NIH Public Access
Author Manuscript
Genet Epidemiol. Author manuscript; available in PMC 2014 December 01.

Published in final edited form as:
Genet Epidemiol. 2013 December ; 37(8): 759–767. doi:10.1002/gepi.21759.

A General Framework for Association Tests With Multivariate Traits in Large-Scale Genomics Studies

Qianchuan He[1], **Christy L. Avery**[2], and **Dan-Yu Lin**[3,*]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

[2]Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America

[3]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, United States of America

Abstract

Genetic association studies often collect data on multiple traits that are correlated. Discovery of genetic variants influencing multiple traits can lead to better understanding of the etiology of complex human diseases. Conventional univariate association tests may miss variants that have weak or moderate effects on individual traits. We propose several multivariate test statistics to complement univariate tests. Our framework covers both studies of unrelated individuals and family studies and allows any type/mixture of traits. We relate the marginal distributions of multivariate traits to genetic variants and covariates through generalized linear models without modeling the dependence among the traits or family members. We construct score-type statistics, which are computationally fast and numerically stable even in the presence of covariates and which can be combined efficiently across studies with different designs and arbitrary patterns of missing data. We compare the power of the test statistics both theoretically and empirically. We provide a strategy to determine genome-wide significance that properly accounts for the linkage disequilibrium (LD) of genetic variants. The application of the new methods to the meta-analysis of five major cardiovascular cohort studies identifies a new locus (*HSCB*) that is pleiotropic for the four traits analyzed.

Keywords

binary traits; genome-wide association studies; meta-analysis; multivariate tests; pleiotropy; quantitative traits; score statistics

Introduction

Pleiotropy, the influence of one gene on multiple traits, is a widespread phenomenon in complex human diseases [Sivakumaran et al., 2011]. Recent years have seen a heightened interest in discovering genetic variants with pleiotropic effects [Gottesman et al., 2012; Lawson et al., 2011; Paaby and Rockman, 2012; Watanabe et al., 2000]. The joint analysis of multiple traits can increase statistical power by aggregating multiple weak effects and

© 2013 Wiley Periodicals, Inc.

*Correspondence to: Dr. Dan-Yu Lin, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420, USA. lin@bios.unc.edu.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

provide new biological insights by revealing pleiotropic variants [Amos and Laing, 1993; Jiang and Zeng, 1995].

The advent of large-scale genetic association studies, particularly genome-wide association studies (GWAS), poses tremendous challenges in analyzing multiple traits. First, there are a huge number of genetic variants to be tested, which may entail considerable computation burden. The inclusion of covariates (e.g., ancestry variables to account for population stratification) may make the computation even more intensive. Second, complex diseases are characterized by a wide variety of traits, some of which are continuous (i.e., quantitative) and some of which are discrete. Third, it is desirable to combine results from multiple studies, some of which may consist of unrelated individuals and some of which may consist of families; the genetic variants and the traits may not be uniformly measured in all studies. Fourth, it is necessary to adjust for multiple testing, but the conventional Bonferroni correction may be overly conservative.

There exist several statistical methods for association analysis of multiple traits, but none of them addresses all the above issues. Ferreira and Purcell [2009] suggested canonical correlation analysis, which is computationally fast but does not accommodate covariates. Liu et al. [2009] suggested a Wald statistic based on generalized estimating equations (GEE) [Liang and Zeger, 1986] for the mixture of one continuous trait and one binary trait. Their method does not accommodate family data, and the Wald statistic requires fitting a regression model for each genetic variant, which can be time consuming. Yang et al. [2010] suggested a linear combination of univariate test statistics with data-dependent weights by estimating the weights from part of the data and calculating the test statistic from the remaining data. The *P*-values are assessed by permutation, which is computationally demanding. Maity et al. [2012] proposed a kernel machine method for joint analysis of multiple genetic variants, which is equivalent to testing the variance component in a multivariate linear mixed model. Recently, van der Sluis et al. [2013] proposed a method called “trait-based association test that uses extended Simes procedure” (TATES). The Simes procedure was originally designed to alleviate the conservativeness of the Bonferroni correction; the TATES extends the Simes procedure to the multivariate-trait analysis by harnessing the correlations among the traits.

In this paper, we provide a very general framework for association analysis of multiple traits, which simultaneously tackles all the aforementioned challenges. Our framework covers both studies of unrelated individuals and family studies and allows any type/mixture of traits. To enhance robustness, we relate the marginal distributions of multivariate traits to genetic variants and covariates through generalized linear models without parametric modeling of the dependence among the traits or family members; we account for the dependence in constructing the test statistics by estimating the correlations empirically from the data. We develop score-type statistics, which are computationally fast and numerically stable even in the presence of covariates and which can be combined efficiently across studies with different designs and arbitrary patterns of missing data. We consider various types of multivariate test statistics and compare their power both theoretically and empirically. We provide a strategy to determine genome-wide significance that properly accounts for the linkage disequilibrium (LD) of genetic variants. We demonstrate the usefulness of the new methods through extensive simulation studies and an application to five GWAS studies involving cardiovascular traits.

Methods

In this section, we present our general framework for association tests with multivariate traits. We first construct the marginal models and the corresponding score-type statistics.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We then show how to combine those statistics to form multivariate test statistics. Finally, we discuss meta-analysis and genome-wide significance thresholds.

Calculating Score Statistics and Their Covariance Matrix

We consider a single study with a total of n unrelated individuals, K (potentially correlated) traits, and p covariates (including the unit component). For $i = 1, \ldots, n$ and $k = 1, \ldots, K$, let Y_{ki} be the kth trait of the ith individual. For $i = 1, \ldots, n$, let X_i be the p-vector of covariates for the ith individual, and let G_i denote the number of minor alleles (or the imputed dosage) the ith individual carries at a particular test locus.

We assume that the marginal distribution of Y_{ki} is related to X_i and G_i through a generalized linear model with mean $\mu_k(\beta_k^{\mathrm{T}} X_i + \gamma_k G_i)$ and dispersion parameter ϕ_k, where $\mu_k(\cdot)$ is a specific function, and β_k and γ_k are unknown regression parameters. We adopt natural link functions such that $\mu_k(x) = x$ for continuous traits and $\mu_k(x) = e^x/(1 + e^x)$ for binary traits.

