Published in final edited form as:
Ann Appl Stat. 2014 March 1; 8(1): 352–376. doi:10.1214/13-AOAS690.

JOINT ANALYSIS OF SNP AND GENE EXPRESSION DATA IN GENETIC ASSOCIATION STUDIES OF COMPLEX DISEASES

Yen-Tsung Huang[*], **Tyler J. VanderWeele**[†], and **Xihong Lin**[†]

[*]Brown University

[†]Harvard University

Abstract

Genetic association studies have been a popular approach for assessing the association between common Single Nucleotide Polymorphisms (SNPs) and complex diseases. However, other genomic data involved in the mechanism from SNPs to disease, e.g., gene expressions, are usually neglected in these association studies. In this paper, we propose to exploit gene expression information to more powerfully test the association between SNPs and diseases by jointly modeling the relations among SNPs, gene expressions and diseases. We propose a variance component test for the total effect of SNPs and a gene expression on disease risk. We cast the test within the causal mediation analysis framework with the gene expression as a potential mediator. For eQTL SNPs, the use of gene expression information can enhance power to test for the total effect of a SNP-set, which are the combined direct and indirect effects of the SNPs mediated through the gene expression, on disease risk. We show that the test statistic under the null hypothesis follows a mixture of χ^2 distributions, which can be evaluated analytically or empirically using the resampling-based perturbation method. We construct tests for each of three disease models that is determined by SNPs only, SNPs and gene expression, or includes also their interactions. As the true disease model is unknown in practice, we further propose an omnibus test to accommodate different underlying disease models. We evaluate the finite sample performance of the proposed methods using simulation studies, and show that our proposed test performs well and the omnibus test can almost reach the optimal power where the disease model is known and correctly specified. We apply our method to re-analyze the overall effect of the SNP-set and expression of the *ORMDL3* gene on the risk of asthma.

Keywords and phrases

Causal Inference; Data Integration; Mediation Analysis; Mixed Models; Score Test; SNP Set Analysis; Variance Component Test

Address of Yen-Tsung Huang, Department of Epidemiology, Brown University, 121 South Main Street, Providence, RI 02912, USA, Yen-Tsung Huang@brown.edu
Address of Tyler J. Vander Weele, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA, tvanderw@hsph.harvard.edu
Address of Xihong Lin, Department of Biostatistics, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA, xlin@hsph.harvard.edu

SUPPLEMENTARY MATERIAL

Supplement:

(). Section 1: Detailed development of causal mediation model and derivations referenced in Sections 4.1 and 4.2; Section 2: derivation of model (4.7) in Sections 4.4; Section 3: asymptotic distribution of *Q* referenced in Section 3.1; Table and Figures referenced in Sections 5.2 and 5.4.

1. Introduction

Genome-wide association studies (GWAS) constitute a popular approach for investigating the association of common Single Nucleotide Polymorphisms (SNPs) with complex diseases. Usually, a large number of SNP markers are tested across the genome. Great interest lies in improving power of testing SNP effects by borrowing additional biological information. Indeed, a major criticism of genetic association studies lies in its agnostic style (Hunter and Chanock, 2010): none of biological knowledge was encoded in the standard genetic association analyses. To overcome such limitations, multi-marker analysis has been advocated to integrate biological information into statistical analyses and to decrease the number of tests (Kwee et al., 2008; Wu et al., 2010). Analysis using SNP-sets grouped by physical locations has better performance than the standard single SNP analysis in re-analyzing the breast cancer GWAS data (Wu et al., 2010). SNPs can also be grouped into a SNP-set according to biological pathways, in which a gene harmonizes with other genes to exert biological functions.

The two factors we try to bridge in genetic association studies are SNPs and disease risk. Despite the success of SNP-set analyses in assembling multiple SNPs based on biological information, mechanistic pathways between SNPs (SNP-sets) and disease are still neglected. Given the availability of multiple sources in genomic data (e.g., gene expression and SNPs) (Moffatt et al., 2007; Cusanovich et al., 2012), it is desirable to perform joint analysis by integrating multiple sources of genomic data. Here we combine the information of SNPs and gene expression by introducing gene expression as a mediator in the causal pathway from SNPs to disease. *Biologically*, gene expression can be determined by the DNA genotype (Morley et al., 2004; Cheung et al., 2005; Fu et al., 2009) and that gene expression can also affect disease risk (Dermitzakis, 2008). Moreover, results from the SNP-set analysis augmented by a biological model can be more scientifically meaningful. *Statistically*, gene expression can help explain variability of the effect of SNPs on disease when there exists an effect of SNPs on disease via gene expression and thus increases the power of detecting the overall effect of SNPs on disease risk.

SNPs that regulate mRNA expression of a gene are so-called expression Quantitative Trait Loci (eQTL) (Schadt et al., 2003). Statistically, eQTL SNPs can be viewed as the SNPs that are correlated with mRNA expression of a gene. *Cis*-eQTL SNPs are the SNPs that are within or around the corresponding gene, and *trans*-eQTL SNPs are those that are far away or even on different chromosomes. Numerous genome-wide eQTL analyses have been reported to comprehensively capture such a DNA-RNA (i.e., SNPs-gene expression) association in the genome in different tissues and organisms (Schadt et al., 2003; Morley et al., 2004; Innocenti et al., 2011). eQTL results can be external information to prioritize the discovery of susceptibility loci in genome-wide association studies (Hsu et al., 2010; Zhong et al., 2010; Zhang et al., 2012). Methods are available to integrate multiple genomic data to draw causal inference on a biological network (Schadt et al., 2005; Zhu et al., 2008; Hageman et al., 2011; Neto et al., 2013). We focus in this paper on *joint analysis* of multiple eQTL SNPs of a gene and their corresponding mRNA expression for their effects on disease phenotypes. Compared with multi-SNP analyses, this approach further incorporates eQTLs into genetic association studies, and accounts for a biological process (from DNA to RNA) within a gene to improve power.

This paper is motivated by an asthma genome-wide association study of subjects of British descent (MRC-A), in which the association between SNPs at *ORMDL3* gene and the risk of childhood asthma was investigated (Dixon et al., 2007; Moffatt et al., 2007). The MRC-A dataset consists of 108 cases and 50 controls with both SNP genotype (Illumina 300K) and gene expression (Affymetrix HU133A 2.0) data available. The original genome-wide study

reported that the 10 typed SNPs on chromosome 17q21 where *ORMDL3* is located, were strongly associated with childhood asthma in MRC-A data, and the results were validated in several other independent studies. The authors also found that each of these 10 SNPs was highly correlated with gene expression of *ORMDL3*, which is also associated with asthma. The 10 SNPs, *ORMDL3* expression and asthma status can be illustrated as the **S**, *G* and *Y*, respectively in Figure 1. Instead of analyzing SNP-expression, expression-asthma, and SNP-asthma associations separately and univariately, here we are interested in assessing the overall genetic effect of *ORMDL3* on the occurrence of childhood asthma, by jointly analyzing SNP and gene expression data and accounting for the possibility that the *ORMDL3* gene expression might be a causal mediator for the association of the SNPs in the *ORMDL3* gene and asthma risk. Our ultimate goal is to integrate multiple sources of genomic data for genetic association analyses.

In this paper, we propose to jointly model a set of SNPs within a gene, a gene expression, and disease status, where a logistic model is used to model the dependence of disease status on the SNP-set and the gene expression, and a linear model is used for the dependence of the gene expression on the SNP-set, both adjusting for covariates. We are primarily interested in testing whether a gene, whose effects are captured by SNPs and/or gene expression, is associated with a disease phenotype. We formulate this hypothesis in the causal mediation analysis framework (Robins and Greenland, 1992; Van-derWeele and Vansteelandt, 2009, 2010; Imai et al., 2010). Note that the previous work on causal mediation analysis is mostly focused on estimation.

We use the joint model to derive the direct and indirect effects of a SNP-set mediated through gene expression on disease risk. For eQTL SNPs, we show that the total effect of a gene on a disease captured by a set of SNPs and a gene expression corresponds to the total effects of the SNP-set, which are the combined direct effects and indirect effects of the SNPs mediated through the gene expression, on a disease. This framework allows us to study how the use of gene expression data can enhance power to test for the total effect of a SNP-set on disease risk. We study the impact of model mis-specification using the conventional SNP-only models when the true model is that both the SNPs and the gene expression affect the disease outcome. For non-eQTL SNPs, the null hypothesis simply corresponds to the joint effects of the SNPs and the gene expression.

Due to potentially a large number of SNPs within a gene and some of them might be highly correlated, i.e., in high linkage disequilibrium (LD), conventional tests, such as the likelihood ratio test, for the total effects of multiple SNPs and a gene expression, do not perform well. We propose in this paper a variance component test to assess the overall effects of a SNP set and a gene expression on disease risk. Under the null that the test statistic follows a mixture of χ^2 distributions, which can be approximated analytically or empirically using a resampling based perturbation procedure (Parzen et al., 1994; Cai et al., 2012). As the true disease model is often unknown, we construct an omnibus test to improve the power by accommodating different underlying disease models.

The rest of the paper is organized as follows. In Section 2, we introduce the joint model for SNPs, a gene expression and disease as well as the null hypothesis of no joint effect of the SNPs and the gene expression on a disease phenotype. In Section 3, we propose a variance component score test for the total effect of SNPs and gene expression, and construct an omnibus test to maximize the test power across different underlying disease models. In Section 4, we interpret the null hypothesis and study the assumptions within the framework of causal mediation modeling for eQTL SNPs and non-eQTL SNPs. In Section 5, we evaluate the finite sample performance of the proposed test using simulation studies and show that the omnibus test is robust and performs well in different situations. In Section 6,

we apply the proposed method to study the overall effect of the *ORMDL3* gene contributed by both the SNPs and the gene expression on the risk of childhood asthma, followed by discussions in Section 7.

