Stone W, Lathrop M, Fontcuberta J, Blangero J. A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. *Am J Hum Genet* 2002;70:567–574. [PubMed: 11805911]
- Tan Q, Christiansen L, Christensen K, Bathum L, Li S, Zhao JH, Kruse TA. Haplotype association analysis of human disease traits using genotype data of unrelated individuals. *Genet Res* 2005;86:223–231. [PubMed: 16454861]
- Waldron ER, Whittaker JC, Balding DJ. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol* 2006;30:170–179. [PubMed: 16385468]
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* 2006;79:792–806. [PubMed: 17033957]
- Williams JT, Begleiter H, Porjesz B, Edenberg HJ, Foroud T, Reich T, Goate A, Van Eerdewegh P, Almasy L, Blangero J. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. II. Alcoholism and event-related potentials. *Am J Hum Genet* 1999;65:1148–1160. [PubMed: 10486334]
- Wiltshire S, Frayling TM, Hattersley AT, Hitman GA, Walker M, Levy JC, O'rahilly S, Groves CJ, Menzel S, Cardon LR, McCarthy MI. Evidence for linkage of stature to chromosome 3p26 in a large U.K. Family data set ascertained for type 2 diabetes. *Am J Hum Genet* 2002;70:543–546. [PubMed: 11753821]
- Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002;53:79–91. [PubMed: 12037407]
- Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002;71:1386–1394. [PubMed: 12439824]

Appendix

Appendix I

To interpret such varying performance of PC-based tests when in different patterns of correlation, we provide the following analytical explanation (for bivariate phenotypes).

By omitting phenotype means and covariate effects, we rewrite the linear models underlying our simulations

$$\begin{aligned} y_1 &= x + \varepsilon_1 \\ y_2 &= x + \varepsilon_2 \end{aligned} ,$$

where x denote locus effects, and ε_1 and ε_2 are residuals (locus effects contribute equally to both traits in our simulations). Also, $\text{var}(\varepsilon_1) = \text{var}(\varepsilon_2) = \text{var}(\varepsilon)$. The first principal component, $y_{pc} = a_1 y_1 + a_2 y_2$, is the combination of the two phenotypes that accounts for the maximal amount of phenotypic variation under the condition $a_1^2 + a_2^2 = 1$. After some manipulation, it can be shown that

$$\begin{aligned} \text{var}(y_{pc}) &= \text{var}(x) + \text{var}(\varepsilon) + [2\text{var}(x) + 2\rho\text{var}(\varepsilon)] a_1 a_2 \\ &\approx \text{var}(x) + \text{var}(\varepsilon) + 2\rho a_1 a_2 \text{var}(\varepsilon), \end{aligned}$$

since $|\rho\text{var}(\varepsilon)| \geq 0.3 \times 0.975 \text{var}(y_{pc}) \approx 0.3 \text{var}(y_{pc}) \gg 2 \text{var}(x) = 0.05 \text{var}(y_{pc})$.

When $\rho > 0$, $\text{var}(y_{pc})$ is maximized at $a_1 = a_2 = \sqrt{2}/2$, in which case the principal component is the sum of the two phenotypes, and should perform well. When $\rho < 0$, on the other hand, $\text{var}(y_{pc})$ is maximized at $a_1 = -a_2 = \sqrt{2}/2$. Then, $y_{pc} = a_1 (\varepsilon_1 - \varepsilon_2)$, which contains no information

regarding locus effects; consequently it had largely decreased power that actually approached type-I error rates.

Figure 1.

Sample $-\log(p\text{-values})$ against the marker number using $p\text{-values}$ from the likelihood ratio test (LRT). The 36th to the 40th SNPs were causal SNPs. A is for window-size of 1; B is for window-size of 3 and C is for window-size of 5.

Table 1
Type-I Error Rates (A is for normal phenotypes and B is for non-normal phenotypes).

A.												
Residual Correlation	Window-Size = 1			Window-Size = 3			Window-Size = 5					
	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR
0.90	5.0	2.9	4.0	5.9	4.0	5.3	4.5	3.6	5.6	4.5	3.6	5.6
0.60	5.2	4.4	6.0	3.2	3.3	4.9	5.9	4.2	5.2	5.9	4.2	5.2
0.30	5.5	4.7	5.2	4.2	4.9	5.1	4.4	3.4	4.3	4.4	3.4	4.3
0.00	3.8	4.1	5.0	4.8	4.1	4.2	5.1	3.9	4.4	5.1	3.9	4.4
-0.30	4.2	3.8	5.1	4.1	4.0	5.2	4.9	4.4	6.0	4.9	4.4	6.0
-0.60	5.0	4.7	5.4	5.5	3.1	5.3	4.3	3.6	4.3	4.3	3.6	4.3
-0.90	4.4	4.5	4.6	5.9	4.5	5.4	4.7	4.7	3.5	4.7	4.7	3.5
B.												
Residual Correlation	Window-Size = 1			Window-Size = 3			Window-Size = 5					
	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR
0.90	3.9	5.0	4.8	3.6	4.4	4.8	3.0	3.3	3.7	3.0	3.3	3.7
0.60	4.4	4.1	4.9	3.6	4.0	3.5	4.2	3.9	3.2	4.2	3.9	3.2
0.30	4.9	3.2	3.9	3.7	3.3	4.8	4.1	4.0	4.4	4.1	4.0	4.4
0.00	4.6	4.2	4.0	3.9	4.0	3.9	3.6	3.3	3.3	3.6	3.3	3.3
-0.30	4.5	3.2	5.0	3.7	3.5	4.6	3.6	4.4	4.4	3.6	4.4	4.4
-0.60	3.5	4.5	4.5	3.2	4.5	4.0	2.9	3.1	5.0	2.9	3.1	5.0
-0.90	4.0	4.9	4.2	3.5	4.5	3.6	2.9	3.3	3.7	2.9	3.3	3.7