To accommodate missing data, we let ξ_{ki} indicate, by the values 1 versus 0, whether Y_{ki} is observed or missing, and let ψ_i indicate, by the values 1 versus 0, whether G_i is observed or missing. It is assumed that the covariates have no missing values. (We recommend to exclude the covariates with substantial missingness and to replace the missing values with their sample means for the remaining covariates.)

The score function for (β_k, γ_k) takes the form

$$S_k(\beta_k, \gamma_k) = \varphi_k^{-1} \sum_{i=1}^{n} \xi_{ki} \{Y_{ki} - \mu_k(\beta_k^{\mathrm{T}} X_i + \gamma_k G_i)\} \begin{bmatrix} X_i \\ \psi_i G_i \end{bmatrix}.$$

Thus, the score statistic for testing the null hypothesis that $\gamma_k = 0$ is

$$U_k = \widehat{\varphi}_k^{-1} \sum_{i=1}^{n} \xi_{ki} \psi_i \{Y_{ki} - \mu_k(\widehat{\beta}_k^{\mathrm{T}} X_i)\} G_i,$$

where $\widehat{\beta_k}$ solves the equation

$$\sum_{i=1}^{n} \xi_{ki} \{Y_{ki} - \mu_k(\beta_k^{\mathrm{T}} X_i)\} X_i = 0,$$

and $\widehat{\phi_k}$ is a sample estimator of ϕ_k. For continuous traits,

$$\widehat{\varphi}_k = \frac{\sum_{i=1}^{n} \xi_{ki} (Y_{ki} - \widehat{\beta}_k^{\mathrm{T}} X_i)^2}{\sum_{i=1}^{n} \xi_{ki} - p};$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

for binary traits, $\hat{\phi_k} = 1$. Note that the construction of the U_ks makes full use of the available data by estimating β_k and ϕ_k from all individuals with nonmissing trait values and is more efficient than the traditional complete-case analysis.

By taking the Taylor series expansion of $S_k(\beta_k, 0)$ at $\beta_k = \hat{\beta_k}$ and applying the law of large numbers, we can show that U_k is asymptotically equivalent to the following sum of n-independent terms:

$$\varphi_k^{-1}\sum_{i=1}^{n}\xi_{ki}\{Y_{ki}-\mu_k(\beta_k^{\mathrm{T}}X_i)\}(\psi_i G_i - A_k^{\mathrm{T}}B_k^{-1}X_i),$$

where A_k and B_k are the limits of $n^{-1}\sum_{i=1}^{n}\psi_i\mu_k'(\beta_k^{\mathrm{T}}X_i)G_iX_i$ and $n^{-1}\sum_{i=1}^{n}\mu_k'(\beta_k^{\mathrm{T}}X_i)X_iX_i^{\mathrm{T}}$, respectively, and $\mu_k'(x)=d\mu_k(x)/dx$. Define the score vector

$$U=\begin{bmatrix} U_1 \\ \vdots \\ U_K \end{bmatrix}.$$

It follows from the multivariate central limit theorem that U is asymptotically K-variate normal with mean 0 and with a covariance matrix that can be estimated by

$$V \equiv \begin{bmatrix} V_{11} & \cdots & V_{1K} \\ \vdots & \vdots & \vdots \\ V_{K1} & \cdots & V_{KK} \end{bmatrix},$$

where

$$V_{kl}=\sum_{i=1}^{n}U_{ki}U_{li}, \quad k,l=1,\ldots,K,$$
$$U_{ki}=\hat{\varphi}_k^{-1}\xi_{ki}\{Y_{ki}-\mu_k(\hat{\beta}_k^{\mathrm{T}}X_i)\}(\psi_i G_i-\hat{A}_k^{\mathrm{T}}\hat{B}_k^{-1}X_i), \quad k=1,\ldots,K;i=1,\ldots,n,$$
$$\hat{A}_k=\sum_{i=1}^{n}\psi_i\mu_k'(\hat{\beta}_k^{\mathrm{T}}X_i)G_iX_i,$$

and

$$\hat{B}_k=\sum_{i=1}^{n}u_k'(\hat{\beta}_k^{\mathrm{T}}X_i)X_iX_i^{\mathrm{T}}.$$

Note that $\hat{\beta_k}$ and $\hat{B_k}$ ($k = 1, \ldots, K$) do not depend on the SNP genotypes and thus need to be calculated only once (before looping through all the SNPs). Note also that, given the $\hat{\beta_k}$s, the calculations of U and V do not involve solving any equations. Thus, the implementation of the proposed score-type statistics is orders of magnitude faster than that of the conventional Wald statistics. In addition, the score-type statistics are numerically more stable and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

statistically more accurate than the Wald statistics, especially when the minor allele frequency (MAF) is low [Lin and Tang, 2011].

We now extend the above results to family studies. Suppose that we have a total of n families, with n_i members in the ith family. For $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, K$, let Y_{kij} denote the kth trait for the jth member of the ith family, X_{ij} denote the p-vector of covariates for the jth member of the ith family, and G_{ij} denote the number of minor alleles (or the imputed dosage) which the jth member of the ith family carries at a particular test locus. We assume that the marginal distribution of Y_{kij} is related to X_{ij} and G_{ij} through the same marginal generalized linear regression model as in the case of unrelated individuals.

