Biostatistics (2014), **15**, 4, *pp.* 587–602
doi:10.1093/biostatistics/kxu014
Advance Access publication on April 4, 2014

Integrative modeling of multiple genomic data from different types of genetic association studies

YEN-TSUNG HUANG
Department of Epidemiology, Brown University, Providence, RI 02912, USA
yen-tsung_huang@brown.edu

SUMMARY

Genome-wide association studies (GWASs) and expression-/methylation-quantitative trait loci (eQTL/mQTL) studies constitute popular approaches for investigating the association of single nucleotide polymorphisms (SNPs) with disease and expression/methylation, respectively. Here, we propose to integrate QTL studies to more powerfully test the SNP effect on disease in GWASs when they are conducted among *different* subjects. We propose a model for the joint effect of SNPs, methylation, and gene expression on disease risk and obtain the marginal model for SNPs by integrating out methylation and expression. We characterize all possible causal relations among SNPs, methylation, and expression and study the corresponding null hypotheses of no SNP effect in terms of the regression coefficients in the joint model. We develop a score test for variance components of regression coefficients to evaluate the genetic effect. We further propose an omnibus test to accommodate different models. We illustrate the utility of the proposed method in an asthma GWAS study, a brain tumor study, and numerical simulations.

Keywords: Data integration; Epigenetics; Mediation analysis; Variance component test.

1. INTRODUCTION

Genetic association studies have been a popular approach for assessing the association of single nucleotide polymorphisms (SNPs) with various phenotypic traits. Genome-wide association studies (GWASs) have been widely used in investigating the genetic etiology of diseases. In addition, genetic methylation or gene expression can also be construed as a kind of phenotypic trait on the molecular scale. Such types of genetic association studies focusing on expression- and methylation-quantitative trait loci, respectively, are so-called eQTL and mQTL studies. Unlike GWASs where usually only peripheral blood samples or buccal cells are required for genotyping, eQTL or mQTL is tissue-specific because the methylation or expression profile varies across different organs and tissues. Methods are available to integrate multiple genomic data to draw inference on the structure of a biological network (Schadt *and others*, 2005; Zhu *and others*, 2008). We focus in this paper on gene-based analyses of multiple eQTL and mQTL SNPs of a gene and

the corresponding methylation and mRNA expression for their effects on disease phenotypes, assuming that the causal structure of SNPs, methylation and expression is known.

As both GWASs and QTL studies become a standard practice in genetic studies, considerable interest emerges in integrating the two. In fact, there have been published studies that combine eQTL studies with GWAS for diseases (Moffatt *and others*, 2007; Xiong *and others*, 2012). These studies consider SNP-disease and SNP-expression associations separately. The association of the QTL SNPs with methylation/expression is not necessarily translated to be a contribution to the disease risk. To address this, we have proposed to jointly analyze SNP and expression data from the *same* subjects (Huang *and others*, 2014). However, a GWAS and QTL study are likely to be conducted in *different* subjects due to the availability of tissue samples and the tissue specificity of expression and DNA methylation. Thus, we extend the methodology to analyze the data from *different* subjects, and propose a general framework to incorporate more than two genomic data.

This article is motivated by an asthma study, in which the association between SNPs at the *ORMDL3* gene and the risk of childhood asthma was discovered in the MRCA dataset, a case–control study, and validated in other studies (Moffatt *and others*, 2007). The MRCE dataset is a case–control study for eczema where SNP and expression data are also both available. Nevertheless, not many GWASs have available expression data. Thus, we take the unique advantage of the MRCA and MRCE datasets to artificially assemble them into one GWAS for asthma (MRCA) and an eQTL study for the *ORMDL3* gene (MRCE or MRCA). We integrate the two studies to jointly model the effect of SNPs and expression on asthma risk.

We are interested in the *ORMDL3* gene, so instead of analyzing individual SNPs, we combine multiple SNPs at *ORMDL3*. Such a multi-SNP approach has been advocated to jointly analyze multiple related SNPs in a gene and to decrease the number of tests, which has also been shown to have better performance in a breast cancer GWAS than the single SNP analysis (Wu *and others*, 2010). In addition to the evidence on the relations among SNPs and gene expression of *ORMDL3* and asthma risk (Moffatt *and others*, 2007), there is also a molecular study on the joint effect of SNPs and methylation on regulation of *ORMDL3* expression (Berlivet *and others*, 2012). Thus, we propose a method to analyze SNPs, methylation, and expression jointly on disease risk. However, there is no publicly available methylation data in MRCA or MRCE, and so we rely on numerical studies to study the performance of our method when analyzing the SNP set, the methylation, and the expression jointly.

In this article, we propose an analytic way of integrating multiple genetic studies (e.g. eQTL and mQTL studies and GWAS) in a regression framework, which is constructed based on biology. We jointly model an SNP set within a gene, its methylation and expression, and the outcome using a regression model, and integrate out methylation and expression to obtain a marginal model. The null hypothesis of no SNP effect in the marginal model corresponds to different zero-coefficients in the joint model depending on the relations of SNPs, methylation, and expression. We enumerate all possible SNP–methylation–expression relations and provide causal interpretations under the framework of mediation modeling (Robins and Greenland, 1992; Mackinnon, 2008). According to SNP–methylation–expression relations, we develop efficient testing procedures for the SNP effect.

The rest of the paper is organized as follows. In Section 2, we introduce the joint model for SNPs, methylation, and gene expression on disease risk, and the marginal model for SNPs. In Section 3, we introduce the null hypothesis of no SNP effect and study how different relationships among SNPs, methylation, and gene expression can affect the correspondence between no SNP effect and coefficients in the joint model. In Section 4, we propose a variance component score test for the SNP effect and construct an omnibus test to optimize the test power across different underlying disease models. In Section 5, we conduct numerical studies to evaluate the performance of our proposed test. In Section 6, we illustrate the utility of our methods in an asthma study and a tumor genomic study. We conclude with a discussion in Section 7.

2. The model

Assume for subject i $(i = 1, \ldots, n)$, the outcome of interest Y_i, with mean associated with q covariates $(\mathbf{X}_i)$, p SNPs $(\mathbf{S}_i)$, one methylation (M_i), one mRNA expression (G_i) of a gene, and their possible cross-product interactions through the following model:

$$\begin{aligned}\mathcal{G}[\mu_{\text{cond}}(Y_i|\mathbf{S}_i, M_i, G_i, \mathbf{X}_i)] = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + h(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\beta}_S + M_i\beta_M + G_i\beta_G + M_i h(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\beta}_{SM} \\ + G_i h(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\beta}_{SG} + M_i G_i\beta_{MG} + M_i G_i h(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\beta}_{SMG} = \eta_{\text{cond},i}, \qquad (2.1)\end{aligned}$$

where $\mathcal{G}(\cdot)$ is a strictly increasing function with differentiable inverse function $\mathcal{G}^{-1}(\cdot)$; $h(\cdot)$ is a differentiable non-constant function; and μ_{cond} is a conditional mean. When the outcome of interest Y_i is dichotomous and $\mathcal{G}$ is the logistic link, it becomes a logistic model (Prentice and Pyke, 1979). Here, we define $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_S{}^{\mathrm{T}}, \beta_M, \beta_G, \beta_{MG}, \boldsymbol{\beta}_{SM}{}^{\mathrm{T}}, \boldsymbol{\beta}_{SG}{}^{\mathrm{T}}, \boldsymbol{\beta}_{SMG}{}^{\mathrm{T}})$. As methylation can be affected by SNPs (Berlivet *and others*, 2012) and gene expression can be affected by SNPs and methylation (Morley *and others*, 2004; Berlivet *and others*, 2012), we further consider the following models for methylation M and expression G:

$$M_i = \mu_{Mi} + \epsilon_{Mi} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\delta}_0 + h_M(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\delta}_S + \epsilon_{Mi}, \qquad (2.2)$$

$$G_i = \mu_{G|Mi} + \epsilon_{G|Mi} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_0 + h_G(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\alpha}_S + M_i\alpha_M + M_i h_G(\mathbf{S}_i)^{\mathrm{T}}\boldsymbol{\alpha}_{SM} + \epsilon_{G|Mi}, \qquad (2.3)$$

where $\epsilon_{Mi} \sim F_M(0, {\sigma_M}^2)$, $\epsilon_{G|Mi} \sim F_{G|M}(0, {\sigma_G}^2)$, F_M, and $F_{G|M}$ can be any arbitrary distribution; $h_M(\cdot)$ and $h_G(\cdot)$ are differentiable non-constant functions. Note $\mu_{Mi} = E(M_i|\mathbf{X}_i, \mathbf{S}_i)$, $\mu_{G|Mi} = E(G_i|\mathbf{X}_i, \mathbf{S}_i, M_i)$ and we further define $\mu_{Gi} = E(G_i|\mathbf{X}_i, \mathbf{S}_i)$ and $\mu_{MGi} = E(M_i G_i|\mathbf{X}_i, \mathbf{S}_i)$. We also assume that ϵ_M and $\epsilon_{G|M}$ are independent of $\mathbf{S}$. By integrating out G_i and M_i, we can obtain the marginal distribution that only depends on the SNPs $\mathbf{S}_i$ and the covariates $\mathbf{X}_i$, $[Y_i|\mathbf{S}_i, \mathbf{X}_i]$:

$$\mathcal{G}[\mu_{\text{marg}}(Y_i|\mathbf{S}_i, \mathbf{X}_i)] = \mathcal{G}\left\{\int\int \mu_{\text{cond}}(Y_i|\mathbf{S}_i, M_i, G_i, \mathbf{X}_i)\, \mathrm{d}F_{G|M}(g|m)\, \mathrm{d}F_M(m)\right\} = \eta_i. \qquad (2.4)$$

We will later develop testing procedures for both the joint model (2.1) and marginal model (2.4). We show in supplementary material available at *Biostatistics* online that the conventional model assuming linearity of SNP effect may misspecify the marginal model, but the two models will coincide under strong assumptions.

