HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2023 August 01.

Published in final edited form as:

Stat Med. 2016 July 20; 35(16): 2831–2844. doi:10.1002/sim.6900.

Integrative genomic testing of cancer survival using semiparametric linear transformation models

Yen-Tsung Huang[*], **Tianxi Cai**[†], **Eunhee Kim**[‡]

[*]Departments of Epidemiology and Biostatistics, Brown University, 121 South Main St., Box G-S121-2, Providence, RI 02912

[†]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115

[‡]Office of Biostatistics, National Institute of Neurological Disorders and Stroke, National Institutes of Health, 10 Center Drive, Building 10/Rm 5N230, Bethesda MD 20892

Abstract

The wide availability of multi-dimensional genomic data has spurred increasing interests in integrating multi-platform genomic data. Integrative analysis of cancer genome landscape can potentially lead to deeper understanding of the biological process of cancer. We integrate epigenetics (DNA methylation and microRNA expression) and gene expression data in tumor genome to delineate the association between different aspects of the biological processes and brain tumor survival. To model the association, we employ a flexible semi-parametric linear transformation model that incorporates both the main effects of these genomic measures as well as the possible interactions among them. We develop variance component tests to examine different coordinated effects by testing various subsets of model coefficients for the genomic markers. A Monte-Carlo perturbation procedure is constructed to estimate the null distribution of the proposed test statistics. We further propose omnibus testing procedures to synthesize information from fitting various parsimonious sub-models to improve power. Simulation results suggest that our proposed testing procedures maintain proper size under the null and outperform standard score tests. We further illustrate the utility of our procedure in two genomic analyses for survival of glioblastoma multiforme patients.

Keywords

integrative genomics; linear transformation model; survival analysis; variance component test

1 Introduction

With advances in high-throughput biotechnology, genomic studies with a wide range of platforms have been performed to identify disease susceptibility loci or biomarkers for various phenotypic traits. Successful examples include gene expression microarray studies, genomewide association studies (GWAS) and epigenome-wide association studies (EWAS).

Yen-Tsung_Huang@brown.edu .

Despite the success of existing single-platform based studies, significant amount of genomic information is lost if one focuses only on a single platform. A new hypothesis has been advocated that the biological process of complex phenotypic traits such as cancer survival can be better characterized by multiple types of genetic, epigenetic and genomic alterations, and each platform provides a different and complementary view of the phenotype [1, 2].

This paper is motivated by The Cancer Genome Atlas (TCGA), a research project with a rich collection of multiplatform genomic data to map the tumor genomes in many types of cancers. We focus on a genomic study of glioblastoma multiforme (GBM), in which the association between DNA methylation and gene expression profile in the *GRB10* gene and the overall survival of GBM patients was reported [3]. It was also established that *GRB10* gene is the target of microRNA, miR-633 [4]. Both DNA methylation and microRNA consist of epigenetic regulation of gene expression and have been found to be associated with gene expression [3, 5]. The example suggests that multiple genomic data are interrelated, e.g., DNA methylation-microRNA-gene expression and may jointly affect cancer survival, illustrated as a causal diagram [6] in Figure 1. We are interested in 1) the effect of DNA methylation of *GRB10* gene on GBM survival mediated through mRNA expression of the gene (the dashed path in Figure 1), 2) the effect of DNA methylation mediated through microRNA expression (the solid path), and 3) the effect of DNA methylation on cancer survival independent of mRNA or microRNA expressions and perhaps through other biological mechanisms (the dotted path).

Hypothesis testing methods of multiple genetic markers on the survival outcome have been developed [7, 8]. These methods largely focus on a single genomic platform such as genetic markers. Moreover, these methods examine the overall effect and are not able to decompose the overall effect into separate components, as illustrated in Figure 1. With the rich collection of tumor genomic data such as TCGA, there has been a pressing need of analyzing multiplatform genomic data to understand their respective contribution to cancer survival. Statistical methods have been proposed under the mediation framework [9, 10, 11, 12] to integrate multiplatform genomic data where the outcome is dichotomous [13, 14]. It has also been shown that the three pathways illustrated in Figure 1 correspond to different sets of coefficients in regression models, and a hypothesis testing method has been developed to examine their effect on dichotomous outcomes [15]. However, the current integrated methods are not able to analyze the time-to-event data due to the challenge of censoring and require additional development prior to applying to the TCGA data. To bridge those gaps, we develop in this paper a new testing procedure for survival data that integrates multi-platform genomic data.