2. The Model and the Null Hypothesis

The statistical problem is to jointly model the effect of a set of SNPs and a gene expression on the occurrence of a disease. Assume for subject i ($i = 1, \ldots, n$), an outcome of interest Y_i is dichotomous (e.g., case/control), whose mean is associated with q covariates ($\mathbf{X}_i$, with the first covariate to be 1, i.e., the intercept), p SNPs in a SNP-set ($\mathbf{S}_i$), mRNA expression of a gene (G_i) and possibly the interactions between the SNPs and the gene expression as:

$$\text{logit}\{P(Y_i=1|\mathbf{S}_i, G_i, \mathbf{X}_i)\}=\mathbf{X}_i^T\boldsymbol{\alpha}+\mathbf{S}_i^T\boldsymbol{\beta}_S+G_i\beta_G+G_i\mathbf{S}_i^T\boldsymbol{\gamma}, \quad (2.1)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$, $\boldsymbol{\beta}_S = (\beta_{S_1}, \ldots, \beta_{S_p})^T$, β_G, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ are the regression coefficients for the covariates, the SNPs, the gene expression, and the interactions of the SNPs and the gene expression, respectively. A SNP-set and gene expression pair can be defined in multiple ways. For example, **S** can be the SNPs in a gene and G is the mRNA expression of the gene. In the asthma data example, **S** are the 10 typed SNPs around *ORMDL3* and G is the expression of *ORMDL3*. Alternatively, one can choose the SNP-set/expression pair based on the eQTL study: eQTL SNP-set/corresponding gene expression.

As SNPs can affect gene expression (Schadt et al., 2003; Morley et al., 2004; Innocenti et al., 2011), for each subject i, we next consider a linear model for the continuous gene expression G_i (i.e., the mediator), which depends on the q covariates ($\mathbf{X}_i$) and the p SNPs ($\mathbf{S}_i$):

$$G_i=\mathbf{X}_i^T\boldsymbol{\phi}+\mathbf{S}_i^T\boldsymbol{\delta}+\varepsilon_i, \quad (2.2)$$

where $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_q)^T$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^T$ are the regression coefficients for the covariates and the SNPs, respectively; and ε_i follows a normal distribution with mean 0 and variance σ_G^2. Again, the p SNPs can be the SNPs within a gene or the eQTL SNPs corresponding to a gene for which the expression level is measured.

Our goal is to test for the total effect of a gene captured by the SNPs in a set **S** and a gene expression G on Y, which can be written using the regression coefficients in model (2.1) as:

$$H_0:\boldsymbol{\beta}_S=\mathbf{0}, \beta_G=0, \boldsymbol{\gamma}=\mathbf{0}. \quad (2.3)$$

Note that the null hypothesis (2.3) only involves the parameters in the [Y |**S**, G, **X**] model (2.1). We use [G|**S**, **X**] model in Section 4 to facilitate interpretation of the null hypothesis (2.3) and study the assumptions within the causal mediation analysis framework. Throughout the paper, we term this null hypothesis as a test for the *total effect of a gene*. Later in Section 4 we will show that it corresponds to the *total effect of SNPs* for eQTL SNPs and simply to the joint effect of SNPs and expression for non-eQTL SNPs.

3. Test for the Total Effects of a Gene

3.1. Test Statistic for the Total Effect of a Gene

We propose in this section to test for the null hypothesis of no total effect of a gene (2.3) under model (2.1). As the number of SNPs (p) in a gene might be large and some might be highly correlated (due to linkage disequilibrium), the likelihood ratio test (LRT) or

multivariate Wald test for the null hypothesis (2.3) uses a large degrees of freedom (DF) and has limited power. To overcome this problem, we assume under model (2.1), the regression coefficients of the individual main SNP effects β_{Sj} are independent and follow an arbitrary distribution with mean 0 and variance τ_S, and the SNP-by-expression interaction coefficients γ_j ($j = 1, \ldots, p$) are independent and follow another arbitrary distribution with mean 0 and variance τ_I. The resulting outcome model (2.1) hence becomes a logistic mixed model. The problem for testing the null hypothesis (2.3) becomes a joint test of variance components ($\tau_S = \tau_I = 0$) and the scalar regression coefficient for the fixed gene expression effect ($\beta_G = 0$) in the induced logistic mixed models as $H_0 : \tau_S = \tau_I = 0$ and $\beta_G = 0$. One can easily show that the scores for τ_S, β_G, and τ_I under the induced logistic mixed models are:

$$U_{\tau_S}=\{\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0\}^T\mathbb{S}\mathbb{S}^T\{\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0\}, U_{\beta_G}=\mathbf{G}^T\{\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0\}, U_{\tau_I}=\{\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0\}^T\mathbb{C}\mathbb{C}^T\{\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0\},$$

where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, $\mathbb{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^T$, $\mathbb{S} = (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n)^T$, $\mathbf{G} = (G_1, G_2, \ldots, G_n)^T$, and $\mathbb{C} = (\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_n)^T = (G_1\mathbf{S}_1, G_2\mathbf{S}_2, \ldots, G_n\mathbf{S}_n)^T$; $\hat{\boldsymbol{\mu}}_0 = (\hat{\mu}_{01}, \cdots, \hat{\mu}_{0n})^T$ and $\hat{\mu}_{0i}=\exp(\mathbf{X}_i^T\hat{\boldsymbol{\alpha}}_0)/\{1+\exp(\mathbf{X}_i^T\hat{\boldsymbol{\alpha}}_0)\}$ is the mean of Y_i under H_0, and $\hat{\boldsymbol{\alpha}}_0$ is the maximum likelihood estimator of $\boldsymbol{\alpha}$ under the null model

$$\text{logit}\{P(Y_i=1|\mathbf{S}_i, G_i, \mathbf{X}_i)\}=\mathbf{X}_i^T\boldsymbol{\alpha}. \quad (3.1)$$

To combine the three scores to test for the null hypothesis $H_0 : \tau_S = \tau_I = 0$ and $\beta_G = 0$, one may consider the conventional score statistic $Q_{conv} = \mathbf{U}^T\mathcal{I}^{-1}\mathbf{U}$, where $\mathbf{U} = (U_{\tau S}, U_{\beta G}, U_{\tau I})^T$ and $\mathcal{I}$ is the efficient information matrix of $\mathbf{U}$. However, this approach has several major limitations. First, notice that the score of the regression coefficient of gene expression $U_{\beta G}$ is a linear function of Y, while the scores of the variance components of main effects of SNPs and SNP-by-expression interactions τ_S and τ_I are quadratic functions of Y. Hence $U_{\beta G}$ has a different scale from ($U_{\tau S}$, $U_{\tau I}$). It follows that ($U_{\tau S}$, $U_{\tau I}$) are likely to dominate $U_{\beta G}$. A combination of the three scores using Q_{conv} is hence not desirable. Second, Q_{conv} involves quartic functions of $\mathbf{Y}$ and the information matrix $\mathcal{I}$ involves the 8th moment of $\mathbf{Y}$. Hence calculations of Q_{conv} are not stable, and it is difficult to analytically approximate the null distribution of Q_{conv}.

We hence propose the following weighted sum of three scores as the test statistic for the null hypothesis (2.3):

$$\begin{aligned} Q&=n^{-1}(a_1U_{\tau_S}+a_2U^2_{\beta_G}+a_3U_{\tau_I})\\ &=n^{-1}(\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0)^T(a_1\mathbb{S}\mathbb{S}^T+a_2\mathbf{G}\mathbf{G}^T+a_3\mathbb{C}\mathbb{C}^T)(\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0), \end{aligned} \quad (3.2)$$

where a_1, a_2, a_3 are some weights. Q is a nice quadratic function of $\mathbf{Y}$. Hence its null distribution can be easily approximated by a mixture of χ^2 distributions. Different weights can be chosen. With an equal weight $a_1 = a_2 = a_3$, Q is equivalent to the variance component test for τ_{common} by assuming β_{Sj}, β_G and γ_j follow a common distribution with mean zero and variance τ_{common}. However, this common distribution assumption is strong, as $\mathbf{S}$, G, and $\mathbf{C}$ generally have different scales and so do their effects β_{Sj}, β_G and γ_j.

Notice that $U_{\tau S}$, $U^2_{\beta_G}$, $U_{\tau I}$ are all quadratic functions of $\mathbf{Y}$ in similar forms, we propose to weight each term $U_{\tau S}$, $U^2_{\beta_G}$, $U_{\tau I}$ using the inverse of the square root of their corresponding variances. This allows each weighted term to have variance 1 and be comparable.

Specifically, the variances for $U_{\tau S}$, $U^2_{\beta_G}$, U_{η} are $I_{\tau S} = \mathbf{1}^T(\mathbb{SS}^T \cdot \mathbb{K} \cdot \mathbb{SS}^T)\mathbf{1}$, $I_G = \mathbf{1}^T(\mathbf{GG}^T \cdot \mathbb{K} \cdot \mathbf{GG}^T)\mathbf{1}$, $I_{\eta} = \mathbf{1}^T(\mathbb{CC}^T \cdot \mathbb{K} \cdot \mathbb{CC}^T)\mathbf{1}$, respectively, where $\mathbb{A}\ \mathbb{B}$ denotes the component-wise multiplication of conformable matrices $\mathbb{A}$ and $\mathbb{B}$, $\mathbf{1}$ denotes a vector of ones, the diagonal and off-diagonal elements of $\mathbb{K}$ are $k_{ii} = -4\hat{\mu}^4_{0i} + 8\hat{\mu}^3_{0i} - 5\mu^2_{0i} + \hat{\mu}_{0i}$ and $k_{ii'} = 2[\mu_{0i}\hat{(}1 - \mu_{0i}\hat{)}][\mu_{0i'}\hat{(}1 - \mu_{0i'}\hat{)}]$, respectively (Lin, 1997).