3. Null hypothesis

The null hypothesis of a genetic association study is that after adjusting for covariates $\mathbf{X}$, the outcome Y of the subjects carrying genotypes $\mathbf{s}_1$ is the same as those carrying $\mathbf{s}_0$:

$$H_0 : \Delta = \mathcal{G}[\mu(Y|\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s_1})] - \mathcal{G}[\mu(Y|\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s_0})] = 0. \qquad (3.1)$$

Different genetic models (dominant, recessive, or additive) follow the same null hypothesis: the outcome of Y of the subjects carrying different genotypes is the same. Under the alternative, (3.1) can be changed to accommodate a different genetic model, e.g. for additive model, $H_A : \Delta = \mu(Y|\mathbf{X}, \mathbf{s}_2) - \mu(Y|\mathbf{X}, \mathbf{s}_1) = \mu(Y|\mathbf{X}, \mathbf{s}_1) - \mu(Y|\mathbf{X}, \mathbf{s}_0) \neq 0$, where $\mathbf{s}_2$, $\mathbf{s}_1$, and $\mathbf{s}_0$ represent two, one, and zero minor alleles for SNPs,

respectively. We will investigate how the SNP–methylation–expression relationships affect the regression coefficients to be tested under the null (3.1). We consider six different conditions:

1. $\boldsymbol{\delta}_S \neq \mathbf{0}$ and $\alpha_M \boldsymbol{\alpha}_S \neq \mathbf{0}$: SNPs are associated with both methylation and gene expression (independent of methylation), and methylation is associated with gene expression.
2. $\boldsymbol{\delta}_S \alpha_M \neq \mathbf{0}$ and $\boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$: SNPs are associated with methylation, and methylation is associated with gene expression. SNPs are associated with gene expression only through methylation.
3. $\boldsymbol{\delta}_S \boldsymbol{\alpha}_S^{\mathrm{T}} \neq \mathbf{0}$, $\alpha_M = 0$ and $\boldsymbol{\alpha}_{SM} = \mathbf{0}$: SNPs are associated with both methylation and gene expression, but methylation is not associated with gene expression conditional on SNPs.
4. $\boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$, $\alpha_M = 0$ and $\boldsymbol{\delta}_S \neq \mathbf{0}$: SNPs are associated with methylation, but both SNPs and methylation are not associated with gene expression.
5. $\boldsymbol{\delta}_S = \mathbf{0}$ and $\boldsymbol{\alpha}_S \neq \mathbf{0}$: SNPs are associated with gene expression but not methylation.
6. $\boldsymbol{\delta}_S = \boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$: SNPs are not associated with methylation or gene expression.

With additional assumptions, we show in the supplementary material available at *Biostatistics* online that Conditions 1–6 correspond to causal diagrams (Robins, 2003) in Figures 1(a)–(f), respectively. Here, we would like to first discuss how these conditions have influences on the null hypothesis. We present in the

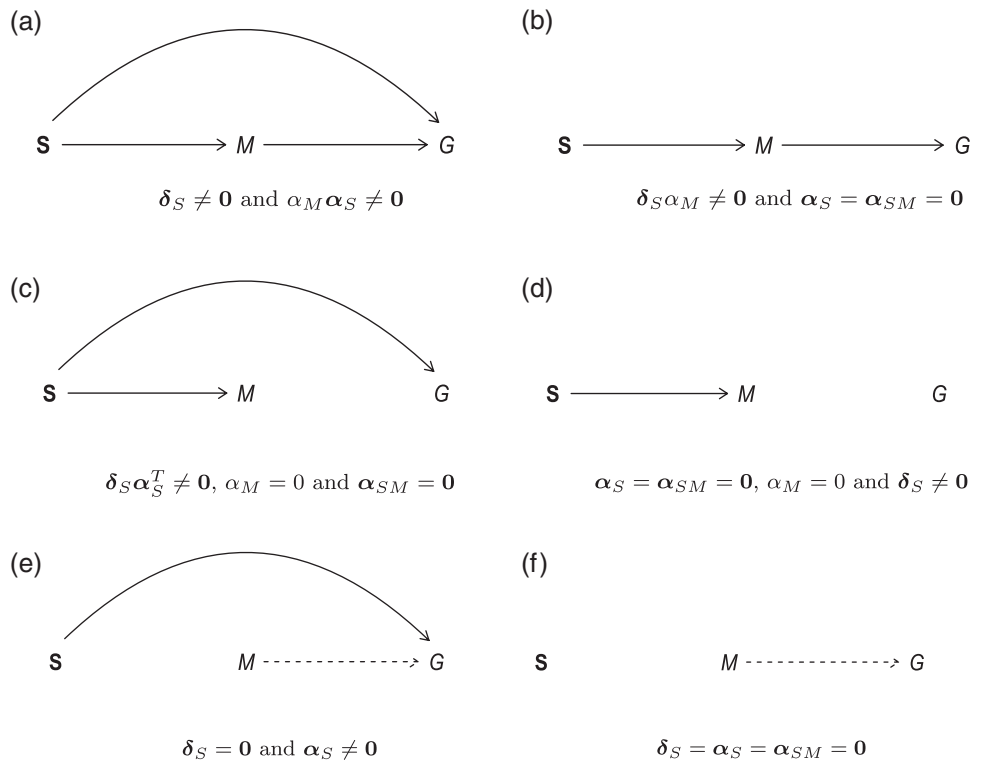


Fig. 1. Directed acyclic graphs of different causal relationships among SNPs **S**, methylation M, and gene expression G. The solid arrow from A to B indicates a non-zero effect of A on B; the dashed arrow indicates the effect may or may not exist; no arrow between A and B indicates no effect. Direct effect is the effect from **S** directly to G ($\mathbf{S} \rightarrow G$), and indirect effect is the effect of **S** on G mediated through M ($\mathbf{S} \rightarrow M \rightarrow G$).

following propositions the correspondence of the null (3.1) with different elements of $\boldsymbol{\beta}$, the coefficients in model (2.1) (proof is provided in the supplementary material available at *Biostatistics* online).

The propositions that we will present next require the following assumptions: there exists $\mathbf{b}$ with $\|\mathbf{b}\| > 0$ such that $\mathrm{E}[\mathcal{G}^{-1'}(\mathbf{b}^{\mathrm{T}}h(\mathbf{S}_i)K)K]$ $(K = M, G, MG)$, $\mathrm{E}[\mathcal{G}^{-1'}(\mathbf{b}^{\mathrm{T}}M)]$, $\mathrm{E}[\mathcal{G}^{-1'}(\mathbf{b}^{\mathrm{T}}G)]$, $\mathrm{E}[\mathcal{G}^{-1'}(\mathbf{b}^{\mathrm{T}}MG)M]$, and $\mathrm{E}[\mathcal{G}^{-1'}(\mathbf{b}^{\mathrm{T}}MG)G]$ are not zero; if two or more elements in $\boldsymbol{\beta}$ are not equal to zero, there does not exist a combination of the non-zero elements for all S such that $\Delta = 0$.

PROPOSITION 3.1 If Y, M, and G follow models (2.1–2.3), respectively, and any of the following: (i) $\boldsymbol{\delta}_S \neq \mathbf{0}$ and $\alpha_M\boldsymbol{\alpha}_S \neq \mathbf{0}$, (ii) $\boldsymbol{\delta}_S\alpha_M \neq \mathbf{0}$ and $\boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$, (iii) $\boldsymbol{\delta}_S\boldsymbol{\alpha}_S^{\mathrm{T}} \neq \mathbf{0}$, $\alpha_M = 0$, and $\boldsymbol{\alpha}_{SM} = \mathbf{0}$ (Conditions 1–3) holds, then $\Delta = 0 \Leftrightarrow \boldsymbol{\beta} = \mathbf{0}$.

In the next three propositions, we further show the influence of Conditions 4–6 on the correspondence of the null (3.1) with $\boldsymbol{\beta}$.

PROPOSITION 3.2 If Y, M, and G follow models (2.1–2.3), respectively, and $\boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$, $\alpha_M = 0$, and $\boldsymbol{\delta}_S \neq \mathbf{0}$ (Condition 4), then $\Delta = 0 \Leftrightarrow \boldsymbol{\beta}_{(-G)} = \mathbf{0}$, where $\boldsymbol{\beta}_{(-G)}$ denotes a vector containing all elements in $\boldsymbol{\beta}$ except β_G.

