Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
J. Genet. Genomics 36 (2009) 733–743
JOURNAL OF GENETICS AND GENOMICS
www.jgenetgenomics.org

Bivariate association analysis for quantitative traits using generalized estimation equation

Fang Yang [a], Zihui Tang [a], Hongwen Deng [a, b, c, *]

[a] *Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha 410081, China*
[b] *Center of Systematic Biomedical Research, Shanghai University of Science and Technology, Shanghai 200093, China*
[c] *Departments of Orthopedic Surgery and Basic Medical Sciences, University of Missouri-Kansas City, Kansas City, MO 64108, USA*

Received for publication 25 May 2009; revised 9 November 2009; accepted 9 November 2009

Abstract

Quantitative traits often underlie risk for complex diseases. Many studies collect multiple correlated quantitative phenotypes and perform univariate analyses on each of them respectively. However, this strategy may not be powerful and has limitations to detect pleiotropic genes that may underlie correlated quantitative traits. In addition, testing multiple traits individually will exacerbate perplexing problem of multiple testing. In this study, generalized estimating equation 2 (GEE2) is applied to association mapping of two correlated quantitative traits. We suppose that a quantitative trait locus is located in a chromosome region that exerts pleiotropic effects on multiple quantitative traits. In that region, multiple SNPs are genotyped. Genotypes of these SNPs and the two quantitative traits affected by a causal SNP were simulated under various parameter values: residual correlation coefficient between two traits, causal SNP heritability, minor allele frequency of the causal SNP, extent of linkage disequilibrium with the causal SNP, and the test sample size. By power analytical analyses, it is showed that the bivariate method is generally more powerful than the univariate method. This method is robust and yields false-positive rates close to the pre-set nominal significance level. Our real data analyses attested to the usefulness of the method.

Keywords: general estimating equation; bivariate; quantitative trait; linkage disequilibrium

Introduction

Complex traits including quantitative traits and complex diseases are under complicated genetic and environmental determination. For some complex diseases, there are multiple quantitative traits that can be used to diagnose or to quantify the degree of risk to the diseases. Often, there are shared genetic determination for these common complex diseases and quantitative traits. So far, a large number of statistical methods have been developed for identifying genes for single quantitative trait in genetic studies. Researchers usually collected a cluster of related risk quantitative traits in relation to complex disease, and test them independently (Turki et al., 1995; Martinez et al., 1997; Kessler et al., 2005). However, such strategy ignores potential genetic correlations between different traits analyzed and hence is difficult to detect pleiotropic genes that are important to the pathogenesis of multiple correlated human diseases, and moreover analyzing traits separately may aggravate multiple testing problems comparing to testing multiple phenotypes jointly.

If two traits are correlated to each other, it is not appropriate to treat them as independent variables and analyze them separately. For related multiple traits, genetic analysis should be carried out by multivariate methods (Lange and Whittaker, 2001, 2002; Lange et al., 2003). Multivariate linkage analyses had been widely adopted (Deng et al.,

* Corresponding author. Tel & Fax: +86-731-8887 2791.
E-mail address: hwdeng@hunnu.edu.cn

DOI: 10.1016/S1673-8527(08)60166-6

2007; Karasik et al., 2007; Tang et al., 2007) and a consistent view had been proposed that multivariate linkage analysis increases statistical power than univariate linkage analysis (Amos and Laing, 1993; Jiang and Zeng, 1995). Multivariate methods for association analyses are available (Liu et al., 2009; Pei et al., 2009) but rare, especially using the population-based genome wide association (GWA) data. Our group proposed a multivariate analysis method for analyzing the mixture phenotypes of continuous and binary traits (Liu et al., 2009). Pei et al. (2009) proposed a method for association test of multivariate quantitative traits using haplotype trend regression. However, this method is designed for haplotype analysis. Multivariate association mapping of two related quantitative traits for single locus has not widely been applied through some basic methodology.

Generalized estimating equation 1 (GEE1) first proposed by Liang and Zeger (1986) is an extension of generalized linear models (GLMs) to accommodate correlated data with various distribution (Zeger et al., 1988). This method was further improved by Zhao and Prentice (1990), referred to as GEE2. GEE2 outperforms GEE1 because GEE2 can estimate regression and association parameters simultaneously under a unified framework while the main goal of GEE1 is the estimation of regression parameter (Zhao and Prentice, 1990). It has been widely applied to multiple-sourced data as well as the phenotypic data obtained in longitudinal studies.

GWA study has been considered as a powerful tool to identify genes for disease or quantitative traits. Multivariate analyses jointly and simultaneously analyze correlated multiple traits in one statistical testing. Thus, they do not need to correct for the number of phenotypes analyzed and are likely more powerful compared with univariate analyses of individual phenotypes respectively. This is particular true in searching for shared pleiotropic genes. Here, we introduced a bivariate association method to detect shared genes with subtle to moderate effects contributing to two related quantitative phenotypes using GEE2 for population-based design study. We also assessed the performance of GEE2-based bivariate analyses including power and false-positive rate in comparison with univariate analyses through extensive simulation studies. This method was further validated in real GWA data analyses for osteoporosis and obesity.

Materials and methods

We consider a biallelic quantitative trait locus (QTL) Q (causal SNP) harbored in a chromosome region. For the convenience of presentation, we just focus the situation that each individual has two phenotypic observations. In a region of the QTL Q, we suppose that multiple single nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD) with the causal SNP are also genotyped. For simplicity, we considered three SNPs with different extents of LD with the causal SNP and a random SNP generated independent of the causal SNP. We suppose that all the SNPs are in Hardy-Weinberg equilibrium (HWE).

Data simulation

To evaluate the performance of bivariate association analysis based on the introduced GEE method, we conducted extensive simulations for bivariate association analyses of two quantitative traits under different parameter designs.

We assume that the causal SNP (SNP1) is in LD with three genotyped SNPs. LD level between the causal SNP and each SNP can be measured by the following formula:

$$r^2 = (p_{12} - p_1 * p_2)^2 / (p_1 * (1 - p_1) * p_2 * (1 - p_2))$$

where r^2 means a measure of LD (Hedrick and Kumar, 2001); p_1 and p_2 denote the minor allele frequency (MAF) at two loci, respectively. Here, p_1 is the MAF of the causal SNP, p_2 is the MAF for the SNPs which are in LD with the causal SNP. p_{12} denotes the haplotype frequency in joint distribution of both alleles. r^2 was given as the simulation parameter and initial p_2 was randomly generated within a range decided by p_1 and r_2 (Supplemental data). Then we can exactly obtain the values of p_{12} and p_2 *via* the above formula. The related SNP sets of SNP2, SNP3, SNP4 were generated randomly under condition of r^2 and allele frequency p_2.

In brief, given r_2 and MAF for each SNP, we can take three steps to generate 5 SNPs as follows:

1) Specify the MAF of the causal SNP, and then the genotypes of the causal SNP can be generated randomly.

2) Given the LD measure of SNPs 2–4 and the range of their MAFs, respectively, the genotypes of these SNPs can be generated by the simulator.

3) For SNP5, which was not linked to the causal SNP, give a MAF of the SNP, the genotypes of this SNP can be simulated independent of the causal SNP.

For those SNPs in high LD with each other, we reconstructed haplotypes from the simulated SNPs. For simplicity, we reconstructed haplotypes just using SNPs 2–4. The algorithm of the haplotype reconstruction is the same as