We derive the asymptotic distribution of Q under the null hypothesis (2.3) by accounting $Q = Q(\hat{\boldsymbol{\alpha}})$ is a function of $\hat{\boldsymbol{\alpha}}$, which is the maximum likelihood estimate of $\boldsymbol{\alpha}$ under the null model (3.1). Define

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{XX} & \mathbf{D}_{XV} \\ \mathbf{D}_{VX} & \mathbf{D}_{VV} \end{bmatrix} = n^{-1}\mathbf{U}^T\mathbf{W}\mathbf{U},$$

where $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n)^T$, $\mathbf{U}_i = (\mathbf{X}_i^T, \mathbf{V}_i^T)$, $\mathbf{V}_i^T = (\sqrt{a_1}\mathbf{S}_i^T, \sqrt{a_2}G_i, \sqrt{a_3}\mathbf{C}_i^T)$, $\mathbf{W} = \text{diag}\{\mu_i(1 - \mu_i)\}$. We show in Section 3 of the Supplementary Material that under the null hypothesis (2.3), the score test statistic Q converges in distribution to $Q(0) = \sum_{l=1}^{2p+1}(\mathbf{A}_l^T\boldsymbol{\varepsilon})^2$, where $\boldsymbol{\varepsilon}$ is a random vector following $N(\mathbf{0}, \mathbf{D})$ and $\mathbf{A}_l$ is the l^{th} row of $\mathbf{A} = [-\mathbf{D}_{XV}^T\mathbf{D}_{XX}^{-1}, \mathbf{I}_{(2p+1)\times(2p+1)}]$. This means under the null hypothesis, Q follows a mixture of χ^2 distributions, which can be approximated using a scaled χ^2 distribution by matching the first two moments (Satterthwaite, 1946) as $Q \sim \kappa\chi^2_{\nu}$, where $\kappa = \text{Var}(Q)/[2\text{E}(Q)]$ and $\nu = 2[\text{E}(Q)]^2/\text{Var}(Q)$, and the expressions of E(Q) and Var(Q) are given in Section 3 of the Supplementary Material. Alternatively, one can approximate the mixture of χ^2 distribution using the characteristic function inversion method (Davies, 1980).

3.2. The Omnibus Test for the Total Effect of a Gene

So far we derive the test statistic Q under the outcome model specified in (2.1), which assumes the disease risk of Y depends on SNPs, gene expression and their interactions. Denote this Q by Q_{SGC}. Suppose that the disease risk of Y depends on SNPs and gene expression but not their interactions, or even depends only on SNPs, then it is more powerful to test for the total SNP effect using the test statistics $Q_{SG} = n^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}_0})^T(a_1\mathbb{SS}^T + a_2\mathbf{GG}^T)(\mathbf{Y} - \boldsymbol{\mu}_0)$, and $Q_S = n^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}_0})^T(a_1\mathbb{SS}^T)(\mathbf{Y} - \hat{\boldsymbol{\mu}_0})$, respectively. Under these simpler disease models, the test statistic Q_{SGC} loses power as it tests for extra unnecessary parameters. On the other hand, if the disease risk indeed depends on SNPs, expression and their interactions, performing tests only using SNPs Q_S or main effects Q_{SG} will lose power, compared to Q_{SGC}

Since in reality we do not know the underlying true disease model, it is difficult to choose a correct model. It is hence desirable to develop a test that can accommodate different disease models to maximize the power. Moreover, in a genome-wide association study, it is almost impossible that one disease model is true for tens of thousands of genes. Thus we further propose an omnibus test where we identify the strongest evidence among the three different models with: 1) only SNPs, 2) SNPs and gene expression, and 3) SNPs, gene expression and their interactions. Specifically, we calculate the p-value under each of the three models, then compute the minimum of the three p-values and compare the observed minimum p-value to its null distribution. Because of the complicated correlation among Q_{SGC}, Q_{SG} and Q_S, it is difficult to analytically derive the null distribution of the minimum p-value. To this end, we resort to a resampling perturbation procedure.

As shown in Section 3.1, Q converges in distribution to $Q(0)=\sum_{l=1}^{2p+1}(\mathbf{A}_l^T\boldsymbol{\varepsilon})^2$. The empirical distribution of $Q(0)$ can be estimated using the resampling method via perturbation (Parzen et al., 1994; Cai et al., 2012). The perturbation method approximates the target distribution of Q by generating random variables $\hat{\boldsymbol{\varepsilon}}$ from the estimated asymptotic distribution of $\boldsymbol{\varepsilon}$. This perturbation procedure is also called the score-based wild bootstrap (Kline and Santos, 2012).

Specifically, set $\hat{\boldsymbol{\varepsilon}}=n^{-1/2}\sum_{i=1}^{n}\mathbf{U}_i^T(Y_i-\hat{\mu}_i)\mathcal{N}_i$, where $\mathcal{N}_i$'s are independent $N(0, 1)$. By generating independent $\mathcal{N}=(\mathcal{N}_1, \ldots, \mathcal{N}_n)$ repeatedly, the distribution of $\hat{\boldsymbol{\varepsilon}}$ conditional on the observed data is asymptotically the same as that of $\boldsymbol{\varepsilon}$. Denoted by $\{\hat{Q}(0)^{(b)}, b = 1, \ldots, B\}$, where B is the number of perturbations. It follows that the empirical distribution of the $\hat{Q}(0)^{(b)}$ is the same as that of $Q(0)$ asymptotically. The p-value can be approximated using the tail probability by comparing $\{\hat{Q}(0)^{(b)}, b = 1, \ldots, B\}$ with the observed Q. Hence one can calculate the p-values of Q_{SGC}, Q_{SG}, and Q_S by setting $\mathbf{V}_i=(\sqrt{a_1}\mathbf{S}_i^T, \sqrt{a_2}G_i, \sqrt{a_3}\mathbf{C}_i^T)^T, (\sqrt{a_1}\mathbf{S}_i^T, \sqrt{a_2}G_i)$ and $\sqrt{a_1}\mathbf{S}_i^T$ and generate their perturbed realizations of the null counterpart as $\{\hat{Q}_{SGC}(0)^{(b)}\}$, $\{\hat{Q}_{SG}(0)^{(b)}\}$, and $\{\hat{Q}_S(0)^{(b)}\}$, and compare them with corresponding observed values, respectively. Note that for each perturbation b, the random normal perturbation variable $\mathcal{N}^{(b)}$ is the same across the three tests such that the correlation among Q_{SGC}, Q_{SG} and Q_S can be preserved.

The p-value of the omnibus test can be easily calculated using the perturbation method. Let $\hat{P}_S = \mathcal{S}_S(Q_S)$, $\hat{P}_{SG} = \mathcal{S}_{SG}(Q_{SG})$, and $\hat{P}_{SGC} = \mathcal{S}_{SGC}(Q_{SGC})$ be the three p-values using the three statistics, where $\mathcal{S}_S(q) = \text{pr}\{\hat{Q}_S(0)^{(b)} > q\}$, $\mathcal{S}_{SG}(q) = \text{pr}\{\hat{Q}_{SG}(0)^{(b)} > q\}$, and $\mathcal{S}_{SGC}(q) = \text{pr}\{\hat{Q}_{SGC}(0)^{(b)} > q\}$. The null distribution of the minimum p-value, $\hat{P}_{\min} = \min(\hat{P}_S, \hat{P}_{SG}, \hat{P}_{SGC})$ can be approximated by $\hat{P}_{\min}^{(b)}=\min\{\mathcal{S}_S(\hat{Q}_S(0)^{(b)}), \mathcal{S}_{SG}(\hat{Q}_{SG}(0)^{(b)}), \mathcal{S}_{SGC}(\hat{Q}_{SGC}(0)^{(b)})\}(b=1, \cdots, B)$. The p-value of the omnibus test hence can be calculated by comparing the observed minimum p-value $\hat{P}_{\min}$ with its empirical null distribution $\{\hat{P}_{\min}^{(b)}\}$.

Note that different from permutation where the observed data are shuffled and resampled to calculate the test statistic Q, the perturbation procedure resamples from the asymptotic null distribution of Q without re-calculating Q using the shuffled data. Thus, it is much more efficient than the permutation method. Using a single CPU (2.53 GHz) to run 100 genes (each with 10 SNPs) and 100 cases/100 controls, the computation time is 134.10 and 809.76 seconds for the perturbation and permutation methods (both with 200 resampling), respectively. Furthermore, covariates can be easily adjusted using the perturbation method, but covariate adjustment is more difficult using permutation. Specifically, the permutation based p-values calculated by simply permuting SNPs and gene expression fail if SNPs/gene expression are correlated with covariates.

4. Understanding the Total Effect of a Gene and the Assumptions of the Test Using the Causal Mediation Analysis Framework

4.1. Characterization of SNPs, Gene Expression and Disease Risk in the Framework of Causal Mediation Modeling

To understand the null hypothesis of no total effect of a gene captured by SNPs in a gene and a gene expression and the underlying assumptions, we discuss in this section how to interpret the null in (2.3) within the causal mediation analysis framework. Causal interpretation can be helpful for understanding genetic etiology of diseases as well as for applications in pharmaceutical research (Li et al., 2010). Although genotype is essentially

fixed at conception, it is at that time effectively randomized, conditional on parental genotypes and could be viewed as subject to have hypothetical intervention. The statistical problem of jointly modeling a set of SNPs, a gene expression and a disease can be presented as a causal diagram (Pearl, 2001; Robins, 2003) in Figure 1 and be framed using a causal mediation model (VanderWeele and Vansteelandt, 2009 and 2010; Imai et al., 2010) based on counterfactuals (Rubin, 1974 and 1978). VanderWeele and Vansteelandt (2010) and Imai et al. (2010) have used the causal mediation analysis for epidemiological and social science studies, respectively, where the exposure of interest is univariate.

One can decompose the *total effect* (TE) of SNPs into the *Direct Effect* (DE) and the *Indirect Effect* (IE). The *Direct Effect* of SNPs is the effect of the SNPs on the disease outcome that is not through gene expression, whereas the *Indirect Effect* of the SNPs is the effect of the SNPs on the disease outcome that is through the gene expression. Within the causal mediation analysis framework, we derive in the Supplementary Material the TE, DE and IE of the SNPs on the disease outcome.