Cox proportional hazards (PH) model is the most popular model for analyzing survival data [16, 17]. Efficient estimation and testing procedures have been developed under the PH model [18]. However, since the PH assumption may be violated in real applications, alternative survival models such as proportional odds (PO) model [19] can be useful for such applications. Both the PH and PO models are special cases of a broader class of linear transformation models, which relates a nonparametric transformation of the failure time to covariates and a parametric random error in a linear form [18]. Various estimating procedures have been proposed for linear transformation models [20, 21], and Zeng and

Lin further proposed a non-parametric maximum likelihood estimator for a more general setting [22]. As most existing work focused primarily on the estimation problem, Tzeng *et al.* recently proposed an efficient testing procedure to examine effects of multiple genetic markers [8]. Although Tzeng's method also concerns multivariate testing, their method, however, is not readily applicable to our motivating example. First, it is not clear how to use the existing method to analyze multi-platforms genomic data. It has been shown that the single platform method is subject to power loss as it fails to account for signals from other platforms [13]. Second, by focusing on the overall effect, the current method is not able to examine specific effects illustrated in Figure 1. Third, it is not clear how to balance between robustness against model misspecification and statistical power while incorporating potential interactions among various platforms. To address these limitations, we propose a testing procedure based on estimating equations that extends Tzeng *et al.*'s work to integrative genomics.

The rest of the paper is organized as follows. In Section 2.2, we introduce a semiparametric linear transformation model for DNA methylation, microRNA and gene expression jointly on failure time, and propose a variance component score testing procedure for an arbitrary set of regression coefficients. We also construct an omnibus test to accommodate different underlying disease models. In Section 3, we provide mechanistic interpretation for various subsets of coefficients in the joint survival model. In Section 4, we conduct numerical studies to examine path-specific effects. In Section 5, we illustrate the utility of our methods with two data applications. We conclude with discussion in Section 6.

2 A multivariate test for the transformation model

2.1 The model

Our overall goal is to understand whether and how a survival time T depends on a p dimensional DNA methylation markers $\mathbf{S}$ within a gene, a microRNA expression M, and a gene expression G, after adjusting for a q dimensional vector of covariates $\mathbf{X}$. We assume a fixed number of DNA methylation markers p, but p may not be small relative to the sample size n in a finite sample. Due to censoring, T is only observable up to a bivariate vector (T^*, δ), where $T^* = \min(T, C)$, $\delta = I(T \leq C)$ and C is the censoring time. Suppose data for analysis consists of n independent and identically distributed random vectors $\{(T_i^*, \delta_i, \mathbf{Z}_i^\top), i = 1, \ldots, n\}$, where i indexes subjects and $\mathbf{Z}_i = (G_i, M_iG_i, G_i\mathbf{S}_i^\top, M_iG_i\mathbf{S}_i^\top, \mathbf{X}_i^\top, \mathbf{S}_i^\top, M_i, M_i\mathbf{S}_i^\top)^\top$.

We model the relationship through a flexible semi-parametric transformation model allowing for interactions among $\mathbf{S}$, M and G:

$$H^*(T_i) = -\boldsymbol{\gamma}^\top \mathbf{Z}_i + \epsilon_i^*, \quad \epsilon_i^* \perp \mathbf{Z}_i \tag{1}$$

where $\boldsymbol{\gamma} = (\beta_G, \beta_{MG}, \boldsymbol{\beta}_{SG}^\top, \boldsymbol{\beta}_{SMG}^\top, \boldsymbol{\beta}_X^\top, \boldsymbol{\beta}_S^\top, \beta_M, \boldsymbol{\beta}_{SM}^\top)^\top$ is the unknown regression parameters representing the effects of the covariates, genomic markers along with their interactions, ϵ_i^* has a specified parametric distribution, and $H^*(\cdot)$ is an unspecified strictly increasing smooth transformation function. The advantage of our proposed model is that after transformation $H^*(\cdot)$, of survival time T, the survival model is a linear model: the outcome