Note: 1000 individuals were simulated under each parameter setting. Type-I error rates were estimated based on 10,000 replications. 'UHTR' denotes the univariate haplotype trend regression testing, 'PCHTR' denotes the principle component based univariate haplotype trend regression testing and 'BHTR' denotes the bivariate haplotype trend regression testing.

Table 2
Power comparison under the scenario that both phenotypes are associated with the causal locus.

Residual Correlation	Window-Size = 1			Window-Size = 3			Window-Size = 5		
	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR
0.90	31.0	35.3	25.5	42.7	48.1	35.1	55.1	58.4	47.7
0.60	38.2	42.2	31.8	37.9	50.0	43.2	54.5	57.6	49.2
0.30	37.3	50.5	37.6	51.9	63.2	51.4	63.7	73.2	62.0
0.00	38.8	41.9	51.7	59.7	55.3	66.5	73.3	67.6	76.4
-0.30	38.9	5.0	70.1	53.8	3.9	86.1	66.0	4.0	92.3
-0.60	39.6	4.7	97.2	54.9	4.3	99.5	64.5	3.9	99.9
-0.90	41.1	4.1	100.0	54.7	3.9	100.0	65.1	4.4	100.0

Note: 1000 individuals were simulated under each parameter setting. Power was estimated based on 5000 replications. 'UHTR' denotes the univariate haplotype trend regression testing, 'PCHTR' denotes the principle component based univariate haplotype trend regression testing and 'BHTR' denotes the bivariate haplotype trend regression testing.

Table 3
Power comparison under the scenario that only one phenotype is associated with the causal locus.

Residual Correlation	Window-Size = 1			Window-Size = 3			Window-Size = 5		
	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR
0.90	20.9	8.4	98.4	27.2	9.2	99.5	31.0	6.1	99.9
0.60	23.5	8.0	41.5	29.5	8.9	43.5	29.3	7.0	50.5
0.30	22.1	12.2	24.6	26.5	9.9	26.2	33.2	11.6	29.9
0.00	21.6	15.3	21.3	28.0	17.3	25.2	31.6	16.4	24.0
-0.30	22.2	11.9	23.7	29.2	11.9	29.4	30.4	10.8	28.3
-0.60	23.8	8.0	35.9	29.3	8.4	46.6	33.5	10.7	50.1
-0.90	25.5	10.1	98.2	27.4	8.9	99.6	32.4	9.1	99.9

Note: 1000 individuals were simulated under each parameter setting. Power was estimated based on 5000 replications. 'UHTR' denotes the univariate haplotype trend regression testing, 'PCHTR' denotes the principle component based univariate haplotype trend regression testing and 'BHTR' denotes the bivariate haplotype trend regression testing.

Table 4
Power comparison under the scenario that the phenotypes deviate from normal distributions.

Residual Correlation	Window-Size = 1			Window-Size = 3			Window-Size = 5		
	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR	UHTR	PCHTR	BHTR
0.90	14.0	16.4	10.9	24.6	28.0	20.4	37.3	39.0	28.3
0.60	15.6	21.0	13.5	28.7	34.1	25.2	37.8	42.2	30.0
0.30	15.6	25.2	17.2	31.1	41.3	31.2	42.3	54.2	41.1
0.00	17.3	21.6	28.9	38.0	37.0	47.0	53.2	52.2	62.1
-0.30	16.2	3.7	46.8	31.5	3.7	71.4	45.1	4.7	84.1
-0.60	16.7	3.9	88.8	32.5	4.9	98.0	43.3	5.1	99.6
-0.90	18.8	3.4	100.0	33.8	5.2	100.0	43.9	5.3	100.0

Note: 1000 individuals were simulated under each parameter setting. Power was estimated based on 5000 replications. 'UHTR' denotes the univariate haplotype trend regression testing, 'PCHTR' denotes the principle component based univariate haplotype trend regression testing and 'BHTR' denotes the bivariate haplotype trend regression testing.

Table 5

Power of trivariate haplotype trend regression analyses under different model specification.

Real\Tested	Independent	Exchangeable	Autoregressive	Unstructured
Independent	23.8	19.9	12.2	20.3
Exchangeable	17.1	23.4	11.7	20.2
Autoregressive	17.3	18.1	22.5	19.6