Specifically, we define the total effect (TE) of SNPs as

$$TE=\text{logit}\{P(Y=1|\mathbf{S}=\mathbf{s}_1,\mathbf{X})\}-\text{logit}\{P(Y=1|\mathbf{S}=\mathbf{s}_0,\mathbf{X})\},$$

i.e., the equation (2.1) marginalizes over gene expression *G*. In Section 1 of the Supplementary Material, we show that for rare diseases, the total effect of the SNPs on the log odds ratio (OR) of disease risk can be expressed in terms of the regression coefficients in models (2.1) and (2.2) and is approximately equal to

$$TE=(\mathbf{s}_1-\mathbf{s}_0)^T\{\boldsymbol{\beta}_\mathbf{s}+\beta_G\boldsymbol{\delta}+\boldsymbol{\gamma}(\mathbf{x}^T\boldsymbol{\phi}+\mathbf{s}_0^T\boldsymbol{\delta}+\beta_G\sigma_G^2)+\boldsymbol{\delta}\mathbf{s}_1^T\boldsymbol{\gamma}\}+\frac{1}{2}\sigma_G^2(\mathbf{s}_1+\mathbf{s}_0)^T\boldsymbol{\gamma}(\mathbf{s}_1-\mathbf{s}_0)^T\boldsymbol{\gamma}. \quad (4.1)$$

We can express the DE and IE of the SNPs on the log odds ratio of disease risk in terms of the regression coefficients in models (2.1) and (2.2). For rare diseases, they are respectively approximately equal to

$$DE=(\mathbf{s}_1-\mathbf{s}_0)^T[\boldsymbol{\beta}_S+\boldsymbol{\gamma}(\mathbf{x}^T\boldsymbol{\phi}+\mathbf{s}_0^T\boldsymbol{\delta}+\beta_G\sigma_G^2)]+\frac{1}{2}\sigma_G^2(\mathbf{s}_1+\mathbf{s}_0)^T\boldsymbol{\gamma}(\mathbf{s}_1-\mathbf{s}_0)^T\boldsymbol{\gamma} \quad (4.2)$$

$$IE=(\mathbf{s}_1-\mathbf{s}_0)^T\boldsymbol{\delta}(\beta_G+\mathbf{s}_1^T\boldsymbol{\gamma}). \quad (4.3)$$

These are derived in Section 1 of the Supplementary Materials using coun-terfactuals under the assumptions of no unmeasured confounding.

The sum of the direct and indirect effects, is equal to the total effect of the SNPs, i.e., TE=DE+IE. As shown in the Supplementary Material and discussed in Section 3.2, identification of the total effect requires a much weaker assumption than those required for the direct and indirect effects.

4.2. Understanding the Null Hypothesis for eQTL SNPs

Under the assumption that the gene expression *G* is associated with the SNPs **S** (i.e., eQTL SNPs), i.e., $\boldsymbol{\delta}\neq 0$, using equations (4.2) and (4.3), the test for the joint effects of SNPs in a SNP set **S** and a gene expression *G* on *Y*, i.e., the total effect of a gene, is equivalent to a test

for the total SNP effect on the outcome (*Y*). In fact, for eQTL SNPs, which have non-zero effects on expression *G* (i.e., $\delta \neq 0$), using the expressions of DE and IE in (4.2) and (4.3), one can show that the null hypothesis of no direct and indirect genetic (SNP) effects is equivalent to the null hypothesis (2.3) that all the regression coefficients ($\boldsymbol{\beta}_S$, β_G and $\boldsymbol{\gamma}$) equal zero:

$$H_0{:}\boldsymbol{\beta}_S = \mathbf{0}, \beta_G = 0, \boldsymbol{\gamma} = \mathbf{0} \iff H_0{:}\mathrm{DE} = 0, \mathrm{IE} = 0.$$

The null hypothesis (2.3) that all the regression coefficients ($\boldsymbol{\beta}_S$, β_G and $\boldsymbol{\gamma}$) are equal to zero is also equivalent to the null hypothesis of no total effect of the SNPs provided $\boldsymbol{\beta}_S + \beta_G \boldsymbol{\delta} \neq 0$ if $\boldsymbol{\beta}_S$ or β_G is not 0 for eQTL SNPs ($\delta \neq 0$), i.e.,

$$H_0{:}\boldsymbol{\beta}_S = \mathbf{0}, \beta_G = 0, \boldsymbol{\gamma} = \mathbf{0} \iff H_0{:}\mathrm{TE} = \mathrm{DE} + \mathrm{IE} = 0. \quad (4.4)$$

We show in Section 1.4 of the Supplementary Material that the null hypothesis (4.4) requires only the assumption of no unmeasured confounding for the effect of eQTL SNPs (**S**) on the outcome (*Y*) after adjusting for the covariates (**X**). Most genetic association studies make this assumption. In other words, we make no stronger assumption than standard SNP only analyses for testing the null hypothesis of no total effect of the SNP set in a gene.

Note that in models (2.1) and (2.2), we allow other covariates (**X**) to affect both the gene expression and the disease. If the covariates **X** affect both expression and disease, ignoring **X** may cause confounding in estimating DE and IE. As shown in Figure 1, if arrows from **X** to *G* and *Y* exist and **X** is not controlled for, assumption (2) in Section 1.2 of the Supplementary Material is violated. But if the covariates **X**, the common causes of expression and disease, do not affect the SNPs **S** (no arrow from **X** to **S**), the estimation and hypothesis testing for TE is still valid. However, if there does exist an effect of **X** on **S**, then it violates the above assumption of no unmeasured confounding for the **S**-*Y* association and thus the test or estimation for TE will be biased.

4.3. Understanding the Null Hypothesis for non-eQTL SNPs

If SNPs have no effect on gene expression ($\boldsymbol{\delta} = \mathbf{0}$), i.e., they are not eQTL SNPs, then there is no indirect effect of the SNPs on *Y*, so that the null hypothesis of no total effect of a gene ($H_0 : \boldsymbol{\beta}_S = \mathbf{0}$, $\beta_G = 0$, $\boldsymbol{\gamma} = \mathbf{0}$) is not equivalent to testing for no total SNP effect on *Y* (H_0 : TE = DE + IE = 0). In this case, what the null hypothesis, $H_0 : \boldsymbol{\beta}_S = \mathbf{0}$, $\beta_G = 0$, $\boldsymbol{\gamma} = \mathbf{0}$ tries to evaluate is simply whether there exists a joint effect of the given set of SNPs **S** and the given gene expression *G*, and possibly their interactive effect, on disease risk. To test for such a joint effect, we need the first two assumptions regarding no unmeasured confounding in Section 1.2 of the Supplementary Material: no unmeasured confounding of the SNPs on the outcome and no unmeasured confounding of the gene expression on the outcome.

4.4. Understanding the Traditional Genetic Analysis Using the SNP Only Model

In standard genetic association analysis, we usually fit the following SNP only model:

$$\mathrm{logit}\{P(Y_i = 1 | \mathbf{S}_i, \mathbf{X}_i)\} = \mathbf{X}_i^T \boldsymbol{\alpha}^* + \mathbf{S}_i^T \boldsymbol{\beta}_S^*, \quad (4.5)$$

which does not take gene expression into account, but simply considers the association between the outcome and SNPs adjusting for covariates. Note for the special case where SNP, *S*, is univariate, the model (4.5) corresponds to single SNP analysis, the most common

approach in GWAS. Kwee et al. (2008) and Wu et al. (2010) have developed tests for a SNP-set for $H_0{:}\boldsymbol{\beta}_S^*=0$ under (4.5), which can be more powerful than individual SNP tests for the association between the joint effects of the SNPs in a gene and the outcome by borrowing information across SNPs within a gene, especially when the SNPs are in good linkage disequilibrium (LD).

Assuming the true models that depend on both SNPs and a gene expression are specified in (2.1) and (2.2), we study in this section how $\boldsymbol{\beta}_S^*$ in the mis-specified standard SNP only model (4.5) is related to the regression parameters $\boldsymbol{\beta}_S$, β_G and $\boldsymbol{\gamma}$ in the true model (2.1) and what the null hypothesis $H_0{:}\boldsymbol{\beta}_S^*=0$ under (4.5) tests for. To focus on the fundamental issues and for simplicity, we first discuss the case of no interaction effect between SNPs and gene expression on disease risk, i.e., $\boldsymbol{\gamma}=\mathbf{0}$ in model (2.1). Under the true [Y |**S**, G, **X**] and [G|**S**, **X**] models in (2.1) and (2.2) assuming no **S** × G interaction ($\boldsymbol{\gamma}=\mathbf{0}$), by plugging (2.2) into (2.1), the true [Y |**S**, G, **X**] model can be rewritten as $\mathrm{logit}\{P(Y_i=1|\mathbf{S}_i,\mathbf{X}_i,\varepsilon_i)\}=\mathbf{X}_i^T(\boldsymbol{\alpha}+\beta_G\boldsymbol{\phi})+\mathbf{S}_i^T(\boldsymbol{\beta}_S+\beta_G\boldsymbol{\delta})+\beta_G\varepsilon_i$. Integrating out $\varepsilon_i\sim N(0,\sigma_G^2)$, we have the true [$Y$ |**S**, **X**] model as

$$\mathrm{logit}[P(Y_i=1|\mathbf{S}_i,\mathbf{X}_i)]\approx c\left\{\mathbf{X}_i^T(\boldsymbol{\alpha}+\beta_G\boldsymbol{\phi})+\mathbf{S}_i^T(\boldsymbol{\beta}_S+\beta_G\boldsymbol{\delta})\right\}, \quad (4.6)$$

where $c=(1+0.35\times\sigma_G^2\beta_G^2)^{-1/2}$ (Zeger et al., 1988).

A comparison of (4.5) with (4.6) shows that $\boldsymbol{\beta}_S^*\approx c(\boldsymbol{\beta}_S+\beta_G\boldsymbol{\delta})$ and that the effect of **S** = $\mathbf{s}_1$ versus $\mathbf{s}_0$ on the outcome Y under the SNP only model (4.5) corresponds to $(\mathbf{s}_1-\mathbf{s}_0)^T\{c(\boldsymbol{\beta}_S+\beta_G\boldsymbol{\delta})\}$, which is proportional to the Total Effect of SNPs in 4.1) when $\boldsymbol{\gamma}=\mathbf{0}$. It follows that testing for $\boldsymbol{\beta}_S^*=\mathbf{0}$ in the SNP only model (4.5) is approximately equivalent to testing for no total effect of the SNPs.