$H^*(T)$ relates to the predictor in a linear form. Under the model (1), the survival function given $\mathbf{Z}$ is

$$S_T(t, \mathbf{Z}) \equiv P(T \geq t \mid \mathbf{Z}) = S_\epsilon(\Lambda(t) e^{\boldsymbol{\gamma}^\top \mathbf{Z}}),$$

where $S_\epsilon(\cdot)$ is a survival function of $\epsilon = e^{\epsilon^*}$ and $\Lambda(\cdot) = e^{H^*(\cdot)}$. It follows that the cumulative hazard and hazard functions, respectively are $\Lambda(T^* \mid \mathbf{Z}) = \mathcal{G}\{e^{\boldsymbol{\gamma}^\top \mathbf{Z}} \Lambda(T^*)\}$ and $d\Lambda(T^* \mid \mathbf{Z}) = \mathcal{G}'\{e^{\boldsymbol{\gamma}^\top \mathbf{Z}} \Lambda(T^*)\} e^{\boldsymbol{\gamma}^\top \mathbf{Z}} d\Lambda(T^*)$ where $\mathcal{G}(\cdot) = -\log S_\epsilon(\cdot)$. We denote $d\Lambda(T_i^*) = \Lambda_i'$. A noteworthy feature of our proposal is that we start from a very general model that incorporates all possible interactions among genomic markers, but then accommodate other parsimonious models later to improve power of the proposed tests.

2.2 Testing procedure for an arbitrary subset of regression parameters

We develop a variance component score-based testing procedure for an arbitrary set of regression coefficients in model (1). We also provide mechanistic interpretation of various subsets of regression coefficients under the framework of causal mediation modeling in Section 3. For illustration, we focus on the testing of whether gene expression G is associated with survival given other markers. This corresponds to testing the hypothesis

$$H_0 : \boldsymbol{\beta} \equiv (\beta_G, \beta_{MG}, \boldsymbol{\beta}_{SG}^\top, \boldsymbol{\beta}_{SMG}^\top)^\top = \mathbf{0}, \tag{2}$$

but note that testing for any arbitrary set of regression coefficients can be developed similarly. Since $\boldsymbol{\beta}$ corresponds to the effect of $\mathbf{V} = (G, MG, G\mathbf{S}^\top, MG\mathbf{S}^\top)^\top$, containing all contributions from G, testing (2) can be used to assess the total effect of G on survival.

2.2.1 Derivation of the test statistic—

To test for H_0 in (2), we first rewrite the model (1) as

$$H^*(T_i) = -(\widetilde{\mathbf{X}}_i^\top \boldsymbol{\alpha} + \mathbf{V}_i^\top \boldsymbol{\beta}) + \epsilon_i^*, \tag{3}$$

where $\widetilde{\mathbf{X}}_i^\top = (\mathbf{X}_i^\top, \mathbf{S}_i^\top, M_i, M_i\mathbf{S}_i^\top)$ and $\boldsymbol{\alpha}^\top = (\boldsymbol{\beta}_X^\top, \boldsymbol{\beta}_S^\top, \beta_M, \boldsymbol{\beta}_{SM}^\top)$. Components of $\mathbf{V}$ may be highly correlated with each other due to correlation within $\mathbf{S}$ and among G, M and $\mathbf{S}$. The conventional approach such as likelihood ratio test or Wald test may not work well due to the instability in fitting model (1) that has a large number, $4p + 3 + q$, of potentially highly correlated predictors, especially when p is not small. Alternatively, one may employ a standard score test, which only requires fitting the null model. However, the type I error of the standard score test is not protected according to our stimulation studies in Section 4, probably due to the relatively large DF, $2p + 2$.

