Joint association analysis of bivariate quantitative and qualitative traits

Mengdie Yuan, Guoqing Diao

From Genetic Analysis Workshop 17
Boston, MA, USA. 13-16 October 2010

Abstract

Univariate genome-wide association analysis of quantitative and qualitative traits has been investigated extensively in the literature. In the presence of correlated phenotypes, it is more intuitive to analyze all phenotypes simultaneously. We describe an efficient likelihood-based approach for the joint association analysis of quantitative and qualitative traits in unrelated individuals. We assume a probit model for the qualitative trait, under which an unobserved latent variable and a prespecified threshold determine the value of the qualitative trait. To jointly model the quantitative and qualitative traits, we assume that the quantitative trait and the latent variable follow a bivariate normal distribution. The latent variable is allowed to be correlated with the quantitative phenotype. Simultaneous modeling of the quantitative and qualitative traits allows us to make more precise inference on the pleiotropic genetic effects. We derive likelihood ratio tests for the testing of genetic effects. An application to the Genetic Analysis Workshop 17 data is provided. The new method yields reasonable power and meaningful results for the joint association analysis of the quantitative trait Q1 and the qualitative trait disease status at SNPs with not too small MAF.

Background

Statistical methods for the univariate association analysis of quantitative and qualitative traits have been well developed in the literature. Complex human diseases are often characterized by multiple traits. These traits tend to be correlated with each other because of common environmental and genetic factors. In the genetic analysis of complex diseases, it is natural to account for the correlations among multiple traits and to model them simultaneously. Joint genetic linkage analysis of multiple correlated phenotypes has been studied by Jiang and Zeng [1], Mangin et al. [2], Amos and Laing [3], Almasy et al. [4], Blangero et al. [5], Wijsman and Amos [6], and Williams et al. [7,8], among others. Joint linkage analysis of multiple correlated traits can potentially improve the power to detect linkage signals at genes that jointly influence a complex disease.

Recently, Liu et al. [9] developed an extended generalized estimating equation method for the bivariate association analysis of continuous and binary traits. Their simulation results demonstrated that, compared with univariate analysis, bivariate analysis may substantially improve power while having comparable type I error rates under certain situations.

In this paper we extend the joint linkage analysis of multivariate qualitative and quantitative traits described by Williams et al. [7,8] to association analysis. Specifically, we assume that a latent variable determines the qualitative trait and that the latent variable and the quantitative trait follow a bivariate normal distribution. With such modeling, we develop likelihood-based inference procedures for testing pleiotropic genetic effects. As an illustration, we perform the joint association analysis of the quantitative trait Q1 and the qualitative trait disease status on chromosome 13 from the Genetic Analysis Workshop 17 (GAW17) data.

Methods

Suppose that the data contain \( n \) independent individuals. Let \( Y_{i1} \) be the qualitative trait and \( Y_{i2} \) be the quantitative trait for the \( i \)th individual. Let \( X_i \) denote a vector of...
covariates, including the intercept and environmental variables, and let $Z$ denote a vector of genotype scores at the major single-nucleotide polymorphism (SNP) locus. We may also include gene by environment interaction terms. We assume that $Y_{i1}$ is determined by a latent continuous variable $\tilde{Y}_{i1}$ such that:

$$Y_{i1} = \begin{cases} 0, & \tilde{Y}_{i1} \leq \alpha, \\ 1, & \tilde{Y}_{i1} > \alpha. \end{cases}$$

(1)

We consider the following model:

$$
\begin{pmatrix}
Y_{i1}^* \\
Y_{i2}^*
\end{pmatrix} = 
\begin{pmatrix}
X_i^T \beta_1 + Z_i^T \gamma_1 \\
X_i^T \beta_2 + Z_i^T \gamma_2
\end{pmatrix} + 
\begin{pmatrix}
e_{i1} \\
e_{i2}
\end{pmatrix},
$$

(2)

where the $\beta$ are regression coefficients for the environmental effects, the $\gamma$ are regression coefficients for the genetic effects, and the subscripts 1 and 2 denote the qualitative and quantitative traits, respectively, and the $e_i = (e_{i1}, e_{i2})$ are independent and identically distributed bivariate normal random variables with mean 0 and variance-covariance matrix $\Sigma$:

$$
\Sigma = \begin{pmatrix}
\sigma_{\epsilon 1}^2 & \rho \sigma_{\epsilon 1} \sigma_{\epsilon 2} \\
\rho \sigma_{\epsilon 1} \sigma_{\epsilon 2} & \sigma_{\epsilon 2}^2
\end{pmatrix},
$$

(3)

where $\sigma_{\epsilon 1}^2$ is the variance of $e_{i1}$, $\sigma_{\epsilon 2}^2$ is the variance of $e_{i2}$, and $\rho$ is the correlation between $e_{i1}$ and $e_{i2}$.