However if there exists a SNP-by-expression interaction on Y and the SNPs are eQTL SNPs, the naive SNP only analysis using (4.5) does not provide obvious correspondence to the total SNP effect. As shown in Section 2 of the Supplementary Material, the induced true [Y |**S**, **X**] model in this setting follows

$$\mathrm{logit}\{P(Y_i=1|\mathbf{S}_i,\mathbf{X}_i)\}\approx c_i^*\left\{\mathbf{X}_i^T(\boldsymbol{\alpha}+\boldsymbol{\phi}\beta_G)+\mathbf{S}_i^T(\boldsymbol{\beta}_S+\boldsymbol{\delta}\beta_G)+\mathbf{X}_i^T\boldsymbol{\phi}\mathbf{S}_i^T\boldsymbol{\gamma}+\mathbf{S}_i^T\boldsymbol{\delta}\mathbf{S}_i^T\boldsymbol{\gamma}\right\}, \quad (4.7)$$

where $c_i^*=\{1+0.35\sigma_G^2(\beta_G+\mathbf{S}_i^T\boldsymbol{\gamma})^2\}^{-1/2}$. This implies that if the [Y |**S**, G, **X**] follows the interaction model (2.1), the induced true [Y |**S**, **X**] model depends not only on the linear terms of **X** and **S** but also on the cross-product terms of **X** and **S** and the second order term of **S**. A comparison of (4.5) with (4.7) shows that the standard SNP only model (4.5) mis-specifies the functional form of the true [Y |**S**, **X**]. The test for $H_0{:}\boldsymbol{\beta}_S^*=\mathbf{0}$ under the mis-specified SNP only model (4.5) will still be valid for testing the total effects of SNPs, because under the null the two models are the same. However, the mis-specified model is subject to power loss, compared to the test based on the correctly specified model. With only an interaction effect ($\boldsymbol{\gamma}\neq\mathbf{0}$, $\boldsymbol{\beta}_S=\mathbf{0}$, $\beta_G=0$), (4.7) can be written as: $\mathrm{logit}[P(Y_i=1|\mathbf{S}_i,\mathbf{X}_i)]\approx c_i^*\left\{\mathbf{X}_i^T\boldsymbol{\alpha}+\mathbf{X}_i^T\boldsymbol{\phi}\mathbf{S}_i^T\boldsymbol{\gamma}+\mathbf{S}_i^T\boldsymbol{\delta}\mathbf{S}_i^T\boldsymbol{\gamma}\right\}$, where $c_i^*=\{1+0.35\sigma_G^2(\mathbf{S}_i^T\boldsymbol{\gamma})^2\}^{-1/2}$. If we assume this is the true model and fit the conventional

GWAS model (4.5) to test for the SNP effect, the test is again still valid under the null, but loses power under the alternative.

4.5. Understanding the relation with Mendelian randomization

The approach here differs in several ways from that based on Mendelian randomization (Smith and Ebrahim, 2003 and 2005) in which genetic markers (SNPs) are instrumental variables to assess the effect of an exposure (in our case, a gene expression value) on an outcome. Here we are interested in using a gene expression to increase power for testing for the total effect of SNPs on a disease outcome. Furthermore, Mendelian randomization makes the assumption that SNPs do not have an effect on an outcome except through an exposure (e.g. gene expression in our case), in other words, no direct effect. No such assumption is being made here. This is because we are interested in testing for a different effect, i.e., the effect of SNPs, rather than the effect of an exposure (gene expression) on disease risk.

5. Simulation Studies

5.1. Simulation Setup

To make the simulation mimic the motivating asthma data (Moffatt et al., 2007), we simulated data using the *ORMDL3* gene on chromosome 17q21. We generated the SNP data in the *ORMDL3* gene by accounting for its linkage disequilibrium structure using HAPGEN based on the CEU sample (Marchini et al., 2007). The genomic location used to generate the SNP data is between 35.22 and 35.39 Mb on chromosome 17, which contains 99 HapMap SNPs. Ten of the 99 HapMap SNPs are genotyped on the Illumina HumanHap300 array, i.e., 10 typed SNPs.

To generate gene expression and the disease outcome, we assumed there is one causal SNP S_{causal} and varied the causal SNP among the 99 HapMap SNPs in each simulation. In Section 5.4, we further perform a simulation study assuming three causal SNPs. For subject i, gene expression G_i was generated by the linear regression model $G_i = 0 + \delta \times S_{causal,i} + \varepsilon_i$, $\varepsilon_i \sim N(0, 1.44)$. The outcome Y_i was generated by the logistic model

$$\text{logit}\{P(Y_i=1|S_{causal,i}, G_i)\} = -0.2 + \beta_S \times S_{causal,i} + \beta_G \times G_i + \gamma \times G_i S_{causal,i}.$$

The parameters and the range of β_S, β_G and γ were based on the empirical estimates from analysis of the asthma data. For each simulation, we first generated a cohort with 1000 subjects, and 100 cases and 100 controls were randomly selected from the 1000 subjects to form a case-control sample.

Two sets of simulation were performed. In the first set, we selected the SNP rs8067378 as the causal SNP, as this SNP is highly associated with asthma in the original GWAS (Moffatt et al., 2007). For each configuration of β_S, β_G, γ and δ, we generated 2000 data sets to calculate the empirical size and power. In the second set of simulation, the causal SNP was chosen one at a time out of the 99 HapMap SNPs. For each selected causal SNP, we generated 1000 data sets and evaluated statistical power for two different disease models: (β_S, β_G, γ) = (0.4, 0, 0), or (0.2, 0.2, 0). In both simulation settings, we used the 10 typed SNPs of the *ORMDL3* gene on the Illumina chip to form the SNP-set for the model (2.1), i.e., $p = 10$, in calculating the test statistics Q_S, Q_{SG}, Q_{SGC} and the omnibus test. For Q_{SG} and Q_{SGC}, both weighted and un-weighted methods were investigated, where $a_1 = 1$, $a_2 = (I_G/I_{\tau S})^{-1/2}$, and $a_3 = (I_{\tau I}/I_{\tau S})^{-1/2}$ for the weighted statistic, and $a_1 = a_2 = a_3 = 1$ for the un-weighted statistic. The p-values were calculated using the scaled χ^2 approximation, the

Davies' method by inverting the characteristic function (Davies, 1980) and the perturbation procedure with 500 perturbations. The results of these approximations were very similar at the significance level of 0.05. We performed the omnibus test by combining the evidence from Q_S, weighted Q_{SG} and weighted Q_{SGC}.

5.2. Size and Power: By Varying Effect Sizes for a Fixed Causal SNP

We first evaluated the sizes of the proposed score tests, where the null distribution was approximated by either the scaled χ^2 approximation or the perturbation procedure (Table 1). Type I errors are well protected using both approximation methods under the three models with statistics Q_S, Q_{SG}, Q_{SGC}. The empirical size is close to 0.05 for the omnibus test and the three models. As the results using different approximation methods are similar at the level of 0.05, we only present in the following the empirical power using the perturbation method. We also evaluate the performance of the proposed tests using the characteristic function inversion method (Davies, 1980) and the perturbation method at smaller sizes ($a = 5 \times 10^{-3}$ and 5×10^{-4}) (Table 1 of the Supplementary Material), and find the methods perform well.

We assumed rs8067378 is the causal SNP and eQTL, and compared the powers of the three statistics Q_S, Q_{SG}, and Q_{SGC} as well as the omnibus test under three different configurations of effect sizes (Figure 2). The first setting assumes both gene expression and the interactions between the SNPs and the gene expression have no effect on the outcome ($\beta_G = 0$ and $\gamma = 0$) (Figure 2(a)). As expected, Q_S shows the optimal power as it correctly specifies the model. The other two tests Q_{SG} and Q_{SGC}, especially Q_{SGC}, over-specify the model and waste DF for testing the effects of expression and interactions, and hence lose power. The performance of the omnibus test is close to the Q_S. As a note, the correctly specified model here means that gene expression or non-linear interaction has been incorporated in the analyses, not that typed SNPs are causal.

The second setting assumes the gene expression has an effect on the outcome but there is no interaction ($\beta_G = 0.2$ and $\gamma = 0$) (Figure 2(b)), while the third setting further assumes an interaction effect ($\beta_G = 0.2$ and $\gamma = 0.2$) (Figure 2(c)). The tests Q_{SG} and Q_{SGC}, respectively have the best power under the correct model. In contrast to the setting 1, Q_S has the worst performance among the three tests when expression has an effect on disease, and the power loss of Q_S is even more in the presence of the interactions (Figure 2(c)). The un-weighted Q_{SG} also does not perform well in these two cases and has considerable loss of power. The rest of the tests have similar power. The performance of the omnibus test in both settings is close to the optimal test with only minimal loss of power.

We also study in Figure 2 the performance of the likelihood ratio test (LRT) for testing for the joint effects of SNPs and gene expression by comparing the model with an intercept, SNPs, gene expression and interactions with the model with only the intercept. In general, our proposed methods outperform the LRT in both power and type I error. The power loss and the incorrect size of the LRT are likely due to the large degrees of freedom relative to the sample size (DF=21; 100 cases/100 controls) and the high LD among some of the typed SNPs.

5.3. Power: By Varying Causal SNPs

In order to investigate the performance under "synthetic association", i.e., the causal variant is untyped (i.e, not on a chip) (Dickson et al., 2010), we assessed the power of the methods when each of the 99 HapMap SNPs was assumed to be causal. In particular, we were interested in evaluating how the correlation between the causal SNP and the 10 typed SNPs affected the statistical powers of the proposed tests. Intuitively, if the causal SNP is untyped

and has low LD with the typed SNPs, one would expect lower power. We considered two settings: the outcome is only associated with the causal SNP: $\beta_S = 0.4$, $\beta_G = 0$, $\gamma = 0$ (Figure 3(a)); and the outcome is associated with the causal SNP and the gene expression but not their interaction: $\beta_S = 0.2$, $\beta_G = 0.2$, $\gamma = 0$ (Figure 3(b)). A total of 1000 simulations were performed.