To overcome the problem, we propose a score test for $\boldsymbol{\beta}$ by imposing a working assumption that the parameters $\{\beta_{SGj}, j = 1, \ldots, p\}$ and $\{\beta_{SMGj}, j = 1, \ldots, p\}$ are $2p$ independent zero-mean random variables with $\text{var}(\beta_{SGj}) = \tau_{SG}$ and $\text{var}(\beta_{SMGj}) = \tau_{SMG}$. The hypothesis test for the null

(2) becomes jointly testing for the variance components (τ_{SG} and τ_{SMG}) [23] and two scalar regression coefficients (β_G and β_{MG}):

$$H_0: \tau_{SG} = \tau_{SMG} = \beta_G = \beta_{MG} = 0. \tag{4}$$

By assuming $\beta_{SGj} \sim F(0, \tau_{SG})$ where F is any arbitrary distribution, one can largely reduce the degree of freedom, i.e., $H_0: \beta_{SG1} = \ldots = \beta_{SGp} = 0$ vs. $H_0: \tau_{SG} = 0$. The score vector for $\boldsymbol{\beta}_{SG}$, $U_{\boldsymbol{\beta}_{SG}}$, is a p-variate normal asymptotically; and the standard score test based on $U_{\boldsymbol{\beta}_{SG}}$ is a p-DF test. The score test for τ_{SG} based on $U_{\tau_{SG}} = \|U_{\boldsymbol{\beta}_{SG}}\|^2$, which follows a mixture of chi-square distribution under the null, has an effective DF typically much lower than p. In finite sample, the distribution of $U_{\tau_{SG}}$ can be better approximated than that of $U_{\boldsymbol{\beta}_{SG}}$. One can show that the scores for τ_{SG}, τ_{SMG}, β_G and β_{MG} are:

$$U_{\tau_{SG}} = \|U_{\boldsymbol{\beta}_{SG}}\|^2 = \left\| n^{-1} \sum_i \kappa_i G_i \mathbf{S}_i \right\|^2, \qquad U_{\tau_{SMG}} = \|U_{\boldsymbol{\beta}_{SMG}}\|^2 = \left\| n^{-1} \sum_i \kappa_i M_i G_i \mathbf{S}_i \right\|^2,$$
$$U_{\beta_G} = n^{-1} \sum_i \kappa_i G_i, \qquad U_{\beta_{MG}} = n^{-1} \sum_i \kappa_i M_i G_i,$$

where

$$\kappa_i = \left\{ \delta_i \frac{\mathcal{G}''(e^{\boldsymbol{\alpha}^\top \widetilde{\mathbf{X}}_i} \Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\alpha}^\top \widetilde{\mathbf{X}}_i} \Lambda(T_i^*))} - \mathcal{G}'(e^{\boldsymbol{\alpha}^\top \widetilde{\mathbf{X}}_i} \Lambda(T_i^*)) \right\} e^{\boldsymbol{\alpha}^\top \widetilde{\mathbf{X}}_i} \Lambda(T_i^*) + \delta_i,$$

$\|U_{\boldsymbol{\beta}_{SG}}\|^2 = \sum_{j=1}^p U_{\beta_{SGj}}^2$, $U_{\beta_{SGj}} = n^{-1} \sum_{i=1}^n \kappa_i G_i S_{ji}$, $\|U_{\boldsymbol{\beta}_{SMG}}\|^2 = \sum_{j=1}^p U_{\beta_{SMGj}}^2$
$U_{\beta_{SMGj}} = n^{-1} \sum_{i=1}^n \kappa_i M_i G_i S_{ji}$. To combine informations from $U_{\tau_{SG}}$, $U_{\tau_{SMG}}$, U_{β_G} and $U_{\beta_{MG}}$, we propose a composite score statistic by taking a weighted sum of $U_{\tau_{SG}}$, $U_{\tau_{SMG}}$, $U_{\beta_G}^2$ and $U_{\beta_{MG}}^2$

$$Q = n(w_1 U_{\beta_G}^2 + w_2 U_{\beta_{MG}}^2 + w_3 U_{\tau_{SG}} + w_4 U_{\tau_{SMG}}) = \left\| n^{-1/2} \sum_i \kappa_i \mathbf{V}_{wi} \right\|^2, \tag{5}$$