PROPOSITION 3.3 If Y, M, and G follow models (2.1–2.3), respectively, and $\boldsymbol{\delta}_S = \mathbf{0}$ and $\boldsymbol{\alpha}_S \neq \mathbf{0}$ (Condition 5), then $\Delta = 0 \Leftrightarrow \boldsymbol{\beta}_{(-M)} = \mathbf{0}$, where $\boldsymbol{\beta}_{(-M)}$ denotes a vector containing all elements in $\boldsymbol{\beta}$ except β_M.

PROPOSITION 3.4 If Y, M, and G follow models (2.1–2.3), respectively, and $\boldsymbol{\delta}_S = \boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$ (Condition 6), then $\Delta = 0 \Leftrightarrow \boldsymbol{\beta}_{(-M,-G,-MG)} = \mathbf{0}$ where $\boldsymbol{\beta}_{(-M,-G,-MG)}$ denotes a vector of all elements in $\boldsymbol{\beta}$ except β_M, β_G, and β_{MG}.

Thus, depending on different associations among $\mathbf{S}$, M, and G, we can evaluate $\Delta = 0$ in (3.1) by developing testing procedures for

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \tag{3.2}$$

$$H_0 : \boldsymbol{\beta}_{(-G)} = \mathbf{0}, \tag{3.3}$$

$$H_0 : \boldsymbol{\beta}_{(-M)} = \mathbf{0}, \tag{3.4}$$

and

$$H_0 : \boldsymbol{\beta}_{(-M,-G,-MG)} = \mathbf{0}. \tag{3.5}$$

4. TEST FOR THE TOTAL GENETIC EFFECT

Using the results in Section 3, we would like to construct a testing procedure for the null (3.2–3.5) using the asthma data where the asthma risk (yes/no) is modeled using logistic link with $h(\mathbf{S}) = h_M(\mathbf{S}) = h_G(\mathbf{S}) = \mathbf{S}$. Other types of outcome or different $\mathcal{G}$ can be easily adapted following a similar development. If the number of SNPs is small, we can perform conventional tests such as Wald test or likelihood ratio test (LRT) for null (3.2–3.5). However, if the number of SNPs (p) in a gene is large or some might be highly correlated due to linkage disequilibrium, the conventional test with a large degree of freedom (DF) has limited power (Huang *and others*, 2014).

4.1 *A score test for variance components*

To overcome the limitation of LRT, we assume a working distribution $\boldsymbol{\beta} \sim F_{\boldsymbol{\beta}}(0, \mathbf{D}(\boldsymbol{\tau}))$ where $\mathbf{D}(\boldsymbol{\tau}) = \text{diag}\{\text{rep}(\tau_S, p), \text{rep}(\tau_{MG}, 3), \text{rep}(\tau_{SM}, p), \text{rep}(\tau_{SG}, p), \text{rep}(\tau_{SMG}, p)\}$, $\text{rep}(A, B)$ indicates that A is repeated B times and $F_{\boldsymbol{\beta}}$ is an arbitrary distribution. Null (3.2) becomes equivalent to:

$$H_0 : \boldsymbol{\tau} = \mathbf{0}. \tag{4.1}$$

We will then construct a testing procedure based on this null hypothesis and other null hypotheses (3.3–3.5) can be viewed as special cases. By a Taylor series at $\boldsymbol{\beta} = \mathbf{0}$, the conditional log-quasilikelihood of model (2.4), l can be approximated as

$$l(\boldsymbol{\beta}) = \sum_i l_i(\mathbf{0}) + \sum_i \left.\frac{\partial l_i}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\mathbf{0}} \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}} \left(\sum_i \left.\frac{\partial l_i}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\mathbf{0}} \sum_i \left.\frac{\partial l_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=\mathbf{0}} + \sum_i \left.\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=\mathbf{0}} \right) \boldsymbol{\beta} + \epsilon,$$

where l_i is the conditional log-quasilikelihood for subject i (Lin, 1997). The marginal log-quasilikelihood can then be expressed with $\boldsymbol{\tau}$:

$$l(\boldsymbol{\tau}) = \mathrm{E}_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_i l_i(\mathbf{0}) + \frac{1}{2}\,\text{tr}\left(\mathbf{Z}^{\mathrm{T}} \left[\frac{\partial l}{\partial \eta}\frac{\partial l}{\partial \eta^{\mathrm{T}}} + \frac{\partial^2 l}{\partial \eta \partial \eta^{\mathrm{T}}} \right] \mathbf{Z}\mathbf{D}(\boldsymbol{\tau}) \right) + \epsilon,$$

where $\partial l/\partial \eta$ is an $n \times 1$ vector whose ith component is $\partial l_i/\partial \eta_i$, $\partial^2 l/\partial \eta \partial \eta^{\mathrm{T}} = \text{diag}\{\partial^2 l_i/\partial \eta_i^2\}$, $\mathbf{Z}^{\mathrm{T}} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, and $\mathbf{Z}_i^{\mathrm{T}} = (\partial \eta_i/\partial \boldsymbol{\beta}_S^{\mathrm{T}},\ \partial \eta_i/\partial \beta_M,\ \partial \eta_i/\partial \beta_G,\ \partial \eta_i/\partial \beta_{MG},\ \partial \eta_i/\partial \boldsymbol{\beta}_{SM}^{\mathrm{T}},\ \partial \eta_i/\partial \boldsymbol{\beta}_{SG}^{\mathrm{T}},\ \partial \eta_i/\partial \boldsymbol{\beta}_{SMG}^{\mathrm{T}})|_{\boldsymbol{\beta}=\mathbf{0}}$.

From (2.4), it can be shown that, evaluating at $\boldsymbol{\beta} = \mathbf{0}$,

$$\begin{aligned} &\frac{\partial \eta_i}{\partial \boldsymbol{\beta}_S} = \mathbf{S}_i, \quad \frac{\partial \eta_i}{\partial \boldsymbol{\beta}_M} = \mu_{Mi}, \quad \frac{\partial \eta_i}{\partial \boldsymbol{\beta}_G} = \mu_{Gi}, \quad \frac{\partial \eta_i}{\partial \boldsymbol{\beta}_{MG}} = \mu_{MGi}, \\ &\frac{\partial \eta_i}{\partial \boldsymbol{\beta}_{SM}} = \mu_{Mi}\mathbf{S}_i, \quad \frac{\partial \eta_i}{\partial \boldsymbol{\beta}_{SG}} = \mu_{Gi}\mathbf{S}_i, \quad \text{and} \quad \frac{\partial \eta_i}{\partial \boldsymbol{\beta}_{SMG}} = \mu_{MGi}\mathbf{S}_i, \end{aligned} \tag{4.2}$$

where $\mu_{Mi} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\delta}_0 + \mathbf{S}_i^{\mathrm{T}}\boldsymbol{\delta}_S$, $\mu_{Gi} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_0 + \mathbf{S}_i^{\mathrm{T}}\boldsymbol{\alpha}_S + \mu_{Mi}(\alpha_M + \mathbf{S}_i^{\mathrm{T}}\boldsymbol{\alpha}_{SM})$, and $\mu_{MGi} = \mu_{Mi}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_0 + \mathbf{S}_i^{\mathrm{T}}\boldsymbol{\alpha}_S) + (\alpha_M + \mathbf{S}_i^{\mathrm{T}}\boldsymbol{\alpha}_{SM})(\mu_{Mi}^2 + \sigma_M^2)$, and that the score, for each τ, U, follows a similar form $(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}\mathbb{K}(\mathbf{Y} - \boldsymbol{\mu}_0) - \text{tr}(\mathbb{K}\mathbf{W})$ with $\mathbb{K} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$, $\mathbf{C}_{MG}\mathbf{C}_{MG}^{\mathrm{T}}$, $\mathbf{C}_{SM}\mathbf{C}_{SM}^{\mathrm{T}}$, $\mathbf{C}_{SG}\mathbf{C}_{SG}^{\mathrm{T}}$, and $\mathbf{C}_{SMG}\mathbf{C}_{SMG}^{\mathrm{T}}$ for τ_S, τ_{MG}, τ_{SM}, τ_{SG}, and τ_{SMG}, respectively, where $\mathbf{W} = \text{diag}\{\mu_i(1 - \mu_i)\}$, $\mathbf{S}^{\mathrm{T}} = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$, $\mathbf{C}_{MG}^{\mathrm{T}} = (\mathbf{C}_{MG,1}, \ldots, \mathbf{C}_{MG,n})$, $\mathbf{C}_{SM}^{\mathrm{T}} = (\mathbf{C}_{SM,1}, \ldots, \mathbf{C}_{SM,n})$, $\mathbf{C}_{SG}^{\mathrm{T}} = (\mathbf{C}_{SG,1}, \ldots, \mathbf{C}_{SG,n})$, and $\mathbf{C}_{SMG}^{\mathrm{T}} = (\mathbf{C}_{SMG,1}, \ldots, \mathbf{C}_{SMG,n})$; $\mathbf{C}_{MG,i} = (\mu_{Mi}, \mu_{Gi}, \mu_{MGi})^{\mathrm{T}}$, $\mathbf{C}_{SM,i} = \mu_{Mi}\mathbf{S}_i$, $\mathbf{C}_{SG,i} = \mu_{Gi}\mathbf{S}_i$, and $\mathbf{C}_{SMG,i} = \mu_{MGi}\mathbf{S}_i$. Also, the corresponding information follows the form $I = \mathbf{1}^{\mathrm{T}}(\mathbb{K} \cdot \mathbf{H} \cdot \mathbb{K})\mathbf{1}$, where $\mathbf{A} \cdot \mathbf{B}$ denotes the component-wise multiplication of conformable matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{1}$ denotes a vector of ones, and the diagonal and off-diagonal elements of $\mathbf{H}$ are $h_{ii} = -4\mu_{0i}^4 + 8\mu_{0i}^3 - 5\mu_{0i}^2 + \mu_{0i}$ and $h_{ii'} = 2[\mu_{0i}(1 - \mu_{0i})][\mu_{0i'}(1 - \mu_{0i'})]$, respectively. We can estimate $\hat{\mu}_{0i} = \text{expit}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^{\mathrm{T}}\mathbf{X}_i)$ and $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ can be obtained from the logistic model under the null.