Let ξ_{kij} indicate whether Y_{kij} is observed or missing, and let ψ_{ij} indicate whether G_{ij} is observed or missing. It is assumed that there are no missing values in the covariates. Under the independence working assumption [Liang and Zeger, 1986], the (pseudo-likelihood) score statistic for testing the null hypothesis that $\gamma_k = 0$ is

$$U_k = \widehat{\varphi}_k^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \xi_{kij} \psi_{ij} \{Y_{kij} - \mu_k(\widehat{\beta}_k^{\mathrm{T}} X_{ij})\} G_{ij},$$

where $\widehat{\beta_k}$ solves the equation

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \xi_{kij} \{Y_{kij} - \mu_k(\beta_k^{\mathrm{T}} X_{ij})\} X_{ij} = 0,$$
$$\widehat{\varphi}_k = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \xi_{kij} (Y_{kij} - \widehat{\beta}_k^{\mathrm{T}} X_{ij})^2 / (\sum_{i=1}^{n} \sum_{j=1}^{n_i} \xi_{kij} - p)$$

for continuous traits, and $\widehat{\phi_k} = 1$ for binary traits. Again, define $U = [U_1, \ldots, U_K]^{\mathrm{T}}$. It follows from the above arguments for the case of unrelated individuals that U is asymptotically K-variate normal with mean 0 and a covariance matrix that can be estimated by $V \equiv \{V_{kl}; k, l = 1, \ldots K\}$, where

$$V_{kl} = \sum_{i=1}^{n} U_{ki} U_{li}, \quad k, l = 1 \ldots, K,$$
$$U_{ki} - \widehat{\varphi}_k^{-1} \sum_{j=1}^{n_i} \xi_{kij} \{Y_{kij} - \mu_k(\widehat{\beta}_k^{\mathrm{T}} X_{ij})\} (\psi_{ij} G_{ij} - \widehat{A}_k^{\mathrm{T}} \widehat{B}_k^{-1} X_{ij}), \quad k = 1, \ldots, K; i = 1, \ldots, n,$$
$$\widehat{A}_k = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \psi_{ij} \mu_k'(\widehat{\beta}_k^{\mathrm{T}} X_{ij}) G_{ij} X_{ij},$$
$$\widehat{B}_k = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \mu_k'(\widehat{\beta}_k^{\mathrm{T}} X_{ij}) X_{ij} X_{ij}^{\mathrm{T}}.$$

Note that the relatedness of family members is accounted for through the empirical correlations of the U_{ki}s.

Performing Multivariate Association Tests

To test the global null hypothesis that $\gamma_1 = \gamma_2 = \cdots = \gamma_K = 0$, we calculate the quadratic form

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$Q=U^{\mathrm{T}}V^{-1}U,$$

which is asymptotically chi-squared with K degrees of freedom. This is a global test statistic that is consistent (i.e., having the power of 1 as the sample size tends to ∞) against any alternative hypotheses.

To enhance power against alternative hypotheses under which genetic effects are similar among the K studies, we calculate a test statistic with one degree of freedom along the lines of O'Brien [1984]. Specifically, let Z be the standardized version of U and let R be the correlation matrix of U. That is, $Z_k=U_k/V_{kk}^{1/2}$ (k = 1, …, K) and $R_{kl} = V_{kl}/(V_{kk}\, V_{ll})^{1/2}$ (k, l = 1, …, K). We then calculate

$$T=\frac{e^{\mathrm{T}}R^{-1}Z}{(e^{\mathrm{T}}R^{-1}e)^{1/2}},$$

where $e = [1, \ldots, 1]^{\mathrm{T}}$. This test statistic is asymptotically standard normal.

The test statistic T maximizes the noncentrality parameter among all linear combinations of the Z_ks. Note that the score test statistic Z_k is asymptotically equivalent to the Wald test statistic, i.e., the estimate of γ_k divided by its standard error. Thus, T is optimal if the limits of the Z_ks or the standardized γ_ks are the same. To detect alternative hypotheses under which the original γ_ks are the same, we define $Z_k^{'}+U_k/V_{kk}(k=1,\ldots,K)$ and $C_{kl} = V_{kl}/(V_{kk}\, V_{ll})$($k$, l = 1, …, K). Note that the limit of U_k/V_{kk} is approximately γ_k. We then calculate

$$T^{'}=\frac{e^{\mathrm{T}}C^{-1}Z^{'}}{(e^{\mathrm{T}}C^{-1}e)^{1/2}},$$

where $Z^{'}=[Z_1^{'},\ldots,Z_{\kappa}^{'}]^{\mathrm{T}}$ and $C = \{C_{kl}; k, l = 1, \ldots, K\}$. This test statistic is also asymptotically standard normal. When using this test statistic, it is important to use comparable scales for the K traits such that it is plausible for the γ_k to be equal. When using either T or T', it is important to code the trait values in such a way that the genetic effects on the K traits are plausibly in the same direction.

If the effects of the SNP are similar among the K traits, then T and T' will tend to be more powerful than Q. If the effects are very different, then Q will likely be more powerful than T and T'. In the Appendix, we derive the asymptotic distributions of Q, T, and T' under alternative hypotheses for the important special case of two continuous traits.

Combining Results From Multiple Studies

We wish to combine results from L-independent studies. For l = 1, …, L, let $U^{(l)}$ and $V^{(l)}$ denote the score vector and its (estimated) covariance matrix from the lth study. Then the overall score vector is $U=\sum_{l=1}^{L}U^{(l)}$, and its covariance matrix is estimated by $V=\sum_{l=1}^{L}V^{(l)}$. Note that $\sum_{l=1}^{L}U^{(l)}$ is the (pseudo-likelihood) score statistic in the joint analysis of the individual-level data of the L studies (allowing nuisance parameters to be

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

different among the studies). Thus, meta-analysis of score statistics is equivalent to the joint analysis of individual-level data. When there are multiple studies, K pertains to the total number of distinct traits, some of which may not be measured in certain studies. (For $L = 2$, we may have four traits that are common between the two studies, two traits that are measured only in the first study, and three traits that are measured only in the second study. Then $K = 9$.) Given U and V, we can calculate Q, T, and T' in the same manner as in the case of a single study.