To ensure the identifiability of the model, we fix $\sigma_1 = 1$ and $\alpha = 0$.

Given $Y_{i2}$, it can be shown that $Y_{i1}^*$ follows a normal distribution with mean

$$
\mu_{e_{i1}} = X_i^T \beta_1 + Z_i^T \gamma_1 + \frac{\rho \sigma_{\epsilon 1}}{\sigma_2} \left(Y_{i2} - X_i^T \beta_2 - Z_i^T \gamma_2 \right),
$$

(4)

and variance:

$$
\sigma_{e_{i1}}^2 = (1 - \rho^2) \sigma_{\epsilon 1}^2.
$$

(5)

Therefore the likelihood function for the unknown parameters $\theta = (\beta_1, \gamma_1, \beta_2, \gamma_2, \sigma_2, \rho)$ takes the form:

$$
L_n(\theta) = \prod_{i=1}^n \left\{ \frac{1}{(2\pi \sigma_2)^{1/2}} \exp \left\{ -\frac{(Y_{i2} - X_i^T \beta_2 - Z_i^T \gamma_2)^2}{2\sigma_2^2} \right\} ight\} p_i^{y_{i1}} (1 - p_i)^{1-y_{i1}},
$$

(6)

where:

$$
p_i = 1 - \Phi \left( - \frac{\mu_{e_{i1}}}{\sigma_{e_{i1}}} \right),
$$

(7)

and $\Phi$ is the cumulative distribution function of the standard normal distribution.

To obtain the maximum-likelihood estimator (MLE) of $\theta$, denoted $\hat{\theta}$, we maximize the log-likelihood by using the quasi-Newton algorithm described by Press et al. [10]. We can perform various association analyses through the likelihood ratio test (LRT) statistic, given by:

$$
LRT = -2 \left[ \log L_n(\hat{\theta}) - \log L_n(\theta) \right],
$$

(8)

where $\hat{\theta}$ is the MLE of $\theta$ under the null hypothesis. The LRT follows a chi-square distribution asymptotically with the degrees of freedom being the difference of the number of free parameters under the null hypothesis and the number under the alternative hypothesis.

**Results**

We first conducted a small simulation study to evaluate the type I error rates of the association test based on the univariate analysis at significance levels of 0.01 and 0.05. Figure 1 presents the results based on 100,000 replicates, where the $x$-axis denotes the correlation between the two Wald test statistics for the two traits. Figure 1 implies that failing to account for the correlation between traits may lead to inflated type I error rate.

We next performed a joint association analysis of the quantitative trait Q1 and the qualitative trait disease status from the GAW17 data. Before conducting the analyses, we had knowledge of the answers. Sex, Age, and Smoking were included as covariates in the model. We evaluated the additive effect of each of the 11 true SNPs on chromosome 13 on each individual trait as well as the pleiotropic effect on both traits. The genotype score was coded as the number of minor alleles at each SNP locus. For comparison, we also performed a univariate association analysis and used the sum of the LRT statistics for testing genetic effect on an individual trait as the test statistic for testing the pleiotropic effect.

Table 1 presents the powers of the tests at the 11 true SNPs on chromosome 13 from the joint association analysis and the powers from the univariate analysis based on 200 replicates. Table 2 presents the type I error rates at 6 randomly chosen SNPs on chromosome 13, excluding the 11 true SNPs. None of these 6 SNPs was in linkage disequilibrium with the 11 true SNPs on chromosome 13. Type I error rates and powers of the association tests appear to depend very much on the minor allele frequency (MAF). Type I error rates of both tests are lower than the nominal significance level when the MAF is small. Essentially no tests have power to detect the SNP effect when the MAF is low and the minor allele is detected in only a few of the 697 individuals. On the other hand, both methods had good power to detect the pleiotropic
Figure 1 Type I error rates of the univariate association test of pleiotropic effects at the significance levels of 0.01 and 0.05.