We can then construct the test statistics for $H_0 : \boldsymbol{\tau} = \mathbf{0}$ as the weighted sum of scores:

$$\begin{aligned} Q &= a_1 U_{\tau_S}^* + a_2 U_{\tau_{MG}}^* + a_3 U_{\tau_{SM}}^* + a_4 U_{\tau_{SG}}^* + a_5 U_{\tau_{SMG}}^* \\ &= (\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathrm{T}}(a_1 \mathbf{S}\mathbf{S}^{\mathrm{T}} + a_2 \mathbf{C}_{MG}\mathbf{C}_{MG}^{\mathrm{T}} + a_3 \mathbf{C}_{SM}\mathbf{C}_{SM}^{\mathrm{T}} + a_4 \mathbf{C}_{SG}\mathbf{C}_{SG}^{\mathrm{T}} + a_5 \mathbf{C}_{SMG}\mathbf{C}_{SMG}^{\mathrm{T}})(\mathbf{Y} - \boldsymbol{\mu}_0), \end{aligned}$$

where U^* is the non-zero-centered counterpart of U: $U^* = U + \text{tr}(\mathbb{K}\mathbf{W})$. Various weighting schemes (a_1-a_5) can be chosen. For example, $a_1 = \cdots = a_5$ correspond to $\mathbf{D}(\boldsymbol{\tau}) = \text{diag}\{\text{rep}(\tau, 4p+3)\}$. Or, we can choose to weight by their respective information $I_\cdot$: $a_1 = I_{\tau_S}^{-1/2}$, $a_2 = I_{\tau_{MG}}^{-1/2}$, $a_3 = I_{\tau_{SM}}^{-1/2}$, $a_4 = I_{\tau_{SG}}^{-1/2}$, $a_5 = I_{\tau_{SMG}}^{-1/2}$, which allows each score to have variance one and be comparable.

We are able to calculate a p-value for the statistic Q if its distribution can be obtained. Note that Q is a quadratic form of Y and, asymptotically, it follows a mixture of χ^2 distributions, which can be approximated with Davies' method by inverting the characteristic function (Davies, 1980). Alternatively, we can perform a resampling perturbation procedure based on the asymptotic distribution of Q (Parzen *and others*, 1994). We show in the supplementary material available at *Biostatistics* online that $Q \xrightarrow{D} \sum_l (\mathbf{A}_l \boldsymbol{\epsilon})^2$, where $\mathbf{D} = \left[\begin{smallmatrix} \mathbf{D}_{XX} & \mathbf{D}_{XV} \\ \mathbf{D}_{VX} & \mathbf{D}_{VV} \end{smallmatrix}\right] = n^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z}$, $\mathbf{V}_i^{\mathrm{T}} = (\sqrt{a_1}\mathbf{S}_i^{\mathrm{T}}, \sqrt{a_2}\mathbf{C}_{MG,i}^{\mathrm{T}}, \sqrt{a_3}\mathbf{C}_{SM,i}^{\mathrm{T}}, \sqrt{a_4}\mathbf{C}_{SG,i}^{\mathrm{T}}, \sqrt{a_5}\mathbf{C}_{SMG,i}^{\mathrm{T}})$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, $\mathbf{Z}_i = (\mathbf{X}_i^{\mathrm{T}}, \mathbf{V}_i^{\mathrm{T}})$, $\mathbf{A}_l$ is the lth row of $\mathbf{A} = [-\mathbf{D}_{XV}^{\mathrm{T}}\mathbf{D}_{XX}^{-1}, \mathbf{I}_{4p+3}]$.

4.2 *Tests for different models*

Note that the above derivation is based on the marginal model (2.4) and the results involve the expectation of M, G, and MG: μ_M, μ_G, and μ_{MG}. If we are able to collect $\mathbf{S}_i$, M, and G from the same subjects, we can also derive the score and information based on the joint model (2.1), which leads to very similar results, except that μ_M, μ_G, and μ_{MG} would be replaced by their observed counterparts M, G, and MG. Thus, we can carry out the proposed testing procedure if SNP, methylation, and expression data are collected from the same subjects. The advantage of the proposed method is that it can still be applied in the setting where methylation and/or expression data are not collected in the subjects of GWAS but their association with SNPs can be consistently estimated from external mQTL and eQTL studies.

In addition to model (2.1) where the main effects and interactive effects of SNPs, methylation and expression on outcome are assumed to exist, one can specify more parsimonious models. For example, if we assume that there is no three-way interaction, we can test $H_0 : \tau_S = \tau_{MG} = \tau_{SM} = \tau_{SG} = 0$. Different model specification depends on our assumption for the true disease model. Here for null (4.1), we consider six disease models: (1) SNPs-only ($H_0 : \tau_S = 0$); (2) main effects with possible methylation-by-expression interaction ($H_0 : \tau_S = \tau_{MG} = 0$); (3) (2) plus SNPs-by-methylation interaction ($H_0 : \tau_S = \tau_{MG} = \tau_{SM} = 0$); (4) (2) plus SNPs-by-expression interaction ($H_0 : \tau_S = \tau_{MG} = \tau_{SG} = 0$); (5) the union of (3) and (4) ($H_0 : \tau_S = \tau_{MG} = \tau_{SM} = \tau_{SG} = 0$); (6) all effects up to three-way interaction ($H_0 : \boldsymbol{\tau} = \mathbf{0}$). Although different parameters are tested under different model specification, they correspond to the same null (3.1) and are all valid tests.

4.3 *Omnibus test*

Since we do not know which one of the above six candidate models is the truth in reality, it is desirable to develop a test that can accommodate different models to maximize the power. Thus, we further propose an omnibus test where we identify the strongest evidence among the six models in Section 4.2. Specifically, we compute the minimum p-value among candidate models and compare the observed minimum p-value to its null distribution, approximated by a resampling perturbation procedure.

As shown in Section 4.1, Q converges in distribution to $Q(0) = \sum_l (\mathbf{A}_l^{\mathrm{T}}\boldsymbol{\epsilon})^2$. The empirical distribution of $Q(0)$ can be estimated using the perturbation (Parzen *and others*, 1994). Set $\hat{\boldsymbol{\epsilon}} = n^{-1/2}\sum_{i=1}^n \mathbf{Z}_i^{\mathrm{T}}(Y_i - \hat{\mu}_i)\mathcal{N}_i$, where $\mathcal{N}_i$'s are independent $N(0, 1)$ ($i = 1, \ldots, n$). By generating independent $\mathcal{N} = (\mathcal{N}_1, \ldots, \mathcal{N}_n)$ repeatedly, the perturbed realization of $Q(0)$ can be obtained, denoted by $\{\hat{Q}(0)^{(b)}, b = 1, \ldots, B\}$, where B is the number of perturbations. 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 the six candidate

models by inputting $\mathbf{V}_i$ with different combinations of $\mathbf{S}$, $\mathbf{C}_{MG}$, $\mathbf{C}_{SM}$, $\mathbf{C}_{SG}$, and $\mathbf{C}_{SMG}$, generating their perturbed realizations of the null counterpart for the candidate model k as $\{\hat{Q}_k(0)^{(b)}\}$, and comparing them with corresponding observed values Q_k ($k = 1, \ldots, 6$). Note that, for each perturbation b, the perturbation random variable $\mathcal{N}^{(b)}$ is the same across the six tests. Let $\hat{P}_k = \mathcal{S}_k(Q_k)$ be the p-value for the candidate model k, where $\mathcal{S}_k(q) = \text{pr}\{\hat{Q}_k(0)^{(b)} > q\}$. The null distribution of the minimum p-value, $\hat{P}_{\min} = \min_k \hat{P}_k$ can be approximated by $\hat{P}^{(b)}_{\min} = \min_k\{\mathcal{S}_k(\hat{Q}_k(0)^{(b)})\}$ ($b = 1, \ldots, B$). The omnibus p-value hence can be calculated by comparing $\hat{P}_{\min}$ with its empirical null distribution.