Determining Genome-Wide Significance

Suppose that we have a total of m SNPs. For $j = 1, \ldots, m$, let Q_j be the value of Q for the jth SNP. If the critical value q_0 for the m test statistics satisfies

$$\Pr(\max_{j=1,\ldots,m} Q_j \geq q_0) = \alpha,$$

then the family-wise error rate will be α. We estimate q_0 by Monte Carlo simulation. At each test locus, we calculate

$$\tilde{U} = \begin{bmatrix} \sum_{i=1}^{n} U_{1i} W_i \\ \vdots \\ \sum_{i=1}^{n} U_{Ki} W_i \end{bmatrix},$$

where $W_1, \ldots, W_n$ are independent standard normal random variables. Let $\tilde{U}_j$ and V_j be the values of $\tilde{U}$ and V for the jth SNP. Define

$$\tilde{Q}_j = \tilde{U}_j^{\mathrm{T}} V_j^{-1} \tilde{U}_j, \quad j = 1, \ldots, m.$$

The joint distribution of $(Q_1, \ldots, Q_m)$ can be approximated by that of $(\tilde{Q_1}, \ldots, \tilde{Q_m})$ [Lin, 2005]. Thus, we determine q_0 by the following equation

$$\Pr\left(\max_{j=1,\ldots,m} \tilde{Q}_j \geq q_0\right) = \alpha.$$

We simulate the normal random sample $(W_1, \ldots, W_n)$ 10,000 times while holding the observed data fixed and set q_0 to be the $10{,}000(1 - \alpha)$th largest value of the resulting $\max_{j=1, \ldots, m} \tilde{Q_j}$'s. We may convert the critical value q_0 to the P-value threshold p_0 by referring q_0 to the chi-squared distribution with K degrees of freedom. We can determine the genome-wide significance thresholds for T, T', and $Z_k (k = 1, \ldots, K)$ in a similar manner.

Results

Simulation Studies

We conducted simulation studies to evaluate the performance of the proposed test statistics. We set G to be the number of minor alleles for a SNP with MAF of 0.4 and set X to be

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

normal with mean $0.1G$ and unit variance. We generated two continuous traits under the bivariate linear model: $Y_1 = 1 + 0.5X + \gamma_1 G + \varepsilon_1$ and $Y_2 = 1 + X + \gamma_2 G + \varepsilon_2$, where ε_1 and ε_2 are zero-mean normal with variances $\sigma_1^2=2$ and $\sigma_2^2=1$, respectively, and with correlation $\rho = 0.5$. We set α to 10^{-4}. To evaluate the type I error, we simulated 10 million data sets under $\gamma_1 = \gamma_2 = 0$. To evaluate the power, we simulated 10,000 data sets under various combinations of γ_1 and γ_2. Each simulated data set consists of 1,000 unrelated individuals. In addition to the three multivariate test statistics, Q, T, and T', we considered two versions of univariate tests, Uni-B and Uni-corr, which adjust for multiple testing (between the traits) by adopting the Bonferroni correction (i.e., dividing the nominal significance level by the number of traits) and by accounting for the correlation between Z_1 and Z_2 (using the multivariate normal distribution of U), respectively. We also included the TATES method (van der Sluis et al., 2013).

The results are summarized in Table 1. The type I error for the TATES is inflated by about 12%. The type I errors for the other five tests are below the nominal significance level. The Q test has reasonable power against all 12 alternatives. As expected, T' is more powerful than the other tests when γ_1 is close to γ_2, and T is more powerful than the others when γ_1/σ_1 and γ_2/σ_2 are close to each other. The Uni-B is expected to have lower power than Uni-corr, but the two tests perform very similarly due to the relatively weak correlation between the two traits. The differences between the two tests become more pronounced as the correlation increases; see supplementary Table SI. The TATES has slightly higher power than Uni-B and Uni-corr but also has inflated type I error, especially when the correlation is high.

We also considered the mixture of a binary trait and a continuous trait. We simulated the binary trait under the logistic regression model $\text{logit}\{P(Y_1 = 1)\} = -1 + 0.5X + \gamma_1 G$ and simulated the continuous trait under the linear model $Y_2 = 1 + X + \gamma_2 G + \varepsilon$, where ε is normal with mean $2Y_1$ and unit variance. (The Pearson correlation between the two traits is about 0.65.) As shown in Table 2, Q tends to have higher power than the two univariate tests. As expected, T' is more powerful than the other tests when $\gamma_1 = \gamma_2$, and T outperforms the others when the means of Z_1 and Z_2 are similar. Again, the TATES has higher power than Uni-B and Uni-corr but at the expense of inflated type I error.

We also considered four continuous traits with a compound-symmetry correlation structure for the error terms. The results are shown in supplementary Table SII. The basic conclusions remain the same. We added the MANOVA method implemented in R to the case of no covariates. As shown in supplementary Table SIII, MANOVA has slightly higher type I error and power than the Q test. This is consistent with the general phenomenon that the likelihood ratio test is more liberal than the score test [Lin and Zeng, 2011]. Finally, we considered family studies with 250 families (two parents and two children in each family) and two continuous traits. As shown in supplementary Table SIV, the conclusions are similar to the case of unrelated subjects.

Cardiovascular Studies

We analyzed the GWAS data on the Caucasian samples from the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, the Cardiovascular Health Study (CHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Framingham Heart Study (FHS), the sample sizes being 9,068, 1,433, 3,892, 2,286, and 2,789, respectively. The FHS is a family study, and the others consist of unrelated individuals. Each individual was genotyped on 250,000 SNPs. We considered four cardiovascular traits: diabetes status, high-density lipoprotein (HDL), low-density lipoprotein (LDL), and triglycerides; the first trait is binary whereas the other three are continuous. These traits are major players in the development of coronary artery

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

diseases and metabolic syndrome [Grundy, 2012; Holmes et al., 1981]. We aimed to identify genetic factors underlying these traits. Since the HDL is "good" cholesterol, we used negative values of HDL in the analysis.

We performed single-SNP analysis with the following covariates: age, gender, study centers, and the top 10 principle components for ancestry. We calculated the score-type statistics and their covariance matrices for each study and then combined the results of the five studies. The (unadjusted) *P*-values of the univariate and multivariate tests are displayed in Figures 1 and 2, respectively. The genome-wide significance thresholds based on the Bonferroni correction and the Monte Carlo procedure are marked in both figures.

We first examine the results based on the Bonferroni correction. For the four traits, more than 10 regions are above the Bonferroni threshold in the Uni-B test (Fig. 1). Compared to Uni-B, the *Q* test identifies one new signal that is located on chromosome (Chr) 9 (Fig. 2). The signal on Chr9 identified by the *Q* test is an accumulation of weak/moderate signals for individual traits. The gene at this locus encodes a protein called *ABCA1*, which is involved in cellular cholesterol removal (Lawn et al., 1999). This gene was previously found to be associated with metabolic syndrome (Avery et al., 2011). Table 3 lists all the loci discovered by the *Q* test. (The *P*-values from the univariate-trait analysis are also shown.) The *T* and T' tests did not identify any additional signals that achieve genome-wide significance, but the two tests yielded more extreme *P*-values for several SNPs than the univariate tests.