Table 1 Power of the association tests at a significance level of 5%

| SNP     | MAF   | Joint analysis |                  | Univariate analysis |                  |
|---------|-------|----------------|-------------------|---------------------|-------------------|
|         |       | $H_0: \gamma_1 = 0$ | $H_0: \gamma_2 = 0$ | $H_0: \gamma_1 = \gamma_2 = 0$ | $H_0: \gamma_1 = 0$ | $H_0: \gamma_2 = 0$ | $H_0: \gamma_1 = \gamma_2 = 0$ |
| C13S399 | 0.0007 | 0.045          | 0.020            | 0.010               | 0.040             | 0.060             | 0.020               | 0.030             |
| C13S479 | 0.0007 | 0.000          | 0.045            | 0.025               | 0.040             | 0.045             | 0.110               |
| C13S505 | 0.0007 | 0.030          | 0.025            | 0.015               | 0.030             | 0.025             | 0.040               |
| C13S514 | 0.0007 | 0.000          | 0.045            | 0.010               | 0.000             | 0.045             | 0.065               |
| C13S547 | 0.0007 | 0.050          | 0.020            | 0.050               | 0.075             | 0.020             | 0.050               |
| C13S567 | 0.0007 | 0.000          | 0.025            | 0.010               | 0.000             | 0.025             | 0.015               |
| C13S320 | 0.0014 | 0.020          | 0.025            | 0.010               | 0.005             | 0.025             | 0.050               |
| C13S324 | 0.0043 | 0.330          | 1.000            | 1.000               | 0.335             | 1.000             | 1.000               |
| C13S431 | 0.0172 | 0.210          | 1.000            | 0.955               | 0.185             | 1.000             | 0.985               |
| C13S522 | 0.0279 | 0.770          | 1.000            | 1.000               | 0.760             | 1.000             | 1.000               |
| C13S523 | 0.0667 | 0.960          | 1.000            | 1.000               | 0.965             | 1.000             | 1.000               |

$\gamma_1$ is the additive SNP effect on the qualitative trait. $\gamma_2$ is the additive SNP effect on the quantitative trait.

Table 2 Type I error rate of the association tests at a significance level of 5%

| SNP     | MAF   | Joint analysis |                  | Univariate analysis |                  |
|---------|-------|----------------|-------------------|---------------------|-------------------|
|         |       | $H_0: \gamma_1 = 0$ | $H_0: \gamma_2 = 0$ | $H_0: \gamma_1 = \gamma_2 = 0$ | $H_0: \gamma_1 = 0$ | $H_0: \gamma_2 = 0$ | $H_0: \gamma_1 = \gamma_2 = 0$ |
| C13S722 | 0.0007 | 0.000          | 0.030            | 0.010               | 0.000             | 0.030             | 0.025               |
| C13S202 | 0.0022 | 0.005          | 0.065            | 0.055               | 0.005             | 0.065             | 0.055               |
| C13S556 | 0.0022 | 0.035          | 0.020            | 0.025               | 0.050             | 0.020             | 0.025               |
| C13S1580| 0.0022 | 0.035          | 0.025            | 0.015               | 0.040             | 0.025             | 0.025               |
| C13S398 | 0.0122 | 0.050          | 0.050            | 0.050               | 0.050             | 0.050             | 0.060               |
| C13S1004| 0.0122 | 0.050          | 0.030            | 0.060               | 0.050             | 0.030             | 0.050               |

$\gamma_1$ is the additive SNP effect on the qualitative trait. $\gamma_2$ is the additive SNP effect on the quantitative trait.
effects at true SNPs with MAF > 0.0043. The estimates of the additive effects and the correlations for the 11 true SNPs are shown in Table 3. As expected, a strong correlation exists between quantitative trait Q1 and disease status with an estimated correlation of 0.68.

**Discussion and conclusions**
In this paper, we described a likelihood-based approach for the joint association analysis of quantitative and qualitative traits in unrelated individuals. As expected from the true data generation model, our association analysis of the GAW17 data reveals that there is strong correlation between the quantitative trait Q1 and the qualitative trait disease status. In this case, the joint association test of pleiotropic effects is more appropriate than the test based on univariate analysis, which ignores the trait correlation.

Two factors might have contributed to the small difference between the joint analysis and the univariate analysis using the GAW17 data set. First, both methods had low power with rare variants, for example, when MAF < 0.0014. On the other hand, when the MAF was moderate and the effect size was relatively large, both methods yielded power close to or equal to 100%; therefore we are unlikely to see the difference between the two methods under such situations. In fact, it may not be fair to compare the powers between the two methods because the univariate analysis tends to inflate type I error rates in the presence of correlation between the qualitative trait and the quantitative trait.