5. Numerical studies

5.1 *Settings*

To mimic the motivating data example of the asthma genetic study, we simulated the data based on the *ORMDL3* gene. We simulated 99 HapMap SNPs at the region where the *ORMDL3* gene is located using HAPGEN (Marchini *and others*, 2007). Nine out of the 99 HapMap SNPs are included in the Illumina HumanHap 300 K array, which are the so-called typed SNPs. We assumed two untyped SNPs ($\mathbf{S}^* = (S_{\text{causal1}}, S_{\text{causal2}})$) out of the 99 HapMap SNPs to be causal. The methylation, gene expression and disease outcome were generated using the two causal untyped SNPs, but the analyses were based on the 9 typed SNPs. The methylation is generated by the model: $M_i = 0 + \mathbf{S}_i^{*\text{T}}\boldsymbol{\delta}_S + \epsilon_{M,i}$, where $\epsilon_{M,i}$ follows a Beta($\alpha = 2, \beta = 5$) distribution that is further standardized to have mean zero and variance 1. The gene expression is generated by the model: $G_i = 0 + \mathbf{S}_i^{*\text{T}}\boldsymbol{\alpha}_S + M_i\alpha_M + M_i\mathbf{S}_i^{*\text{T}}\boldsymbol{\alpha}_{SM} + \epsilon_{G,i}$, where $\epsilon_{G,i} \sim N(0, \text{sd} = 0.4) + 0.5 \times U(-0.3, 0.3)$; U denotes a uniform distribution. The outcome Y was generated by a logistic regression model: $\text{logit}[P(Y_i = 1|\mathbf{S}_i^*, M_i, G_i)] = -1.2 + \mathbf{S}_i^{*\text{T}}\boldsymbol{\beta}_S + M_i\beta_M + G_i\beta_G + M_i\mathbf{S}_i^{*\text{T}}\boldsymbol{\beta}_{SM} + G_i\mathbf{S}_i^{*\text{T}}\boldsymbol{\beta}_{SG} + M_iG_i\beta_{MG} + M_iG_i\mathbf{S}_i^{*\text{T}}\boldsymbol{\beta}_{SMG}$. Note G and M follow arbitrary distributions to illustrate the flexibility of the proposed method.

For each simulation, we generated a cohort with 1000 subjects, from which we selected 150 cases and 150 controls for the genetic association study of the disease Y (case–control GWAS data) and selected another 300 subjects to study the association among $\mathbf{S}$, M, and G using models (2.2) and (2.3) (QTL data). As our primary interest is to study the genetic etiology of the disease, QTL data serve as an external source to study the relationship of $\mathbf{S}$, M, and G. We investigated the performance of our methods using the nine typed SNPs and the observed or estimated methylation and gene expression in case–control data, i.e. ($\mathbf{S}$, M, G, MG) or ($\mathbf{S}$, $\hat{\mu}_M$, $\hat{\mu}_G$, $\hat{\mu}_{MG}$), where $\hat{\mu}_{Mi} = \mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\delta}}_0 + \mathbf{S}_i^{\text{T}}\hat{\boldsymbol{\delta}}_S$, $\hat{\mu}_{Gi} = \mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\alpha}}_0 + \mathbf{S}_i^{\text{T}}\hat{\boldsymbol{\alpha}}_S + \hat{\mu}_{Mi}(\hat{\alpha}_M + \mathbf{S}_i^{\text{T}}\hat{\boldsymbol{\alpha}}_{SM})$, and $\hat{\mu}_{MGi} = \hat{\mu}_{Mi}(\mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\alpha}}_0 + \mathbf{S}_i^{\text{T}}\hat{\boldsymbol{\alpha}}_S) + (\hat{\alpha}_M + \mathbf{S}_i^{\text{T}}\hat{\boldsymbol{\alpha}}_{SM})(\hat{\mu}^2_{Mi} + \hat{\sigma}^2_M)$; $\hat{\boldsymbol{\delta}}_0, \hat{\boldsymbol{\delta}}_S, \hat{\boldsymbol{\alpha}}_S, \hat{\alpha}_M, \hat{\boldsymbol{\alpha}}_{SM}$, and $\hat{\sigma}^2_M$ are least squares estimates from the QTL data, and $\mathbf{X}_i$ and $\mathbf{S}_i$ are from the case–control GWAS data. By setting different configurations of δ's and α's, we were able to generate data according to different conditions illustrated in Figure 1. Different configurations of β's will be studied.

5.2 *Size and power*

For both observed and estimated M and G, the size of the test is well protected using either Davies' method or perturbation when the null hypotheses are correctly specified (Table 1). However, the tests are biased if we use the observed M and G test for the null (3.2): $\boldsymbol{\beta} = \mathbf{0}$ while the data are generated under Figures 1(d)–(f). But the type I errors are still protected if the estimated $\hat{\mu}_M$ and $\hat{\mu}_G$ are used, as explained in supplementary material available at *Biostatistics* online.

We also compared our proposed testing procedure with the conventional LRT and Fisher combination of marginal analyses for SNPs and gene expression (Fisher, 1925). The size of $p < 0.05$ is slightly inflated for the SNP-only analyses ($H_0 : \boldsymbol{\beta}_S = 0$), but as the number of parameters becomes large, the size is lower

Table 1. *The empirical size of the proposed test using perturbation*

	a	b	c	d	d	e*	e**	e*	e**	f*	f**	f*	f**
	$\boldsymbol{\beta}=\mathbf{0}$	$\boldsymbol{\beta}=\mathbf{0}$	$\boldsymbol{\beta}=\mathbf{0}$	$\boldsymbol{\beta}=\mathbf{0}$	$\boldsymbol{\beta}_{-G}=\mathbf{0}$	$\boldsymbol{\beta}=\mathbf{0}$		$\boldsymbol{\beta}_{-M}=\mathbf{0}$		$\boldsymbol{\beta}=\mathbf{0}$		$\boldsymbol{\beta}_{-M,-G,-MG}=\mathbf{0}$	
VCT: observed methylation M and gene expression G													
$\boldsymbol{\tau}_1:\tau_S$	5.10	5.10	5.10	5.00	4.70	4.50	4.50	5.15	5.15	4.85	4.85	5.05	4.60
$\boldsymbol{\tau}_2:\tau_S,\tau_{MG}$	4.80	4.75	4.90	100	4.55	100	100	4.45	4.60	100	100	5.05	4.60
$\boldsymbol{\tau}_3:\tau_S,\tau_{MG},\tau_{SM}$	4.50	4.85	4.30	100	4.60	100	100	4.20	4.30	100	100	5.05	4.15
$\boldsymbol{\tau}_4:\tau_S,\tau_{MG},\tau_{SG}$	4.85	4.70	4.90	100	4.60	100	100	4.00	4.10	100	100	5.05	4.25
$\boldsymbol{\tau}_5:\tau_S,\tau_{MG},\tau_{SM},\tau_{SG}$	4.65	4.60	4.95	100	4.40	100	100	3.95	4.15	100	100	5.00	4.10
$\boldsymbol{\tau}_6:\tau_S,\tau_{MG},\tau_{SM},\tau_{SG},\tau_{SMG}$	4.60	4.55	4.65	100	4.45	100	100	4.25	3.75	100	100	4.85	2.90
omb: omnibus	5.25	5.05	5.10	100	4.80	100	100	5.20	5.30	100	100	5.10	4.25
VCT: expected methylation μ_M and gene expression μ_G													
$\boldsymbol{\tau}_1:\tau_S$	5.10	5.10	5.10	5.00	4.10	4.50	4.50	3.55	3.55	4.85	4.85	4.30	4.20
$\boldsymbol{\tau}_2:\tau_S,\tau_{MG}$	5.40	5.40	5.25	4.90	2.75	4.40	4.35	3.60	3.55	4.15	3.90	4.30	4.20
$\boldsymbol{\tau}_3:\tau_S,\tau_{MG},\tau_{SM}$	5.70	5.35	5.55	4.70	3.50	4.85	4.75	3.20	3.05	4.60	4.35	4.45	4.20
$\boldsymbol{\tau}_4:\tau_S,\tau_{MG},\tau_{SG}$	5.40	5.00	5.60	4.40	2.65	4.55	4.70	3.60	3.55	4.45	4.55	4.20	4.55
$\boldsymbol{\tau}_5:\tau_S,\tau_{MG},\tau_{SM},\tau_{SG}$	5.35	5.15	5.40	4.00	2.90	4.45	4.35	3.10	3.30	4.65	5.20	4.60	4.60
$\boldsymbol{\tau}_6:\tau_S,\tau_{MG},\tau_{SM},\tau_{SG},\tau_{SMG}$	5.40	5.25	5.45	4.45	2.60	4.70	4.85	3.55	3.40	4.45	4.70	4.70	4.40
omb: omnibus	6.15	6.15	6.05	5.45	3.45	5.15	5.00	3.75	3.80	4.95	5.05	4.60	4.55
LRT: observed methylation M and gene expression G													
$\boldsymbol{\beta}_S$	6.15	6.15	6.15	7.15	7.15	6.40	6.40	6.40	6.40	6.50	6.50	7.40	7.40
$\boldsymbol{\beta}_S,\beta_M,\beta_G,\beta_{MG}$	6.75	7.05	7.20	100	6.40	100	100	6.45	6.70	100	100	5.85	5.30
$\boldsymbol{\beta}_S,\beta_M,\beta_G,\beta_{MG},\boldsymbol{\beta}_{SM}$	1.75	1.70	1.70	100	2.70	100	100	4.30	4.10	100	100	4.25	5.15
$\boldsymbol{\beta}_S,\beta_M,\beta_G,\beta_{MG},\boldsymbol{\beta}_{SG}$	1.45	1.40	1.25	100	4.30	100	100	2.20	2.40	100	100	4.55	4.40
All $\boldsymbol{\beta}$ except $\boldsymbol{\beta}_{SMG}$	0.90	0.65	0.70	100	2.45	100	100	2.55	2.45	100	100	2.95	3.85
All $\boldsymbol{\beta}$	0.55	0.35	0.40	100	1.75	100	100	1.70	1.85	100	100	2.75	3.70
Fisher's combination of univariate p-values: observed M and gene expression G													
$\boldsymbol{\beta}_{Sj},\beta_M,\beta_G$	15.70	15.30	15.80	100	17.35	100	100	15.90	31.35	100	100	15.25	15.85
$\boldsymbol{\beta}_S,\beta_M,\beta_G$	8.00	6.15	6.70	100	7.20	100	100	6.80	63.90	100	100	6.50	7.40

*$\alpha_M=0$; **$\alpha_M=0.3$.