Not surprisingly, the Monte Carlo procedure reduced the genome-wide significance thresholds for all tests. For the univariate test on the HDL, one SNP on Chr20 becomes significant by the Monte Carlo criterion. For the *T* test, one SNP (rs5752792) on Chr22 is above the Monte Carlo threshold. This SNP resides near gene *HSCB*, which is mainly expressed in liver, muscle and heart [Sun et al., 2003] and is involved in the biogenesis of an elementary metabolic function unit [Rouault and Tong, 2008]. The expression pattern and biological function of *HSCB* strongly suggest that this gene is pleiotropic.

We have provided a very general and flexible approach to association testing with multivariate traits. An earlier version of this approach (focusing on the *Q* test for continuous traits) was recently used to successfully identify genes associated with metabolic syndrome [Avery et al., 2011]. The new application presented in this paper further demonstrates the usefulness of the proposed approach. It only took several hours to calculate all the *P*-values shown in Figures 1 and 2. We have posted our software online at http://dlin.web.unc.edu/software.

When the number of traits is very large, we recommend to reduce the dimension through principal component analysis [Avery et al., 2011]. Although we have focused on main effects of genetic variants, our approach can be easily modified to test gene–environment interactions. It can also be extended to perform burden tests on rare variants (Lin and Tang, 2011).

Univariate-trait analysis and multivariate-trait analysis are complementary to each other. The former is easier to implement and can be used to rapidly screen a large number of genetic variants. The multivariate-trait analysis provides a useful tool to uncover pleiotropic variants that have weak or moderate effects on individual traits. This is particularly important for dissecting the genetic basis of complex diseases, as most of the genetic variants with strong effects and high MAFs might have already been identified.

There is no uniformly most powerful test for analyzing multivariate traits. If the effects of a genetic variant are similar across the traits, then *T* and T' are generally preferable. If the effects are considerably different or even in opposite directions, then *Q* is preferable. The

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

theoretical results of the Appendix offer useful insights into the relative power of the three test statistics and can be used to determine the power and sample size for future studies.

For family data, we adopted the marginal models with an independence working correlation matrix. A more efficient approach would be a random-effect model which utilizes the family relationships. We adopted marginal models instead of random-effects models for several reasons. First, the association tests under marginal models are more robust to model misspecification. Second, it is much faster to fit marginal models than random-effects models. Third, marginal models can easily handle mixtures of continuous and binary traits.

Adjustment for multiple testing is an important issue in genetic association analysis. The Monte Carlo procedure considered in this paper accounts for the correlations among the test statistics and is thus less conservative than the conventional Bonferroni correction. Some existing methods, such as the TATES, may yield inflated type I error. We have focused on determining the genome-wide significance threshold rather than calculating individual adjusted *P*-values. The former only requires several thousands Monte Carlo samples whereas the latter would entail millions of Monte Carlo samples to estimate extremely small *P*-values. If the number of SNPs is small, the joint distribution of the test statistics can be evaluated through numerical integration [Conneely and Boehnke, 2007, 2010].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the National Institutes of Health grants R01 CA082659, U01 HG004803, and R00-HL-098458. We thank two reviewers for their helpful comments.

References

Amos CI, Laing AE. A comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol. 1993; 10:671–676. [PubMed: 8314079]

Avery CL, He Q, North KE, Ambite JL, Boerwinkle E, Fornage M, Hindorff LA, Kooperberg C, Meigs JB, Pankow JS, et al. A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. PLoS Genet. 2011; 7:e1002322. [PubMed: 22022282]

Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of p-values for multiple correlated tests. Am J Hum Genet. 2007; 81:1158–1168. [PubMed: 17966093]

Conneely KN, Boehnke M. Meta-analysis of genetic association studies and adjustment for multiple testing of correlated SNPs and traits. Genet Epidemiol. 2010; 34:739–746. [PubMed: 20878715]

Ferreira M, Purcell S. A multivariate test of association. Bioinformatics. 2009; 25:132–133. [PubMed: 19019849]

Gottesman O, Drill E, Lotay V, Bottinger E, Peter I. Can genetic pleiotropy replicate common clinical constellations for cardiovascular disease and risk? PLoS ONE. 2012; 7:e46419. [PubMed: 23029515]

Grundy SM. Pre-diabetes, metabolic syndrome, and cardiovascular risk. J Am Coll Cardiol. 2012; 59:635–643. [PubMed: 22322078]

Holmes DR, Elveback LR, Frye RL, Kottke BA, Ellefson RD. Association of risk factor variables and coronary artery disease documented with Angiography. Circulation. 1981; 63:293–299. [PubMed: 7449053]

Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics. 1995; 140:1111–1127. [PubMed: 7672582]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Lawn RM, Wade DP, Garvin MR, Wang X, Schwartz K, Porter JG, Seilhamer JJ, Vaughan AM, Oram JF. The Tangier disease gene product ABC1 controls the cellular apolipoprotein-mediated lipid removal pathway. J Clin Invest. 1999; 104:R25–R31. [PubMed: 10525055]

Lawson HA, Cady JE, Partridge C, Wolf JB, Semenkovich CF, Cheverud JM. Genetic effects at pleiotropic loci are context-dependent with consequences for the maintenance of genetic variation in populations. PLoS Genet. 2011; 7:e1002256. [PubMed: 21931559]

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics. 2005; 21:781–787. [PubMed: 15454414]

Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011; 89:354–367. [PubMed: 21885029]

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. 2009; 33:217–227. [PubMed: 18924135]

Maity A, Sullivan PF, Tzeng JY. Multivariate phenotype association analysis by marker-set kernel machine regression. Genet Epidemiol. 2012; 36:686–695. [PubMed: 22899176]

O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984; 40:1079–1087. [PubMed: 6534410]