In its current form, the single-SNP joint analysis is more ideal for common variants and less ideal for rare variants. To improve the power of the association tests with rare variants, we could apply the approaches of Morris and Zeggini [11] based on accumulations of rare variants within the same functional unit. The idea is to treat the proportion of rare variants at which an individual carries a minor allele as a covariate in the joint analysis model and to test for the covariate effects on the qualitative and quantitative traits jointly.

**Acknowledgments**
This research was supported by National Institutes of Health grant R15 CA150698. This article has been published as part of BMC Proceedings Volume 5 Supplement 9, 2011: Genetic Analysis Workshop 17. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/5/S9/suppl.

**Authors’ contributions**
GD was responsible for the study design and helped to draft the manuscript. MY performed the statistical analysis and drafted the manuscript. Both authors read and approved the final manuscript.

**Competing interests**
The authors declare that there are no competing interests.

**Published:** 29 November 2011

**References**
1. Jiang C, Zeng ZB: Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 1995, 140:1111-1127.
2. Mani B, Thoquet P, Grimsley N: Pleiotropic QTL analysis. Biometrics 1998, 54:88-99.
3. Amos CI, Laird AF: A comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol 1993, 10:671-676.
4. Almasy L, Dyer TD, Blangero J: Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet Epidemiol 1997, 14:953-958 [http://onlineibrary.wiley.com/doi/10.1002/(SICI)1098-2272(1997%2914:6%3C953::AID-GEPI6%3E3.0.CO;2-K/pdf].
5. Blangero J, Almasy L, Williams JT, Porjesz B, Reich T, Begleiter H, COGA Collaborators: Incorporating quantitative traits in genomic scans of psychiatric diseases: alcoholism and event-related potentials. Am J Med Genet 1997, 74:602.
6. Wijmjan EM, Amos CI: Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 1997, 14:719 [http://onlineibrary.wiley.com/doi/10.1002/(SICI)1098-2272(1997%2914:6%3C719::AID-GEPI8%3E3.0.CO;2-X/pdf].
7. Williams JT, Eerdekens PV, Almasy L, Blangero J: Joint multipoint linkage analysis of multivariate quantitative and qualitative traits. I. Likelihood formulation and simulation results. Am J Hum Genet 1999, 65:1134-1147.
8. Williams JT, Begleiter H, Porjesz B, Edenberg HJ, Foroud T, Reich T, Goaite A, van Eerdewegh P, Almasy L, Blangero J: Joint multipoint linkage analysis of multivariate quantitative and qualitative traits. II. Alcoholism and event-related potentials. Am J Hum Genet 1999, 65:1148-1160.
9. Liu J, Pei Y, Papasian CJ, Deng H: Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet Epidemiol 2009, 33:217-227.
10. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: Numerical recipes in C: the art of scientific computing, Cambridge, UK, Cambridge University Press, 2nd 1992.
11. Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 2010, 34:188-193.

**Table 3 Estimate of the additive effects and the correlation for the 11 true SNPs on chromosome 13**

| SNP     | Joint analysis | Univariate analysis |
|---------|----------------|---------------------|
|         | γ1  | γ2  | ρ     | γ1  | γ2  |
| C13S520 | 1.721 | 0.296 | 0.687 | 2.311 | 0.296 |
| C13S399 | -1.548 | 0.149 | 0.687 | -2.760 | 0.149 |
| C13S431 | 0.300 | 0.821 | 0.690 | 0.329 | 0.821 |
| C13S479 | -0.459 | 0.598 | 0.687 | -1.230 | 0.598 |
| C13S505 | -2.283 | 0.065 | 0.687 | -3.865 | 0.065 |
| C13S514 | 0.575 | 0.320 | 0.686 | 0.304 | 0.320 |
| C13S522 | 0.616 | 1.134 | 0.684 | 0.617 | 1.134 |
| C13S523 | 0.586 | 1.017 | 0.679 | 0.597 | 1.017 |
| C13S524 | 0.672 | 1.894 | 0.688 | 0.700 | 1.894 |
| C13S547 | -0.926 | 0.290 | 0.687 | -3.332 | 0.290 |
| C13S567 | -0.689 | -0.124 | 0.687 | -0.826 | -0.124 |

γ1 is the additive SNP effect on the qualitative trait, γ2 is the additive SNP effect on the quantitative trait, ρ is the correlation between the quantitative trait and the qualitative trait.