Figures 3(a) and (b) show that the pattern of simulation results is very similar to those in the previous section. The test assuming the correct model performs the best. The omnibus test nearly reaches the optimal power obtained under the true model in both settings. In addition, the test using the weighted statistic derived under the model with SNPs and gene expression as predictors (weighted Q_{SG}) performs well even when the interaction model is true (data not shown), although it has some loss of power in setting 1 when only SNPs are associated with the outcome. As the model (2.1) can be written as $\text{logit}[P(Y_i=1|\mathbf{S}_i, G_i, \mathbf{X}_i)]=\mathbf{X}_i^T\boldsymbol{\alpha}+\mathbf{S}_i^T\boldsymbol{\beta}^*+G_i\beta_G$ where $\boldsymbol{\beta}^* = \boldsymbol{\beta}_S + G_i\boldsymbol{\gamma}$, the main effect only analyses can still capture the interactive effect γ even though the model is not correctly specified. So the simpler test Q_{SG} can be used as an alternative to the omnibus test if the computation cost is a concern.

Figure 3 also shows that statistical power depends on the correlation between the causal SNP (which might be untyped) and the 10 typed SNP. The power rises as the correlation between the causal SNP and the typed SNPs increases. For example, the statistical power is high if a causal SNP is in the LD block spanned between the second to the third typed SNPs (marked as the second and third black triangles from left to right, according to the physical location, in Figure 3), as it has good correlation with the typed SNPs. The power is generally low if a causal variant lies in the region between the first and second typed SNPs as it has little correlation with the typed SNPs. In this case, it is virtually not possible to detect genetic effects using the typed SNPs on the chip no matter what method one uses. Although the typed SNPs might not include the underlying causal SNPs, it still provides a valid testing procedure due to the same model under the null. However, the typed SNPs may or may not provide a consistent estimate for the effect of the causal SNP, depending on the LD pattern of the causal SNP and the typed SNPs.

5.4. Additional Simulations: Model mis-specification, Multiple Causal Variants, Varying LD structures

We performed additional simulations to assess how model mis-specification influences our proposed test. Gene expression G_i was generated without the normality assumption $G_i = 0 + \delta S_{causal,i} + \varepsilon_i$, $\varepsilon_i \sim N(0, 1.44) + uniform(-0.3, 0.3)$. Two outcome models are explored. The first model generates the outcome Y_i by the logistic model assuming non-linear effects of SNPs and G as $\text{logit}\{P(Y_i = 1|S_{causal,i}, G_i)\} = -100^{0.9} + (100 + \beta_S S_{causal,i} + \beta_G G_i + \gamma G_i S_{causal,i})0.9$, and the second model generates Y_i by a probit model $\Phi^{-1}\{P(Y_i = 1|S_{causal,i}, G_i)\} = -0.2 + \beta_S S_{causal,i} + \beta_G G_i + \gamma G_i S_{causal,i}$. Although the model is not correctly specified in our proposed test under these settings, the joint analyses of SNPs and expression still outperform the SNP-only analyses when the gene expression contributes to the risk of developing disease. Similarly, the performance of the omnibus test is very close the the optimal test obtained under the true model for different scenarios (Figure 4).

We conducted two additional simulation studies by varying the number of causal variants and LD structures. The pattern of the results from these additional studies is very similar to what is presented above (Figures 1 and 2 in the Supplementary Material). The first additional study is similar to the study in Section 5.2, except that there are three causal SNPs in the *ORMDL3* gene instead of a single causal SNP. Using the same ten typed SNPs for the analyses, we again found that the test performs the best when the model is correctly

specified and the omnibus test approaches the optimal test obtained under the true model with limited power loss (Figure 1 of the Supplementary Material).

Similar to analyses in Section 5.3, the second additional study investigates the performance of the proposed test at 15q24-15q25.1 where SNPs have a different LD pattern from the *ORMDL3* gene. Assuming one causal SNP at a time, we used the same ten typed SNPs to perform our proposed test. Again, the test performs the best when the model is correctly specified, and the omnibus test is robust and approaches the optimal test obtained by assuming the true model, and the power depends on the correlation of the causal SNP and the typed SNPs (Figure 2 of the Supplementary Material).

6. Analysis of the Asthma Data

We applied the proposed testing procedures to re-investigate the genetic effects of the *ORMDL3* gene on the risk of childhood asthma in the MRC-A data (Dixon et al., 2007; Moffatt et al., 2007). This subset of the data contained 108 asthma cases and 50 controls where we have complete data of the 10 typed SNPs and gene expression of *ORMDL3*. The SNP data were genotyped using the Illumina 300K chip and the gene expression was collected using the Affymetrix Hu133A 2.0. We analyzed the data using both additive and dominant modes: in the additive mode, the genotype was coded as the number of the minor allele (i.e., 0, 1, 2), whereas in the dominant mode, the genotype was coded as whether or not the minor allele was present (i.e., 0, 1).

We applied the proposed tests for the total SNP effect of *ORMDL3* using the SNP and gene expression data. There are strong associations between the SNPs and the gene expression (8 out of 10 with p-value<0.05 and the other two with p-values of 0.076 and 0.21), i.e., the SNPs are eQTL SNPs. We considered six test statistics: Q_S, un-weighted Q_{SG}, weighted Q_{SG}, un-weighted Q_{SGC}, weighted Q_{SGC}, and omnibus test (Table 2). We compared our methods with the standard multivariate or univariate methods: the multivariate Wald test, which has 10, 11 and 21 degrees of freedom under the three models (SNPs only, SNPs and gene expression, and SNPs, gene expression and interactions). We also included in the comparison the test using the smallest p-value from the 10 single SNP analyses with the Bonferroni adjustment or the adjustment using the permutation procedure to account for the correlation among the SNPs.

The results in Table 2 show that our proposed methods give smaller p-values compared to the standard testing procedures. The test Q_{SGC}, which accounts for the effects of SNPs, gene expression and their interactions, gives the smallest p-value compared to the tests only using SNPs, in both additive and dominant modes. For example, the p-values using weighted Q_{SGC} and Q_S are 0.0028 and 0.044 respectively using the additive SNP model. The omnibus test calculated using the perturbation procedure by computing the minimum p-value from Q_S, weighted Q_{SG} and weighted Q_{SGC} also provides a more significant signal than those only considering SNPs, with the p-value being 0.0055. These results are consistent with the findings in simulation studies.

We also performed genome-wide analyses for both SNP-sets and single SNP. We first paired eQTL SNPs with their corresponding gene expression (Dixon et al., 2007) and performed SNP-only analyses and our proposed method. For single SNP analyses, after adjustment of multiple comparison using false discovery rate (FDR; Storey, 2002), 56 SNPs with FDR <0.1 were identified in SNP-only analyses and 97 SNPs were identified from the proposed omnibus test. For SNP-set analyses, we grouped eQTL SNPs that correspond to the same gene as a SNP-set, and we identified 5 and 15 SNP-sets (FDR<0.1) from SNP-only analyses and omnibus tests, respectively.

7. Discussion

We proposed in this paper to integrate SNP and gene expression data to improve power for genetic association studies. The major contributions of this paper are: 1) to formulate the data integration problem of different types of genomic data as a mediation problem; 2) to propose a powerful and robust testing procedure for the total effect of a gene contributed by SNPs and a gene expression; and 3) to relax the assumptions required for mediation analyses in the test for the total effect.

Specifically, as shown in Figure 1, we are able to integrate the information of SNPs and gene expression as a biological process through the mediation model. Our proposed variance component score test for the total effect of SNPs and a gene expression circumvents the instability of estimation of the joint effects of multiple SNPs and gene expression, because only the null model needs to be fit. Mediation analysis to estimate direct and indirect effects generally requires additional unmeasured confounding assumptions, and previous work mainly focused on estimation. Here we focus on testing for the total effect of a gene using SNPs and a gene expression. For eQTL SNPs, we show that the total effect of a gene contributed by SNPs and a gene expression is equivalent to the total effect of SNPs, which is the sum of direct and indirect effects of SNPs mediated through gene expression. Testing for the total SNP effect only requires one assumption: no un-measured confounding for the effect of SNPs on the outcome, which is the same assumption as the standard GWAS and thus no stronger assumption is required.

We characterize the relation among SNPs, gene expression and disease risk in the framework of causal mediation modeling. This framework allows us to understand the null hypothesis of no total effect of a gene contributed by SNPs and gene expression, and the underlying assumptions of the test for both eQTL SNPs and non-eQTL SNPs. We propose a variance component score test for the total effects of a gene on disease. This test allows to jointly test for the effects of SNPs, gene expression and their interactions. We show that the proposed test statistic follows a mixture of χ^2 distributions asymptotically, and proposed to approximate its finite sample distribution using a scaled χ^2 distribution, a characteristic function inversion method or a perturbation method.

We considered three tests: using only SNPs (Q_S), SNPs and gene expression main effects (Q_{SG}), and SNPs, gene expression and their interactions (Q_{SGC}). Our simulation study shows that all three tests have the correct type I error for testing for the overall SNP effect. The relative power of these tests depends on the underlying true relation between the predictors (SNPs, gene expression and their interactions) and disease. As the underlying biology is often unknown, we further constructed the omnibus test that identifies the most powerful test among the three disease models, and proposed to use the perturbation method to calculate the p-value for the omnibus test. Further, the test using only the main effects of SNPs and gene expression loses limited power compared to the omnibus test and can be used as a simple alternative.

Our results also show that to test for the total effects of a gene, the tests that incorporate both SNP and gene expression information, such as Q_{SG} and Q_{SGC}, are more powerful if SNPs are associated with gene expression than if they are not. In other words, it is even more beneficial to incorporate gene expression data with SNP data to detect genetic effects on disease if gene expression is a good causal mediator for the SNPs. To achieve this, a natural way is to select SNPs located within or at the neighborhood of a gene, since it has been well-established that the SNP within a gene can alter its expression value via transcription regulation (Lee et al., 2008). Alternatively, one can restrict the joint SNP-expression analysis to known eQTL SNPs. If selection of eQTL SNPs is based on statistical

significance, one also needs to be aware of the possibility that cis-action may be a confounding effect of SNPs on array hybridization (Li et al., 2006).