The size is estimated based on 2000 simulations. The alphabet in columns indicates the part of Figure 1, by which the data are generated. The variance components specified in each row are those to be tested ($=0$).

than 5% (Table 1). The conservativeness due to the large DF is also observed in power (cf. Table S1 of supplementary material available at *Biostatistics* online with Table 2). The size for Fisher's combination is also inflated, which may be again due to the large DF for SNP-set analyses. Moreover, the independence assumption of Fisher's combination for SNPs, methylation and expression effects is obviously violated as SNPs are correlated to each other and to expression/methylation.

Table 2 presents the power when analyzing data generated under the causal diagram in Figure 1(a), and the results corresponding to Figures 1(b)–(f) are provided in the supplementary material available at *Biostatistics* online. In general, the tests can reach or almost reach the optimal performance when models are correctly specified, and the omnibus tests are very close to the optimal one. In Table 2, for example, under the setting of $\boldsymbol{\beta}_S^{\mathrm{T}} = (0.5, -0.25)$ and other β's equal zero, the test for $\tau_S = 0$ has the optimal power (68.8%) and the omnibus test has power of 69.1%; under the setting of $\boldsymbol{\beta}_S^{\mathrm{T}} = (0.1, -0.05)$, $\beta_M = \beta_G = \beta_{MG} = 0.2$, $\boldsymbol{\beta}_{SM}^{\mathrm{T}} = \boldsymbol{\beta}_{SG}^{\mathrm{T}} = (0.2, -0.1)$, $\boldsymbol{\beta}_{SMG} = \mathbf{0}$, the test for $\tau_S = \tau_{MG} = \tau_{SM} = \tau_{SG} = 0$ has the optimal power (87.0% and 60.3%) and the omnibus test has power of 82.4% and 56.4% for the observed and expected methylation/expression, respectively.

A few settings that reflect different interesting biology requires further attention. For example, we may observe SNPs to be eQTL and mQTL ($\boldsymbol{\delta}_S \neq \mathbf{0}$, $\boldsymbol{\alpha}_S \neq \mathbf{0}$) but expression and methylation do not affect the outcome ($\beta_M = \beta_G = 0$) if we measure the irrelevant tissue. Under this setting (the first row of Table 2), the joint analyses are subject to power loss compared to SNP-only analyses. In addition, the fourth row of the Table 2 indicates that SNPs have no direct effect on the disease outcome ($\boldsymbol{\beta}_S = \mathbf{0}$) and their effect is only through gene expression and methylation ($\beta_M \neq 0$, $\beta_G \neq 0$) [again, the gene expression and methylations are affected by the SNPs ($\boldsymbol{\delta}_S \neq \mathbf{0}$, $\boldsymbol{\alpha}_S \neq \mathbf{0}$) in Figure 1]. The second and third rows are the special cases. Under these settings, the SNP-only analysis does not perform well and the joint analyses perform much better. Also note that model (2.1), we start with is a very general model that can reflect different biological mechanisms with different parameter configurations.

The tests using the estimated $\hat{\mu}_M$ and $\hat{\mu}_G$ have power loss as compared to those using the observed M and G. The power loss between observed and expected methylation and gene expression depends on how well they are associated with SNPs. In Figures 1(a) and (c), SNPs are good determinants of methylation and gene expression, so the power loss from the observed ones is less than that in Figure 1(b) where gene expression can only be determined by SNPs through methylation, which also needs to be estimated.

We also study the performance of our method with only SNPs and expression data or only SNPs and methylation data when the true model depends on all three of them (Table S7 of supplementary material available at *Biostatistics* online). Without including either methylation or expression, the type I error is well protected. The tests not including methylation lose power when methylation indeed has an effect on the outcome by comparing the results between Table 2 and Table S7 of supplementary material available at *Biostatistics* online. However, if only SNPs and gene expression but not methylation affect the outcome, then ignoring methylation performs better than the joint analyses of the three. Similar results for the setting without gene expression.

6. DATA APPLICATIONS

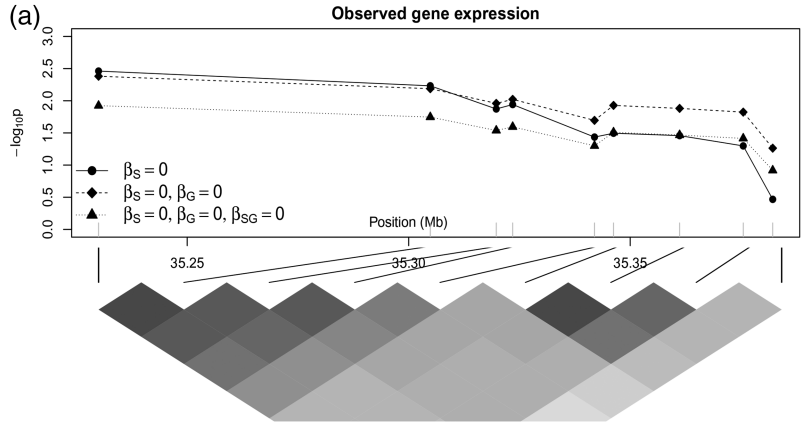
6.1 *ORMDL3 gene and asthma risk*

We demonstrate the utility of the theoretical results and the proposed testing procedure in single-SNP analyses of the *ORMDL3* gene (Figure 2), SNP-set analyses of *ORMDL3* (Table 3) and genome-wide SNP-set analyses of MRCA data (Figure S1 of supplementary material available at *Biostatistics* online) to investigate the risk of childhood asthma (Dixon *and others*, 2007; Moffatt *and others*, 2007). We used another dataset, the MRCE data to study the association between SNPs and expression of the *ORMDL3* gene. The MRCA dataset actually also collected gene expression data, so we can compare the results

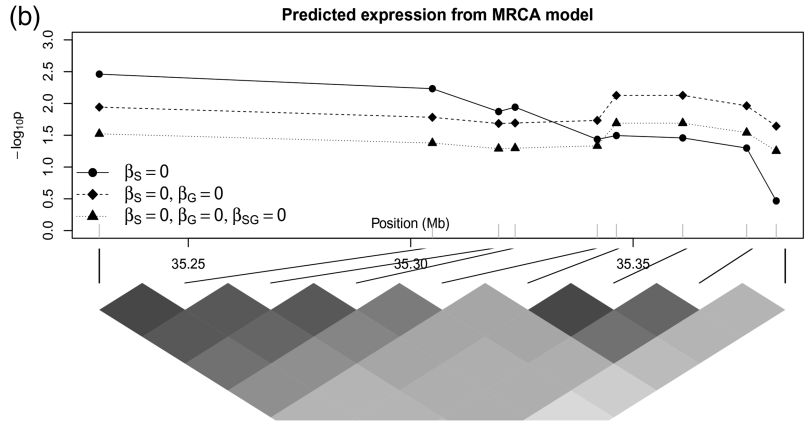
Table 2. *The empirical power of the proposed tests and competitors under the scenario of Figure 1a*