where $\mathbf{V}_{wi}^\top = (\sqrt{w_1} G_i, \sqrt{w_2} M_i G_i, \sqrt{w_3} G_i \mathbf{S}_i^\top, \sqrt{w_4} M_i G_i \mathbf{S}_i^\top)$. Different weighting schemes for $\{w_1, w_2, w_3, w_4\}$ can be implemented to reflect the prior knowledge regarding the relative contributions of various genomic effects. If no such knowledge is available, we propose to weight each term using the inverse of its standard deviation. The asymptotic variances for $U_{\tau_{SG}}$, $U_{\tau_{SMG}}$, $U_{\beta_G}^2$ and $U_{\beta_{MG}}^2$ can be estimated from a Monte-Carlo perturbation procedure described in Section 2.2.2. Equal weighting $w_1 = w_2 = w_3 = w_4$ is equivalent to testing $H_0: \tau = 0$ where τ is a common variance of all elements in $\boldsymbol{\beta}^\top = (\beta_G, \beta_{MG}, \boldsymbol{\beta}_{SG}^\top, \boldsymbol{\beta}_{SMG}^\top)$, which is still a valid test but may not be powerful in practice since the information from different genomic markers may not be comparable due to different scales.

To calculate Q, one needs to estimate $\boldsymbol{\alpha}$ and $\Lambda(\cdot)$ under H_0 by fitting the null model:

$$H^*(T_i) = -\widetilde{\mathbf{X}}_i^{\mathsf{T}}\boldsymbol{\alpha} + \epsilon_i^*. \tag{6}$$

Estimating procedures to estimate $\boldsymbol{\alpha}$ and Λ' such as Expectation-Maximization (EM) algorithm to obtain the nonparametric maximum likelihood estimate (NPMLE) have been proposed [22]. However, a challenge remains in estimating $\boldsymbol{\alpha}$ as its dimension is large $(q+1+2p)$. We use a ridge regression to stablize the estimation by introducing an L_2 penalty on the coefficients corresponding to methylation related components. The penalized log-likelihood under the null model (6) is $l_p(\boldsymbol{\psi}) = l_n(\boldsymbol{\psi}) - \frac{1}{2}\lambda\boldsymbol{\beta}_S^{\mathsf{T}}\boldsymbol{\beta}_S - \frac{1}{2}\lambda\boldsymbol{\beta}_{SM}^{\mathsf{T}}\boldsymbol{\beta}_{SM}$ where $l_n(\boldsymbol{\psi}) = \sum_{i=1}^{n} l_i(\boldsymbol{\psi})$, l_i is the unit log likelihood under the null model (6), λ is a tuning parameter and $\boldsymbol{\psi}^{\mathsf{T}} = (\boldsymbol{\alpha}^{\mathsf{T}}, \Lambda'^{\mathsf{T}})$. The estimation of $\boldsymbol{\psi}$ can be achieved by solving the estimating equation $U_{\boldsymbol{\psi}}(\boldsymbol{\psi}) - \lambda\mathbf{I}_2\boldsymbol{\psi} = 0$ where $U_{\boldsymbol{\psi}}^{\mathsf{T}}(\boldsymbol{\psi}) = (U_{\boldsymbol{\alpha}}^{\mathsf{T}}, U_{\Lambda}^{\mathsf{T}})$, $U_{\boldsymbol{\alpha}}$ and U_{Λ_j} are provided in Appendix, $\mathbf{I}_2$ is $(q+1+2p+m)\times(q+1+2p+m)$ block diagonal matrix with the top $(q+1+2p)\times(q+1+2p)$ block diagonal matrix being $\mathbf{I}_{(q+1+2p)\times(q+1+2p)}$ and the bottom $m\times m$ block diagonal matrix being 0 with m being the number of events. For selection of the tuning parameter λ, we use generalized cross-validation (GCV) [24, 25] to estimate λ as the minimizer of the GCV function $\frac{l_n(\widehat{\boldsymbol{\psi}})}{n\{1-n^{-1}\mathrm{tr}(\mathbf{H})\}^2}$, where $\mathbf{H} = (\frac{\partial U_{\boldsymbol{\psi}}}{\partial\boldsymbol{\psi}} + \lambda\mathbf{I}_2)^{-1}\frac{\partial U_{\boldsymbol{\psi}}}{\partial\boldsymbol{\beta}}$. λ is searched within a range of $[0, \sqrt{n}\ /\ \log(n)]$ to ensure $\widehat{\lambda} = o(\sqrt{n})$, an assumption that we later use to derive the asymptotic distribution of $\widehat{Q}$, the estimate of Q. By plugging in the estimates of $\boldsymbol{\alpha}$ and Λ', one can obtain $\widehat{Q} = Q(\widehat{\boldsymbol{\psi}})$.