We mainly focus on testing for the total effect of a gene in this paper. The proposed method can be easily extended to test for direct and indirect effects separately for eQTL SNPs. Using equation (4.2), to test for the direct effect of the SNPs, one can test $H_0 : \boldsymbol{\beta}_S = 0, \boldsymbol{\gamma} = 0$. Using the notation in equation (3.2), one can test this null hypothesis using the statistic $Q_{DE} = n^{-1}(a_1 U_{\tau_S} + a_3 U_{\tau_I})$, where the null model is a logistic model with $\boldsymbol{X}$ and $\boldsymbol{G}$. To test for the indirect effects of the SNPs, using equation (4.3), one can test $H_0 : \beta_G = 0, \boldsymbol{\gamma} = 0$. Using the notation in equation (3.2), one can test this null hypothesis using the statistic $Q_{IE}=n^{-1}(a_2 U^2_{\beta_G} + a_3 U_{\tau_I})$, where the null model is a logistic model with $\boldsymbol{X}$ and $\boldsymbol{S}$. As the number of SNPs $\boldsymbol{S}$ might be large and some SNPs might be highly correlated (with high LD), standard regression to fit the null model might not work well. One can fit the null model using ridge regression. To perform these tests, one will need to make the four unmeasured confounding assumptions required for estimating direct and indirect effects of SNPs stated in Section 1.2 of the Supplementary Material.

Gene expression may not be the only mediator for the relation between SNPs and disease. Other biomarkers, such as DNA methylation, proteins, metabolites of the gene product in the blood, immunological or biochemical markers in the serum, and environmental factors can also serve as potential mediators, depending on the context or the disease to be studied. For instance, epigenetic variations have been reported to exert heritable phenotypic effects (Johannes et al., 2008). Furthermore, our proposed test can be applied to address many other scientific questions as long as there exist a causal relationship as illustrated in Figure 1. For example, the SNP-gene-disease relations can be replaced by the DNA copy number-protein-cancer stage (early vs. late) in tumor genomics studies to assess if copy number can have any effect on the clinical stage of cancer. It is advantageous to set up a biologically meaningful model before applying our proposed test procedure, which makes the best use of the prior knowledge.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work is supported by grants from the National Cancer Institute (R37-CA076404 and P01-CA134294) and the National Institute of Environmental Health P42-ES016454. The authors would like to thank the editor, the associate editor and the referees for their helpful comments that have improved the paper.

References

1. Cai T, Lin X, Carroll R. Identifying Genetic Marker Sets Associated with Phenotypes via an Efficient Adaptive Score Test. Biostatistics. 2012 In press.
2. Carlo C. Oncogene and cancer. New England Journal of Medicine. 2008; 358:502–511. [PubMed: 18234754]
3. Cheung V, Spielman R, Ewens K, Weber T, Morley M, Burdick J. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005; 437:1365–1369. [PubMed: 16251966]
4. Cusanovich DA, Billstrand C, Zhou X, Chavarria C, De Leon S, Michelini K, et al. The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. Human Molecular Genetics. 2012; 21:2111–2123. [PubMed: 22286170]

5. Davies R. The distribution of a linear combination of chi-square random variables. Applied Statistics. 1980; 29:323–333.
6. Dermitzakis E. From gene expression to disease risk. Nature Genetics. 2008; 40:492–493. [PubMed: 18443581]
7. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biology. 2010; 8:e1000294. [PubMed: 20126254]
8. Dixon A, Liang L, Moffatt M, Chen W, Heath S, Wong K, et al. A genome-wide association study of global gene expression. Nature Genetics. 2007; 39:1202–1207. [PubMed: 17873877]
9. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, Ward JL, Beale MH, de Vos RC, Dijkstra M, Scheltema RA, Johannes F, Koornneef M, Vreugdenhil D, Breitling R, Jansen RC. System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nature Genetics. 2009; 41:166–167. [PubMed: 19169256]
10. Hageman RS, Leduc MS, Korstanje R, Paigen B, Churchill GA. A Bayesian Framework for Inference of the Genotype-Phenotype Map for Segregating Populations. Genetics. 2011; 4 :1163–1170. [PubMed: 21242536]
11. Hunter D, Chanock S. Genome-wide association studies and “the art of the soluble”. Journal of National Cancer Institute. 2010; 102:1–2.
12. Hsu Y, Zillilkens M, Wilson S, Farber C, Demissie S, Soranzo N, et al. An integration of genome-wide association study and expression profiling to prioritize the discovery of susceptibility loci for osteoporosis-related traits. PLoS Genetics. 2010; 6:e1000977. [PubMed: 20548944]
13. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science. 2010; 25:51–71.
14. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, et al. Identification, replication and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genetics. 2011; 7:e1002078. [PubMed: 21637794]
15. Johannes F, Colot V, Jansen RC. Epigenome dynamics: a quantitative genetics perspective. Nature Reviews Genetics. 2008; 9:883–890.
16. Kline P, Santos A. A score based approach to wild bootstrap inference. Journal of Econometric Methods. 2012:1, 2341, 2156–6674.10.1515/2156-6674.1006
17. Kwee L, Liu D, Lin X, Ghosh D, Epstein M. A powerful and flexible multilocus association test for quantitative traits. The American Journal of Human Genetics. 2008; 82:386397.
18. Lee P, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Research. 2008; 36:D820–D824. [PubMed: 17986460]
19. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazen-donk E, Prins P, Plasterk RH, Jansen RC, Breitling R, Kammenga JE. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. PLoS Genetics. 2006; 2:e222. [PubMed: 17196041]
20. Lin X. Variance component test in generalised linear models with random effects. Biometrika. 1997; 84:309–326.
21. Li Y, Tesson B, Churchill G, Jansen R. Critical reasoning on causal inference in genome-wide linkage and association studies. Trends in Genetics. 2010; 26:493–498. [PubMed: 20951462]
22. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics. 2007; 39:906–913. [PubMed: 17572673]
23. Moffatt M, Kabesch M, Liang L, Dixon A, Strachan D, Heath S, et al. Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. Nature. 2007; 448:470–473. [PubMed: 17611496]
24. Morley M, Molony C, Weber T, Devlin J, Ewens K, Spielman R, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430:743–747. [PubMed: 15269782]
25. Neto EC, Broman AT, Keller MP, Attie AD, Zhang B, Zhu J, Yandell BS. Modeling causality for pairs of phenotypes in system genetics. Genetics. 2013; 193:1003–1013. [PubMed: 23288936]
26. Parzen M, Wei L, Ying Z. A resampling method based on pivotal functions. Biometrika. 1994; 81:341–350.

27. Pearl, J. Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. Morgan Kaufmann; San Francisco: 2001. Direct and indirect effects; p. 411-420.
28. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–411.
29. Robins J, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992; 3:143–155. [PubMed: 1576220]
30. Robins, J. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P.; Hjort, NL.; Richardson, S., editors. Highly Structured Stochastic Systems. New York, NY: Oxford University Press; 2003. p. 7081
31. Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology. 1974; 66:688–701.
32. Rubin D. Bayesian inference for causal effects. Annals of Statistics. 1978; 6:34–58.
33. Satterthwaite F. An approximate distribution of estimates of variance components. Biometrics Bulletin. 1946; 2:110–114. [PubMed: 20287815]
34. Schadt E, Monks S, Drake T, Lusis A, Che N, Colinayo V, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422:297–302. [PubMed: 12646919]
35. Schadt E, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics. 2005; 37:710–717. [PubMed: 15965475]
36. Smith DG, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology. 2003; 32:1–22. [PubMed: 12689998]
37. Smith DG, Ebrahim S. What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? British Medical Journal. 2005; 330:1076–1079. [PubMed: 15879400]
38. Storey J. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B. 2002; 64:479–498.
39. VanderWeele T, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface (Special Issue on Mental Health and Social Behavioral Science). 2009; 2:457–468.
40. VanderWeele T, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. American Journal of Epidemiology. 2010; 172:1339–1348. [PubMed: 21036955]
41. Wu M, Kraft P, Epstein M, Taylor D, Chanock S, Hunter D, et al. Powerful SNP set analysis for case-control genomewide association studies. American Journal of Human Genetics. 2010; 86:929–942. [PubMed: 20560208]
42. Zeger S, Liang K, Albert P. Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics. 1988; 44:1049–1060. [PubMed: 3233245]
43. Zhang M, Liang L, Morar N, Dixon AL, Lathrop GM, Ding J, et al. Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. Human Genetics. 2012; 131:615–623. [PubMed: 22006220]
44. Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, MacNeil DJ, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genetics. 2010; 6:e1000932. [PubMed: 20463879]
45. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics. 2008; 40:854–861. [PubMed: 18552845]

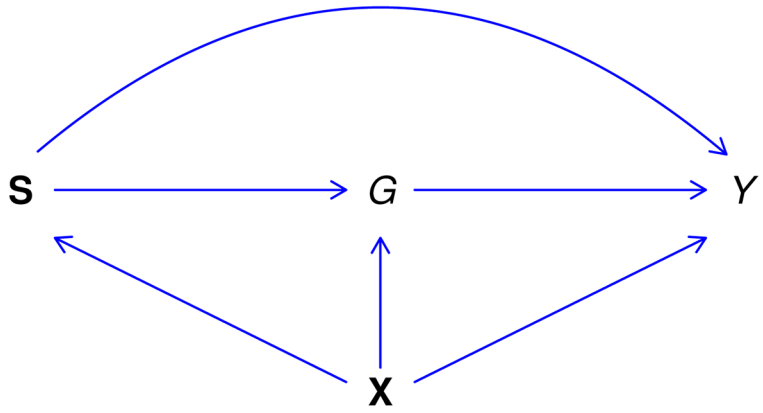
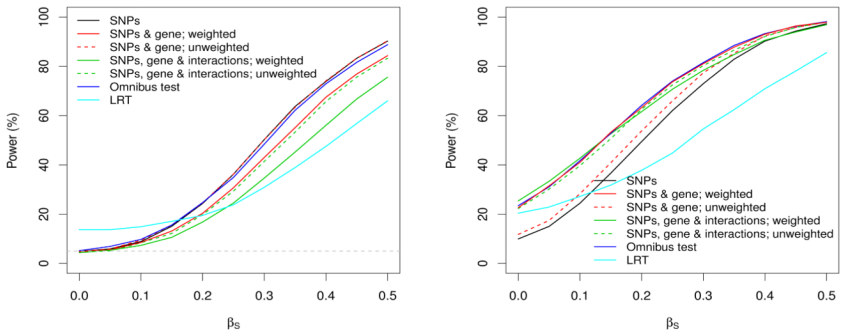
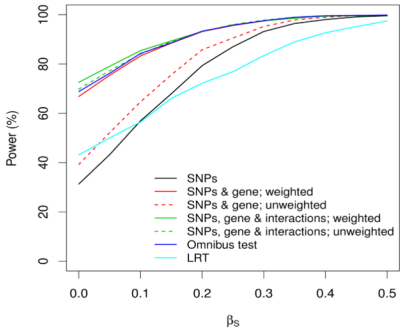


Fig 1.