$\boldsymbol{\beta}_S$	β_M	β_G	β_{MG}	$\boldsymbol{\beta}_{SM}$	$\boldsymbol{\beta}_{SG}$	$\boldsymbol{\beta}_{SMG}$	$\boldsymbol{\tau}_1$	$\boldsymbol{\tau}_2$	$\boldsymbol{\tau}_3$	$\boldsymbol{\tau}_4$	$\boldsymbol{\tau}_5$	$\boldsymbol{\tau}_6$	omb	$\boldsymbol{\tau}_{KD}$	2-Stage	$\boldsymbol{\tau}_{1w}$	Fisher
(0.5, −0.25)	0	0	0	(0, 0)	(0, 0)	(0, 0)	68.8	68.6	55.3	58.0	49.4	48.2	69.1	58.6	68.8	31.4	63.3
							68.8	70.4	66.4	64.5	62.5	60.3	74.1	65.5			
(0, 0)	0.4	0	0	(0, 0)	(0, 0)	(0, 0)	24.9	83.6	90.6	83.2	89.2	87.5	86.1	83.5	24.9	30.8	90.6
							24.9	56.9	59.6	58.7	59.2	57.7	52.9	52.4			
(0, 0)	0	0.4	0	(0, 0)	(0, 0)	(0, 0)	35.2	83.2	83.4	89.5	87.8	85.4	84.7	85.1	35.2	45.2	89.0
							35.2	74.4	75.4	75.2	75.4	73.8	72.0	68.5			
(0, 0)	0.1	0.3	0	(0, 0)	(0, 0)	(0, 0)	32.4	83.2	84.5	88.2	88.1	85.4	84.1	83.1	32.4	41.3	89.1
							32.4	70.9	70.6	72.0	70.8	69.9	67.7	65.1			
(0.2, −0.1)	0.2	0.2	0	(0, 0)	(0, 0)	(0, 0)	49.1	73.4	75.2	74.8	75.6	74.2	73.6	72.2	49.1	52.0	77.8
							49.1	59.6	58.6	58.7	58.0	54.8	58.4	51.7			
(0, 0)	0.1	0.2	0.4	(0, 0)	(0, 0)	(0, 0)	18.4	97.1	97.3	97.4	97.6	99.6	98.7	99.1	18.4	25.8	74.6
							18.4	63.5	65.5	66.6	66.6	74.2	67.3	69.3			
(0.1, −0.05)	0.2	0.2	0.2	(0, 0)	(0, 0)	(0, 0)	40.8	92.8	94.5	94.0	94.8	96.2	94.0	95.5	40.8	47.4	89.3
							40.8	73.3	73.5	74.0	74.1	75.3	71.8	71.7			
(0.1, −0.05)	0.2	0.2	0.2	(0.2, −0.1)	(0.2, −0.1)	(0, 0)	41.7	73.9	84.3	81.9	87.0	85.2	82.4	84.9	41.7	44.4	81.3
							41.7	53.9	58.0	58.6	60.3	58.3	56.4	58.4			
(0.1, −0.05)	0.2	0.2	0.2	(0.2, −0.1)	(0.2, −0.1)	(0.05, −0.025)	42.9	67.5	78.3	74.8	80.2	78.1	74.8	80.6	42.9	42.8	75.7
							42.9	47.1	52.2	53.7	53.7	51.5	52.0	53.8			

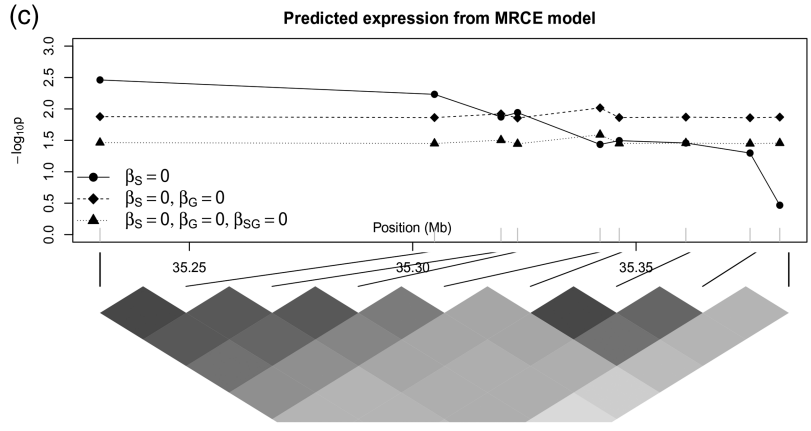
The power is estimated based on 1000 simulations. For each configuration, the upper row is the power (%) using the observed M and G and the lower row is that using the predicted ones. Here $\boldsymbol{\tau}_1$ is the test for $\tau_S = 0$; $\boldsymbol{\tau}_2$ is the test for $\tau_S = \tau_{MG} = 0$; $\boldsymbol{\tau}_3$ is the test for $\tau_S = \tau_{MG} = \tau_{SM} = 0$; $\boldsymbol{\tau}_4$ is the test for $\tau_S = \tau_{MG} = \tau_{SG} = 0$; $\boldsymbol{\tau}_5$ is the test for $\tau_S = \tau_{MG} = \tau_{SM} = \tau_{SG} = 0$; $\boldsymbol{\tau}_6$ is the test for $\boldsymbol{\tau} = \mathbf{0}$; omb is the omnibus test for the above tests; $\boldsymbol{\tau}_{KD}$ is the test for $\tau_S = \tau_{MG} = \tau_{SM} = \tau_{SG} = \tau_{SMG} = 0$, the variance component test implementing the idea proposed by (Kraft *and others*, 2007; Dai *and others*, 2012); 2-stage is the test that requires p-values < 0.05 in both $H_0 : \tau_S = 0$, and either $H_0 : \boldsymbol{\delta}_S = \mathbf{0}$ or $H_0 : \boldsymbol{\alpha}_S = \boldsymbol{\alpha}_{SM} = \mathbf{0}$; $\boldsymbol{\tau}_{1w}$ is an SNP-only test with the SNP effect on disease weighted by its eQTL and mQTL effects; Fisher, the Fisher's combination of p-values of marginal effects of SNP, methylation, and expression. The left part of the table is the configuration of β in model (2.1) used to simulate the data.



Observed gene expression G



Expected expression $\mu_{G,MRCA}$, using MRCA eQTL model



Expected expression $\mu_{G,MRCE}$, using MRCE eQTL model

Fig. 2. Single SNP analyses of the 9 typed SNPs at *ORMDL3*. The lower panel is the linkage disequilibrium plot where the darker indicates higher correlation between SNPs. (a) Observed gene expression G. (b) Expected expression $\mu_{G,\mathrm{MRCA}}$, using the MRCA eQTL model. (c) Expected expression $\mu_{G,\mathrm{MRCE}}$, using the MRCE eQTL model.

Table 3. *p-values of re-analyzing the 9 SNPs at ORMDL3 gene with MRCA data*

	Dominant			Additive		
	G	$\mu_{G,\text{MRCA}}$	$\mu_{G,\text{MRCE}}$	G	$\mu_{G,\text{MRCA}}$	$\mu_{G,\text{MRCE}}$
$\tau_S = 0$		0.0036			0.0108	
$\tau_S = \tau_G = 0$	0.0015	0.0015	0.0015	0.0045	0.0043	0.0049
$\tau_S = \tau_G = \tau_{SG} = 0$	0.0062	0.0028	0.0018	0.0119	0.0050	0.0065
Omnibus	0.0028	0.0028	0.0022	0.0088	0.0076	0.0063
$\tau_S = \tau_{SG} = 0$	0.0137	0.0015	0.0024	0.0451	0.0054	0.0112
$\tau_S = 0$ with eQTL weighting		0.0039			0.0110	
Fisher: SNP, expression		0.0140			0.0749	

G: joint analysis with observed expression of the *ORMDL3* gene; $\mu_{G,\text{MRCA}}$: analysis with the mean expression estimated from the model developed using the MRCA data; $\mu_{G,\text{MRCE}}$: analysis with the mean expression estimated from the model developed using the MRCE data. The omnibus test chooses the optimal test among $H_0 : \tau_S = 0$, $H_0 : \tau_S = \tau_G = 0$, and $H_0 : \tau_S = \tau_G = \tau_{SG} = 0$. Fisher, the Fisher combination of p-values of marginal effects of SNP and expression.

between using the observed gene expression and using the expected expression from the MRCE and MRCA data. The MRCA data contain 219 asthma cases and 99 controls. Genotype data were collected with the Illumina 300 K chip and gene expression was collected using the Affymetrix U133 Plus 2.0 in the MRCA data and Illumina Human6V1 in the MRCE data. Gene expression data were collected from an EBV-transformed lymphoblastoid cell line from study subjects and normalized with the RMA algorithm for Affymetrix array and quantile normalization for Illumina array (Liang *and others*, 2013).

There are 9 SNPs at the *ORMDL3* gene found to be highly associated with asthma risk in MRCA and also genotyped in MRCE data. We first model the association between the *ORMDL3* expression and the nine SNPs (i.e. eQTL model) in the MRCE data ($n = 487$) using the weighted least squares estimator. Because the MRCE data are case–control data designed for studying eczema, to obtain unbiased estimates, we need to reweight the case and control by π/d and $(1-\pi)/(1-d)$, respectively, where π is the prevalence of eczema and d is the proportion of eczema cases in the MRCE data. Since expression data were actually collected in the MRCA data, we can also evaluate the SNP-expression association in MRCA ($n = 318$). The association is highly significant in both datasets: $p < 2.20 \times 10^{-16}$ ($R^2 = 0.30$) in MRCE; $p = 5.16 \times 10^{-10}$ ($R^2 = 0.19$) in MRCA. Since the SNPs and expression are highly associated, we should evaluate the $H_0 : \tau_S = \tau_G = \tau_{SG} = 0$, the equivalent of null (4.1) for two genomic data. *ORMDL3* is differentially expressed in cases and controls ($p = 0.0085$). With these two eQTL models, we predict the gene expression in MRCA using the 9 SNPs, denoted as $\mu_{G,\text{MRCE}}$ and $\mu_{G,\text{MRCA}}$. To study the SNP effects of the 9 SNPs on asthma risk in the MRCA data, we perform joint analysis of SNPs at *ORMDL3* and its gene expression, including the observed expression G, the expected expression using eQTL models of MRCA ($\mu_{G,\text{MRCA}}$) and MRCE ($\mu_{G,\text{MRCE}}$).