Paaby AB, Rockman MV. The many faces of pleiotropy. Trends Genet. 2012; 29:66–73. [PubMed: 23140989]

Rouault TA, Tong WH. Iron-sulfur cluster biogenesis and human disease. Trends Genet. 2008; 24:398–407. [PubMed: 18606475]

Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H. Abundant pleiotropy in human complex disease and traits. Am J Hum Genet. 2011; 89:607–618. [PubMed: 22077970]

Sun G, Gargus JJ, Ta DT, Vickery LE. Identification of a novel candidate gene in the iron-sulfur pathway implicated in ataxia-susceptibility: human gene encoding HscB, a J-type co-chaperone. J Hum Genet. 2003; 48:415–419. [PubMed: 12938016]

van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013; 9:e1003235. [PubMed: 23359524]

Watanabe RM, Ghosh S, Langefeld CD, Valle TT, Hauser ER, Magnuson VL, Mohlke KL, Silander K, Ally DS, Chines P, et al. The Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics (FUSION) Study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. Am J Hum Genet. 2000; 67:1186–1200. [PubMed: 11032784]

Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet Epidemiol. 2010; 34:444–454. [PubMed: 20583287]

APPENDIX

Asymptotic Distributions of *Q*, *T*, *T'* for Two Quantitative Traits

We consider a study of unrelated individuals and two quantitative traits satisfying the bivariate linear model:

$$Y_{ki}=\beta_k^{\mathrm{T}} X_i+\gamma_k G_i+\varepsilon_{ki},\ \ k=1,2; i=1,\ldots,n.$$

where ε_{1i} and ε_{2i} are bivariate zero-mean normal with covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In the absence of missing values, the score statistic for testing $\gamma_k = 0$ takes the form

$$U_k = \frac{1}{\hat{\sigma}_k^2} \sum_{i=1}^{n} (Y_{ki} - \hat{\beta}_k^{\mathrm{T}} X_i) G_i,$$

where $\hat{\beta}_k$ and $\hat{\sigma}_k^2$ are the least-squares estimators of β_k and σ_k^2. Simple algebraic manipulation yields

$$U_k = \frac{1}{\hat{\sigma}_k^2} \sum_{i=1}^{n} \left\{ G_i - \left(\sum_{j=1}^{n} G_j X_j^{\mathrm{T}} \right) \left(\sum_{j=1}^{n} X_j X_j^{\mathrm{T}} \right)^{-1} X_i \right\} \times (\varepsilon_{ki} + \gamma_k G_i).$$

Assume that γ_k is in the order of $n^{-1/2}$. By the multivariate central limit theorem and the law of large numbers, $(U_1, U_2)^{\mathrm{T}}$ is approximately bivariate normal with mean

$$\omega \begin{bmatrix} \frac{\gamma_1}{\sigma_1^2} \\ \frac{\gamma_2}{\sigma_2^2} \end{bmatrix}$$

and covariance matrix

$$\omega \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{\rho}{\sigma_1 \sigma_2} \\ \frac{\rho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix},$$

where $\omega = \sum_{j=1}^{n} G_j^2 - \left(\sum_{j=1}^{n} G_j X_j^{\mathrm{T}} \right) \left(\sum_{j=1}^{n} X_j X_j^{\mathrm{T}} \right)^{-1} \left(\sum_{j=1}^{n} G_j X_j \right)$.

It follows from the above result that Q is approximately chi-squared with 2 degrees of freedom and with noncentrality parameter

$$\lambda = \omega \frac{\gamma_1^2 \sigma_2^2 - 2\rho\sigma_1\sigma_2\gamma_1\gamma_2 + \gamma_2^2 \sigma_1^2}{(1-\rho^2)\sigma_1^2\sigma_2^2}.$$

In addition, T is approximately normal with mean

$$\mu_T = \sqrt{\frac{w}{2(1+\rho)}} \left(\frac{\gamma_1}{\sigma_1} + \frac{\gamma_2}{\sigma_2} \right)$$

and unit variance, and T' is approximately normal with mean

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mu_{T'}=\sqrt{\frac{\omega}{(1-\rho^2)(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2)}}\times\left\{\gamma_1\left(\frac{\sigma_2}{\sigma_1}-\rho\right)+\gamma_2\left(\frac{\sigma_1}{\sigma_2}-\rho\right)\right\}$$

and unit variance.

In the special case of $\gamma_1/\sigma_1 = \gamma_2/\sigma_2 = s$,

$$\begin{aligned}\lambda&=\omega\frac{2s^2}{1+\rho},\\ \mu_T^2&=\omega\frac{2s^2}{1+\rho},\\ \mu_{T'}^2&=\omega\frac{(\gamma_1+\gamma_2)^2(1-\rho)}{(1-\rho)(\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2)}.\end{aligned}$$

Clearly, $\lambda=\mu_T^2$. It can be shown that $\mu_T^2 \geq \mu_{T'}^2$, where the equality holds if and only if $\sigma_1 = \sigma_2$ (assuming that $|\rho| < 1$).

In the special case of $\gamma_1 = \gamma_2 = \gamma$,

$$\begin{aligned}\lambda&=\omega\frac{\gamma^2(\sigma_1^2-2\rho\sigma_1\sigma_2+\sigma_2^2)}{(1-\rho^2)\sigma_1^2\sigma_2^2},\\ \mu_T^2&=\omega\frac{\gamma^2(\sigma_1+\sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho)},\\ \mu_{T'}^2&=\omega\frac{\gamma^2(\sigma_1^2-2\rho\sigma_1\sigma_2+\sigma_2^2)}{(1-\rho^2)\sigma_1^2\sigma_2^2}\end{aligned}$$

Note that $\lambda=\mu_{T'}^2$. It can be shown that $\mu_{T'}^2 \geq \mu_T^2$, where the equality holds if and only if $\sigma_1 = \sigma_2$ (assuming that $|\rho| < 1$).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