Causal diagram of the mediation model. **S** is a set of correlated exposure, e.g., SNP set; *G* is a mediator, e.g., gene expression; *Y* is an outcome, e.g., disease (yes/no); and **X** are covariates, including the true and potential confounders.



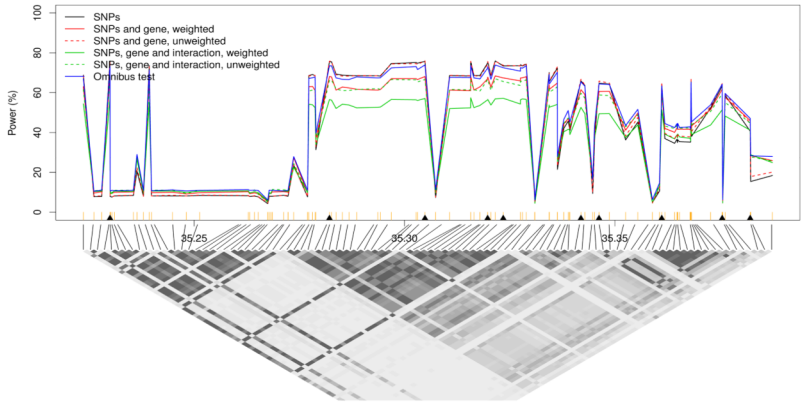
(a) $\beta_G = 0, \gamma = 0; \delta = 1.0$ (b) $\beta_G = 0.2, \gamma = 0; \delta = 1.0$



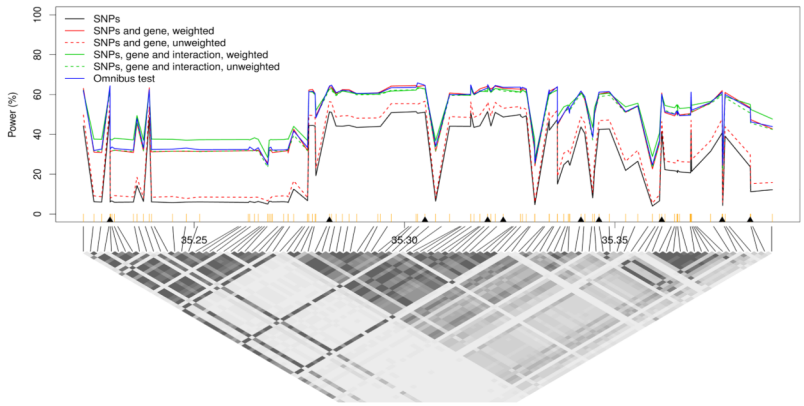
(c) $\beta_G = 0.2, \gamma = 0.2; \delta = 1.0$

Fig 2.

Empirical power. SNPs are assumed to be eQTL SNPs ($\delta = 1$). Each figure plots the powers of the proposed tests as a function of the main effect of the SNP (β_s). The three figures correspond to the three different true models, the model with only SNP effects, the model with main effects without interaction, and the model with SNPs, gene expression and their interaction effects. The dashed line in (a) indicates 5% type I error rate.



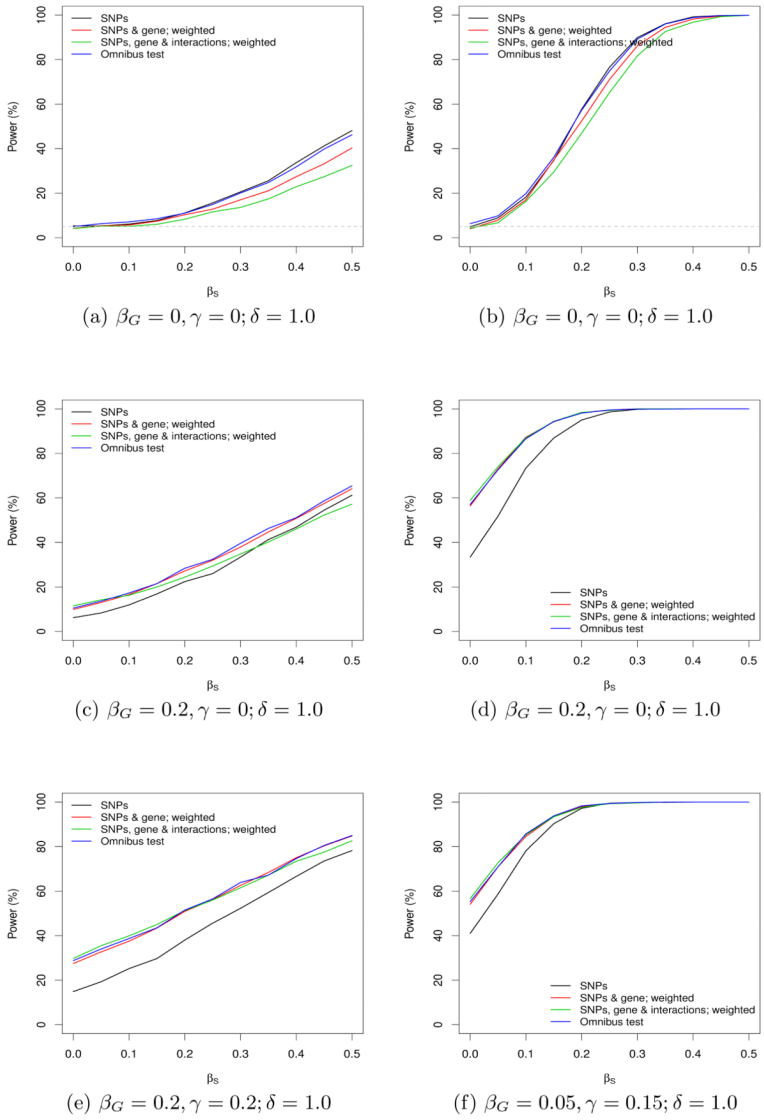
(a) $\beta_S = 0.4, \beta_G = 0, \gamma = 0$



(b) $\beta_S = 0.2, \beta_G = 0.2, \gamma = 0$

Fig 3.

Simulated power curves for evaluating how different choices of causal SNPs affect the powers of the proposed tests. The x-axis indicates the physical location (Mb) of the 99 HapMap SNPs at 17q21. The orange vertical bar indicates the relative locations of the causal SNP and the black triangles indicate the ten typed SNPs. Different lines indicate the power of different tests. The lower panel of each subfigure is the plot for linkage disequilibrium, measured as r^2 ranging from 0 (white) to 1 (black).



(a) $\beta_G = 0, \gamma = 0; \delta = 1.0$ (b) $\beta_G = 0, \gamma = 0; \delta = 1.0$

(c) $\beta_G = 0.2, \gamma = 0; \delta = 1.0$ (d) $\beta_G = 0.2, \gamma = 0; \delta = 1.0$

(e) $\beta_G = 0.2, \gamma = 0.2; \delta = 1.0$ (f) $\beta_G = 0.05, \gamma = 0.15; \delta = 1.0$

Fig 4.

Empirical power under model mis-specification. SNPs are assumed to be eQTL SNPs (δ = 1). Each figure plots the powers of the proposed tests as a function of the main effect of SNP (β_S). The six figures correspond to the different true models: the model with only SNP effects ((a) and (b)), the model with main effects of SNP and gene expression ((c) and (d)), and the model with SNPs, gene expression and their interaction effects ((e) and (f)). (a) (c) (e) are simulated under logit[P (Y_i = 1|$S_{causal,i}$, G_i)] = $-100^{0.9}$ + (100 + $\beta_S S_{causal,i}$ + $\beta_G G_i$ + $\gamma G_i S_{causal,i}$)$^{0.9}$ and (b), (d), (f) are simulated under the probit model Φ^{-1}[P (Y_i = 1|$S_{causal,i}$, G_i)] = −0.2 + $\beta_S S_{causal,i}$ + $\beta_G G_i$ + $\gamma G_i S_{causal,i}$. The dashed lines in (a) and (b) indicate 5% type I error rate.

Table 1

Empirical sizes (%) of the proposed tests using scaled χ^2 approximation and the perturbation. The size was calculated at the significance level of 0.05 based on 2000 simulations.

	Scaled χ^2 approximation		Perturbation	
	Un-weighted	**Weighted**	**Un-weighted**	**Weighted**
Gene expression depends on SNPs				
SNPs	4.65		4.90	
SNPs and expression	4.80	4.95	4.80	4.80
SNPs, expression and interaction	4.75	4.60	4.60	4.35
Omnibus test	-			5.15
Gene expression and SNPs are independent				
SNPs	4.83		4.97	
SNPs and expression	4.87	4.90	4.83	4.87
SNPs, expression and interaction	4.60	4.80	4.77	5.07
Omnibus test	-			5.13

Ann Appl Stat. Author manuscript; available in PMC 2014 April 09.

Table 2

p-values of the effects of the 10 typed SNPs at *ORMDL3* on the risk of childhood asthma. Different rows indicate the predictors to be tested. $P_{\min}$, calculates the minimum p-value using individual SNP analyses; VCT, the proposed variance component test.

	Multivariate Wald	**Bonferroni-adjusted P_{min}**	**Permutation-adjusted P_{min}**	**VCT unweighted**	**VCT weighted**
Additive model					
SNPs	0.122	0.102	0.039	0.044	
SNPs, gene	0.342	0.194	0.057	0.039	0.033
SNPs, gene and interaction	0.013	0.303	0.093	0.0025	0.0028
Omnibus test	-			0.0055	
Dominant model					
SNPs	0.018	0.015	0.018	0.0045	
SNPs, gene	0.094	0.031	0.018	0.0040	0.0040
SNPs, gene and interaction	0.131	0.098	0.048	0.0035	0.0023
Omnibus test	-			0.0030	

Ann Appl Stat. Author manuscript; available in PMC 2014 April 09.