For single-SNP analyses, inclusion of gene expression using LRT provides smaller p-values in many SNPs compared to SNP-only analyses (Figure 2). For multi-SNP analyses of *ORMDL3*, we applied our proposed score test for variance components. As shown in Table 3, joint analyses of SNPs and expression yield more significant results than SNP-only analyses in both additive and dominant models. Gene expression that is actually observed, estimated internally or externally can all improve the significance level. The analyses focusing on main effects of SNPs and gene expression provide the most significant results across different settings and the omnibus test can almost approach them.

We further perform a genome-wide analyses of the entire MRCA data (Figure S1 of supplementary material available at *Biostatistics* online). We first choose the SNP-expression pairs with false discovery rate (FDR) <1% from the *cis*-eQTL results (Liang *and others*, 2013). To illustrate the utility of combining different studies, we estimate the gene expression using the *cis*-SNP set, and for each gene, we group the *cis*-SNP set and its estimated gene expression and apply our proposed procedure to investigate its effect on the gene level. Our proposed omnibus test identifies 25 genes that are highly associated with asthma risk with FDR < 10%, whereas the SNP-only analyses without eQTL weighting identify 5 genes. The omnibus test from the joint analyses outperforms the SNP-only approach even with different cutoffs: for FDR <5%, 8 genes are identified in omnibus tests and 3 genes in SNP-only tests; for FDR <15%, 35 and 18 genes, respectively; with Bonferroni correction, 2 and 1 genes, respectively.

6.2 *GRB10 methylation and mortality of glioblastoma multiforme*

Here, we would like to illustrate the utility of our method beyond GWAS. Glioblastoma multiforme (GBM) is the most common malignant brain tumor that is rapidly fatal with a median survival time between 12.1 and 14.6 months (Stupp *and others*, 2005). It is thus important to identify genes that may be associated with its poor prognosis. Multiple genomic data of GBM as well as its survival information have been archived in The Cancer Genome Atlas. We integrate DNA methylation, micro-RNA, and gene expression data to jointly model the survival of 271 GBM patients. The survival is dichotomized at 390 days (the median survival). From our unpublished analyses of methylation and gene expression, we have found that the *GRB10* gene is significantly associated with GBM survival (Smith *and others*, 2014). Here, we combine 12 methylation loci within *GRB10* and its expression value and micro-RNA miR-633 expression to perform a gene-based analysis. Based on our statistical analyses of *GRB10* and miR-633, we set up a model as in Figure 1a with S, M, and G being 12 methylation loci of *GRB10*, miR-633 expression and *GRB10* expression, respectively.

The joint effect of *GRB10* on GBM survival under the main effect model ($p = 0.004$) is more significant than models with only methylation ($p = 0.146$) or with higher-order interactions ($p = 0.038$ and 0.054 with 2-way and 3-way, respectively). Owing to the non-convergence issue from the multi-locus analyses using LRT, we compare our methods with single-locus analyses where we calculate the permutation-adjusted minimum p-values from LRT. The p-values of our proposed omnibus test for variance components are more significant than the omnibus p-value from permutation-adjusted single-locus analyses ($p = 0.006$ vs. 0.026). We conclude that *GRB10* methylation has a significant effect on GBM survival, which may be through miR-633 and/or *GRB10* expression.

7. DISCUSSION

This paper has two major contributions. First, we propose an integrative approach to model genetic effect on clinical outcome. In genetic association studies, SNPs and the disease status are collected, but not gene expression/methylation, and in QTL studies, SNPs and gene expression/methylation are collected, but not the disease status. Here we develop a method that can integrate (1) multiple genomic data (e.g., SNPs, methylation and gene expression) and (2) different studies (e.g. GWAS, eQTL, and mQTL studies) to investigate genetic etiology for complex diseases. Secondly, we characterize all possible relationships among multiple genomic measures and investigate its correspondence to the regression parameters under the null. We further develop an efficient testing procedure that accounts for multiple correlated genetic markers and accommodates different underlying disease models.

Both methylation and gene expression are tissue-specific, but most GWASs only collect blood samples. Thus, for most GWASs, it may be difficult to obtain DNA and RNA samples from the ideal target tissue

for methylation and expression. The advantage of the proposed method is that as long as we are able to obtain consistent estimates of the QTL association parameters from other studies to estimate μ_G and μ_M, we can still perform the joint analysis. However, we rely on the assumption that the two studies are randomly sampled from a common base population, which needs to be carefully evaluated when assembling different studies.

The inclusion of SNP by gene expression or SNP by methylation interaction is biologically plausible. For example, single nucleotide change of an oncogene can lead to a detrimental mutation that has a synergistic effect from both undue biological consequences of the gene product and its uncontrolled expression (Carlo, 2008): the combination of the aberrant gene product due to the nucleotide change (i.e. mutation) and its high expression would lead to uncontrolled cell growth, which may not occur if only either one condition exists.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The author would like to thank the editor, the associate editor, and the referees for their helpful comments. The author thanks Professors Tyler VanderWeele and Joseph Hogan for helpful discussion. *Conflict of Interest*: None declared.

Funding

The work was supported by Richard B. Salomon Award Fund, Brown University.

References

Berlivet, S., Moussette, S., Ouimet, M., Verlaan, D. J., Koka, V., Tuwaijrl, A. A., Kwan, T., Sinnett, D., Pastinen, T. and Naumova, A. K. (2012). Interaction between genetic and epigenetic variation defines gene expression patterns at the asthma-associated locus 17q12–q21 in lymphoblastoid cell lines. *Human Genetics* **131**, 1161–1171.

Carlo, C. (2008). Oncogene and cancer. *New England Journal of Medicine* **358**, 502–511.

Dai, J. Y., Logsdon, B. A., Huang, Y., Hsu, L., Reiner, A. P., Prentice, R. L. and Kooperberg, C. (2012). Simultaneously testing for marginal genetic association and gene-environment interaction. *American Journal of Epidemiology* **176**, 164–173.

Davies, R. (1980). Algorithm AS 155: the distribution of a linear combination of χ^2 random variables. *Applied Statistics* **29**, 323–333.

Dixon, A., Liang, L., Moffatt, M., Chen, W., Heath, S., Wong, K., Taylor, J., Burnett, E., Gut, I., Farrall, M. *and others* (2007). A genome-wide association study of global gene expression. *Nature Genetics* **39**, 1202–1207.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Huang, Y. T., VanderWeele, T. J. and Lin, X. (2014). Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Annals of Applied Statistics* (in press).

Kraft, P., Yen, Y.-C., Stram, D. O., Morrison, J. and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63**, 111–119.

LIANG, L., MORAR, N., DIXON, A. L., LATHROP, G. M., ABECASIS, G. R., MOFFATT, M. F. AND COOKSON, W. O. C. (2013) A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research* **23**, 716–726.

LIN, X. (1997). Variance component test in generalised linear models with random effects. *Biometrika* **84**, 309–326.

MACKINNON, D. (2008). *Introduction to Statistical Mediation Analysis*. New York: Taylor & Francis.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. AND DONNELLY, P. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics* **39**, 906–913

MOFFATT, M., KABESCH, M., LIANG, L., DIXON, A., STRACHAN, D., HEATH, S., DEPNER, M., VON BERG, A., BUFE, A., RIETSCHEL, E. *and others* (2007). Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473.

MORLEY, M., MOLONY, C., WEBER, T., DEVLIN, J., EWENS, K., SPIELMAN, R. AND CHEUNG, V. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.

PARZEN, M., WEI, L. AND YING, Z. (1994). A resampling method based on pivotal functions. *Biometrika* **81**, 341–350.

PRENTICE, R. L. AND PYKE, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66**, 403–411.

ROBINS, J. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P., Hjort, N. L. and Richardson, S. (editors), *Highly Structured Stochastic Systems*. New York, NY: Oxford University Press, pp. 70–81.

ROBINS, J. AND GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155.

SCHADT, E., LAMB, J., YANG, X., ZHU, J., EDWARDS, S., GUHATHAKURTA, D., SIEBERTS, S. K., MONKS, S., REITMAN, M., ZHANG, C. *and others* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717.

SMITH, A. A., HUANG, Y.-T., ELIOT, M., HOUSEMAN, A., MARSIT, C. J., WIENCKE, J. K. AND KELSEY, K. T. (2014). A novel approach to the discovery of survival biomarkers in glioma using a joint analysis of DNA methylation and gene expression. *Epigenetics* **9**, (in press).

STUPP, R., MASON, W. P., VAN DEN BENT, M. J., WELLER, M., FISHER, B., TAPHOORN, M. J., BELANGER, K., BRANDES, A. A., MAROSI, C., BOGDAHN, U. *and others* (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine* **352**, 987–996.

WU, M., KRAFT, P., EPSTEIN, M., TAYLOR, D., CHANOCK, S., HUNTER, D. AND LIN, X. (2010). Powerful SNP set analysis for case–control genome-wide association studies. *American Journal of Human Genetics* **86**, 929–942.

XIONG, Q., ANCONA, N., HAUSER, E. R., MUKHERJEE, S. AND FUREY, T. S. (2012). Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Research* **22**, 386–397.

ZHU, J., ZHANG, B., SMITH, E. N., DREES, B., BREM, R. B., KRUGLYAK, L., BUMGARNER, R. E. AND SCHADT, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics* **40**, 854–861.

[*Received October 8, 2013; revised December 18, 2013; accepted for publication February, 24 2014*]