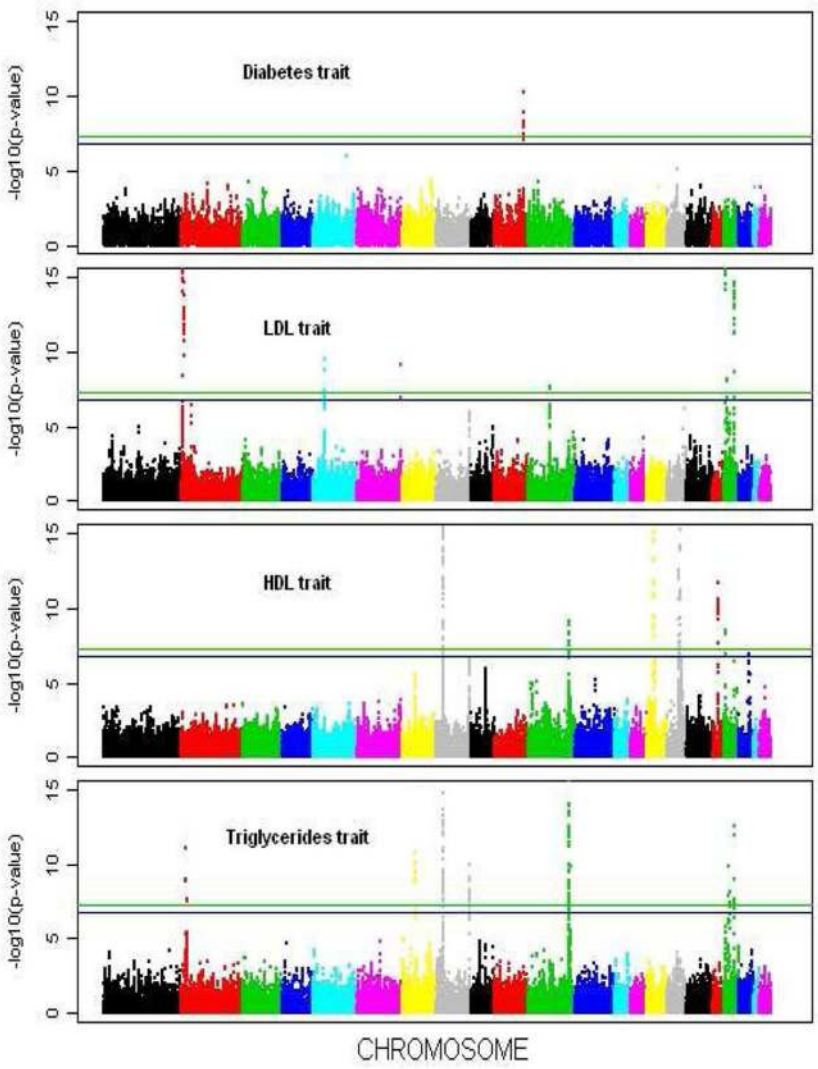

Figure 1.
Univariate tests of the diabetes status and the LDL, HDL, and triglyceride levels in the ARIC, CARDIA, CHS, MESA, and FHS GWAS studies. Genome-wide significance thresholds based on the Bonferroni correction and the Monte Carlo procedure are shown in green and blue, respectively.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

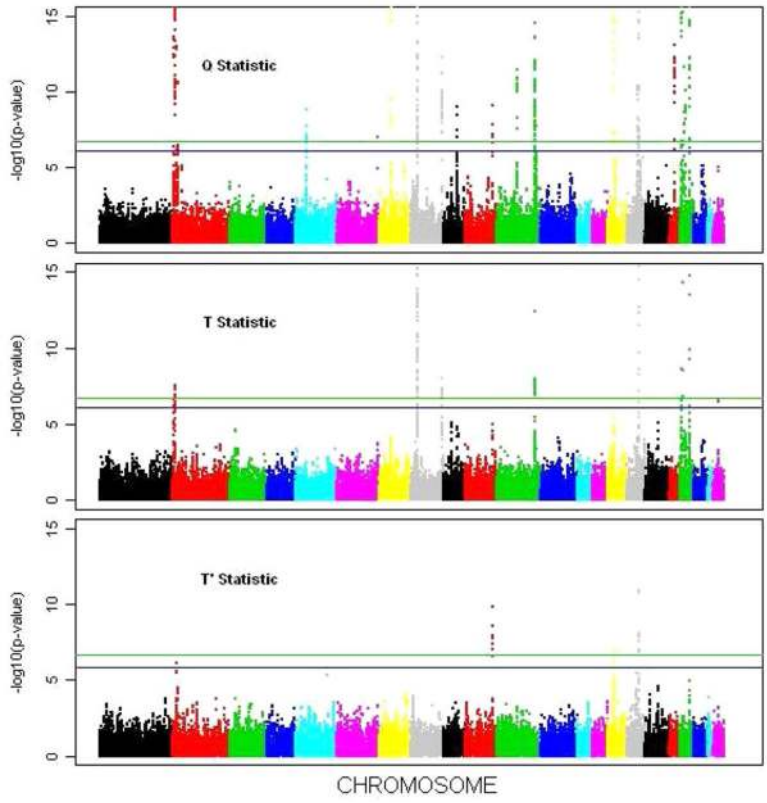

Figure 2.
Multivariate tests of the diabetes status and the LDL, HDL, and triglyceride levels in the ARIC, CARDIA, CHS, MESA, and FHS GWAS studies. Genome-wide significance thresholds based on the Bonferroni correction and the Monte Carlo procedure are shown in green and blue, respectively.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

Type I error and power of univariate versus multivariate tests for two continuous traits (ρ= 0.5)