2.2.2 Distribution of $Q(\widehat{\boldsymbol{\psi}})$—Denote $\boldsymbol{\theta}^{\mathsf{T}} = (\boldsymbol{\beta}^{\mathsf{T}}, \boldsymbol{\psi}^{\mathsf{T}})$ and $\boldsymbol{\psi}_0$, $\boldsymbol{\beta}_0 (= \mathbf{0})$ and $\boldsymbol{\theta}_0$ to be true parameters under the null (4) for their counterparts $\boldsymbol{\psi}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Q can be re-expressed as an L_2 norm of the score for $\boldsymbol{\beta}$:

$$\widehat{Q} = \|n^{-1/2}U_{\beta}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}})\|^2.$$

Note that the weight w is involved in the test statistic. As expressed in $\mathbf{V}_{wi}$, the weighting scheme can be conceived as a pre-determined variable standardization before fitting the model. We show in Appendix that

$$n^{-1/2}U_{\beta}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}}) = n^{-1/2}\mathbf{A}U_{\zeta}(\boldsymbol{\theta}_0) + o_p(1)\cdot\mathbf{J}. \tag{7}$$

By continuous mapping theorem, asymptotic distribution of $\widehat{Q}$ is a function of the estimating equation U_{ζ}:

$$\widehat{Q} \xrightarrow{d} \|n^{-1/2}\mathbf{A}U_{\zeta}(\boldsymbol{\theta}_0)\|^2. \tag{8}$$

$n^{-1/2}\mathbf{A}U_{\zeta}(\boldsymbol{\theta}_0)$ can be approximated by a perturbation procedure [26, 27] using the estimating equation $n^{-1/2}\widehat{\mathbf{A}}\sum_i U_{\zeta i}(\widehat{\boldsymbol{\psi}})\mathcal{N}_i$ where $\overrightarrow{\mathcal{N}} = (\mathcal{N}_1, \ldots, \mathcal{N}_n)^{\mathsf{T}}$ is a vector of n independent standard

normal random variables; $\widehat{\mathbf{A}}$ is the empirical version of $\mathbf{A}$ by plugging in $\widehat{\boldsymbol{\psi}}$, the estimate under the null model (6) with L_2 penalty; $\mathbf{A} = \left[\mathbf{I}_{2p+2\times 2p+2}, \quad \frac{\partial U_{\beta}}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}^*}\left(-\frac{\partial U_{\psi}}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}^*} + \lambda \mathbf{I}_2\right)^{-1}\right]$ with $\boldsymbol{\psi}^*$ between $\widehat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}_0$; and $\frac{\partial U_{\beta}}{\partial \boldsymbol{\psi}}\frac{\partial U_{\psi}}{\partial \boldsymbol{\psi}}$, $U_{\xi} = \sum_i U_{\xi i}$ are provided in Appendix.

2.2.3 The omnibus test—While testing procedures derived under the three-way interaction model is robust to model misspecification, power may be compromised when the true underlying model does not involve certain interactions. Hence, it is desirable to develop a test that can accommodate different models to optimize statistical power. We propose an omnibus test that combines multiple p-values from testing under a range of models that incorporate different layers of interactions yet are all correct under the null. Specifically, we compute the minimum of these p-values from multiple models and compare the observed minimum p-value to its null distribution, approximated by a resampling perturbation procedure. The test statistic Q in (5) is derived under the outcome model (1), which assumes all possible two-way and three-way interactions. In this section, we denote the test statistic (5) as Q_4. Suppose that the outcome Y does not depend on the three-way interaction ($\boldsymbol{\beta}_{SMG} = \mathbf{0}$), or it does not depend on the three-way interaction, SNP-by-methylation or the SNP-by-expression interaction $\boldsymbol{\beta}_{SMG} = \boldsymbol{\beta}_{SM} = \boldsymbol{\beta}_{SG} = \mathbf{0}$), or it depends only on the main effect of gene expression ($\boldsymbol{\beta}_{SMG} = \boldsymbol{\beta}_{SM} = \boldsymbol{\beta}_{SG} = \mathbf{0}$ and $\beta_{MG} = 0$), then it is more powerful to test for $H_0 : \beta_G = \beta_{MG} = 0$, $\boldsymbol{\beta}_{SG} = \boldsymbol{\beta}_{SMG} = \mathbf{0}$ using the test statistics Q_3, Q_2, and Q_1, respectively, with corresponding $\mathbf{V}_{wi}^{\mathrm{T}} = (\sqrt{w_1}G_i, \sqrt{w_2}\boldsymbol{M}_i G_i, \sqrt{w_3}G_i\mathbf{S}_i^{\mathrm{T}})$, $(\sqrt{w_1}G_i, \sqrt{w_2}\boldsymbol{M}_i G_i)$ and $(\sqrt{w_1}G_i)$. Q_1-Q_4 all provide valid tests under the null. Under those more parsimonious models, the test statistic Q_4 loses power as it tests for unnecessary parameters. However, if the outcome model is truly determined by all two-way and three-way interactions as (1), Q_1-Q_3 will lose power compared to Q_4.

As shown in Section 2.2.2, the null distribution of Q can be estimated based on the empirical distribution of the perturbed statistics $\|n^{-1/2}\widehat{\mathbf{A}}\sum_i U_{\xi i}(\widehat{\boldsymbol{\psi}})\mathcal{N}_i\|^2$ conditional on the observed data. By generating independent $\overrightarrow{\mathcal{N}}$ repeatedly, the perturbed realization of Q can be obtained, denoted by $\{\widehat{Q}^{(b)}, b = 1, \ldots, B\}$, where B is the number of perturbations. The p-value can be approximated as the tail probability by comparing $\{\widehat{Q}^{(b)}\}$ with the observed $\widehat{Q}$. Hence one can calculate the p-values of the four candidate models by inputting $U_{\theta i}$ with $\mathbf{V}_{wi}^{\mathrm{T}} = (\sqrt{w_1}G_i)$, $(\sqrt{w_1}G_i, \sqrt{w_2}\boldsymbol{M}_i G_i)$, $(\sqrt{w_1}G_i, \sqrt{w_2}\boldsymbol{M}_i G_i, \sqrt{w_3}G_i\mathbf{S}_i^{\mathrm{T}})$ and $(\sqrt{w_1}G_i, \sqrt{w_2}\boldsymbol{M}_i G_i, \sqrt{w_3}G_i\mathbf{S}_i^{\mathrm{T}}, \sqrt{w_4}\boldsymbol{M}_i G_i\mathbf{S}_i^{\mathrm{T}})$, respectively for Q_1-Q_4, generating their perturbed realizations of the null counterpart for the candidate model k as $\{\widehat{Q}_k^{(b)}\}$, and comparing them with corresponding observed values $\widehat{Q}_k\,(k = 1, \ldots, 4)$. Note that for each perturbation b, the random normal perturbation variable $\overrightarrow{\mathcal{N}}^{(b)}$ is the same across the four tests. Let $\widehat{P}_k = \mathcal{S}_k(\widehat{Q}_k)$ be the p-value for the candidate model k, where $\mathcal{S}_k(q) = \mathrm{pr}\{\widehat{Q}_k^{(b)} > g\}$. The null distribution of the minimum p-value, $\widehat{P}_{\min} = \min_k \widehat{P}_k$ can be approximated by the empirical distribution of $\left\{\widehat{P}_{\min}^{(b)} = \min_k\{\mathcal{S}_k(\widehat{Q}_k^{(b)})\}, b = 1, \ldots, B\right\}$ given the observed data. The p-value of the omnibus test hence can be calculated by comparing $\widehat{P}_{\min}$ with $\{\widehat{P}_{\min}^{(b)}\}$.