(γ_1, γ_2)	$(\gamma_1/\sigma_1, \gamma_2/\sigma_2)$	Uni-B	Uni-corr	TATES	*Q*	*T*	*T′*
(0, 0)	(0, 0)	8.6×10^{-5}	8.7×10^{-5}	1.12×10^{-4}	9.0×10^{-5}	9.0×10^{-5}	8.8×10^{-5}
(0.3, 0)	(0.21, 0)	0.6861	0.6865	0.714	0.854	0.095	0.003
(0.3, 0.1)	(0.21, 0.1)	0.6865	0.6869	0.715	0.638	0.487	0.187
(0.25, 0.18)	(0.18, 0.18)	0.5714	0.5723	0.606	0.594	0.700	0.641
(0.3, 0.25)	(0.21, 0.25)	0.9358	0.936	0.945	0.942	0.967	0.966
(0.2, 0.2)	(0.14, 0.2)	0.6222	0.6229	0.651	0.594	0.632	0.702
(0.2, 0.25)	(0.14, 0.25)	0.9054	0.9056	0.917	0.881	0.828	0.922
(0.25, 0.25)	(0.18, 0.25)	0.9128	0.913	0.925	0.906	0.917	0.947
(0, 0.25)	(0, 0.25)	0.9034	0.9036	0.915	0.977	0.197	0.701
(0, 0.3)	(0, 0.3)	0.9902	0.9903	0.992	0.999	0.402	0.914
(0.1, 0.25)	(0.07, 0.25)	0.9034	0.9036	0.915	0.907	0.523	0.841
(0.1, 0.3)	(0.07, 0.3)	0.9902	0.9903	0.992	0.994	0.742	0.968
(0.2, 0.3)	(0.14, 0.3)	0.9903	0.9904	0.992	0.987	0.940	0.990

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Type I error and power of univariate versus multivariate tests for one binary and one continuous traits

(γ_1, γ_2)	$(\bar{Z_1}, \bar{Z_2})^a$	Uni-B	Uni-corr	TATES	Q	T	T'
(0, 0)	—	8.7×10^{-5}	9.0×10^{-5}	1.02×10^{-4}	8.7×10^{-5}	9.7×10^{-5}	9.3×10^{-5}
(0.3, 0)	(3.07, 2.09)	0.171	0.173	0.171	0.141	0.141	0.026
(0.3, 0.1)	(3.07, 3.65)	0.377	0.380	0.398	0.334	0.410	0.382
(0.25, 0.18)	(2.54, 4.52)	0.685	0.687	0.709	0.664	0.488	0.758
(0.3, 0.25)	(3.07, 5.88)	0.973	0.973	0.977	0.975	0.845	0.986
(0.2, 0.2)	(2.02, 4.49)	0.676	0.678	0.700	0.709	0.372	0.764
(0.2, 0.25)	(2.02, 5.24)	0.892	0.893	0.907	0.938	0.534	0.937
(0.25, 0.25)	(2.54, 5.56)	0.943	0.944	0.952	0.958	0.707	0.967
(0, 0.25)	(0.02, 4.02)	0.489	0.493	0.519	0.880	0.046	0.655
(0, 0.3)	(0.02, 4.79)	0.782	0.784	0.808	0.987	0.104	0.888
(0.1, 0.25)	(1.01, 4.62)	0.722	0.725	0.750	0.899	0.210	0.832
(0.1, 0.3)	(1.01, 5.37)	0.917	0.919	0.930	0.990	0.347	0.960
(0.2, 0.3)	(2.02, 5.97)	0.979	0.979	0.984	0.994	0.688	0.992

[a] $\bar{Z_1}$ and $\bar{Z_2}$ are the sample means of Z_1 and Z_2, respectively.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

P-values of the genetic loci identified by the *Q* test

				P **-values of single-trait analysis**			
Chr	**SNP**	**Gene**	***P*** **-value of Q test**	**Diabetes**	**LDL**	**HDL**	**Trig**
2	rs515135	*APOB*	6.3×10^{-17}	4.0×10^{-1}	3.4×10^{-19}	5.8×10^{-1}	3.1×10^{-1}
2	rs1260326	*GCKR*	2.2×10^{-16}	3.7×10^{-4}	3.6×10^{-1}	6.1×10^{-1}	8.9×10^{-12}
5	rs12916	*HMGCR*	1.7×10^{-9}	8.2×10^{-1}	3.0×10^{-10}	2.0×10^{-1}	9.4×10^{-1}
6	rs10455872	*LPA*	1.1×10^{-7}	9.3×10^{-1}	8.7×10^{-10}	1.8×10^{-1}	8.7×10^{-1}
7	rs7777102	*MLXIPL*	2.2×10^{-19}	4.6×10^{-1}	8.4×10^{-1}	3.8×10^{-4}	1.2×10^{-20}
8	rs1011685	*LPL*	5.3×10^{-37}	5.2×10^{-1}	7.6×10^{-1}	2.3×10^{-17}	4.6×10^{-35}
8	rs2954021	*TRIB1*	5.9×10^{-13}	6.4×10^{-1}	1.1×10^{-6}	2.0×10^{-5}	1.1×10^{-10}
9	rs2575876	*ABCA1*	9.5×10^{-10}	2.4×10^{-1}	1.9×10^{-3}	2.0×10^{-6}	4.8×10^{-1}
10	rs7903146	*TCF7L2*	8.4×10^{-10}	5.8×10^{-11}	8.0×10^{-2}	8.3×10^{-1}	6.4×10^{-1}
11	rs174538	*FEN1*	3.5×10^{-12}	1.9×10^{-1}	2.9×10^{-8}	3.2×10^{-3}	6.6×10^{-4}
11	rs964184	*ZNF259*	2.7×10^{-29}	8.6×10^{-1}	2.9×10^{-1}	3.6×10^{-17}	3.3×10^{-27}
15	rs1077835	*LIPC*	1.1×10^{-29}	2.2×10^{-1}	2.9×10^{-1}	5.2×10^{-21}	3.1×10^{-3}
16	rs247616	*CETP*	6.7×10^{-99}	7.3×10^{-2}	5.5×10^{-3}	3.5×10^{-96}	2.5×10^{-4}
18	rs4121823	*LIPG*	8.0×10^{-14}	5.3×10^{-2}	7.3×10^{-1}	2.4×10^{-12}	8.2×10^{-1}
19	rs6511720	*LDLR*	5.8×10^{-37}	8.6×10^{-1}	3.3×10^{-39}	6.6×10^{-4}	6.0×10^{-1}
19	rs10401969	*SUGP1*	2.4×10^{-12}	5.1×10^{-1}	1.1×10^{-5}	3.5×10^{-1}	1.6×10^{-10}
19	rs445925	*APOC1*	2.1×10^{-74}	8.6×10^{-1}	2.2×10^{-60}	6.3×10^{-4}	1.1×10^{-9}