3 Implication of testing a subset of coefficients

In this section, we provide mechanistic interpretation of our testing procedure. The effect on Y contributed by G can be examined by testing all the parameters related to G, as null (2). Similarly, those contributed by $\mathbf{S}$ and M, respectively, can be evaluated by testing

$$H_0 : \boldsymbol{\beta}_S = \boldsymbol{\beta}_{SM} = \boldsymbol{\beta}_{SG} = \boldsymbol{\beta}_{SMG} = \mathbf{0}$$
$$H_0 : \beta_M = \beta_{MG} = 0, \boldsymbol{\beta}_{SM} = \boldsymbol{\beta}_{SMG} = \mathbf{0}.$$

By testing different subsets of regression coefficients, we are able to examine the significance of various genomic effects on the survival outcome. The proposed integrative testing procedure helps identify useful biomarkers across multiple genomic data, which can also be potential therapeutic targets.

Furthermore, we can interpret the results under the framework of causal mediation modeling. In our data example, there are three path-specific effects (Figure 1): 1) the effect of DNA methylations on the outcome mediated through gene expression but not through microRNA, denoted by $\Delta_{\mathbf{S}\to G\to Y}$; 2) the effect of methylations mediated through microRNA and possibly through gene expression, denoted by $\Delta_{\mathbf{S}\to MY}$; and 3) the alternative effect of DNA methylations on the outcome, not through microRNA or mRNA gene expression, denoted by $\Delta_{\mathbf{S}\to Y}$. With identifiability assumptions discussed in Supplementary Materials[28], it has been shown that under the structure that M is determined by $\mathbf{S}$, G is determined by M, and G is also determined by $\mathbf{S}$ independent of M, $\Delta_{\mathbf{S}\to G\to Y}$ corresponds to all regression coefficients for G: β_G, β_{MG}, $\boldsymbol{\beta}_{SG}$ and $\boldsymbol{\beta}_{SMG}$; $\Delta_{\mathbf{S}\to MY}$ corresponds to all regression coefficients for M and G: β_M, β_G, β_{MG}, $\boldsymbol{\beta}_{SM}$, $\boldsymbol{\beta}_{SG}$ and $\boldsymbol{\beta}_{SMG}$; $\Delta_{\mathbf{S}\to Y}$ corresponds to all regression coefficients for $\mathbf{S}$: $\boldsymbol{\beta}_S$, $\boldsymbol{\beta}_{SM}$, $\boldsymbol{\beta}_{SG}$ and $\boldsymbol{\beta}_{SMG}$; the overall effect $\Delta_{overall}$ corresponds to all regression coefficients: $\boldsymbol{\beta}_S$, β_M, β_G, β_{MG}, $\boldsymbol{\beta}_{SM}$, $\boldsymbol{\beta}_{SG}$ and $\boldsymbol{\beta}_{SMG}$[15]. With these results, the testing procedures in Section 2.2 can be used to examine path-specific effects and thus have mechanistic implication. For example, the test for $H_0 : \beta_G = \beta_{MG} = 0, \boldsymbol{\beta}_{SG} = \boldsymbol{\beta}_{SMG} = \mathbf{0}$ is equivalent to that for $H_0 : \Delta_{\mathbf{S}\to G\to Y} = 0$; and the test statistic (5) assesses the effect of methylation $\mathbf{S}$ on the survival time T mediated through gene expression G. More discussions on path-specific effects under mediation analyses can be found in Supplementary Materials.

4 Simulation

We have conducted extensive simulation studies to evaluate the performance of the proposed methods and compare with the conventional score test. We investigate $p = 12$ DNA methylation markers of *GRB10*, microRNA miR-633 and mRNA expression of *GRB10* in $n = 271$ simulated subjects. To mimic the motivating data example of the survival study for glioblastoma multiforme or GBM, we simulate the data focusing on *GRB10* gene. We obtain 12 DNA methylation markers at *GRB10* from 271 GBM patients of TCGA data and simulate microRNA miR-633 expression, mRNA gene expression of *GRB10* and failure time based on the real methylation data. We assume cg25915982 at 50.85 Mb of chromosome 7 to be the causal methylation marker S_{causal}. MicroRNA miR-633 expression, mRNA expression of *GRB10* and survival time are generated using the causal marker, but

analyses are based on all 12 methylation markers, assuming we do not know the causal marker. miR-633 expression value M is generated by a model: $M_i = 5.75 + S_{causal,i} \times \delta_S + \epsilon_{M,i}$, where $\epsilon_{M,i}$ follows normal distribution with mean zero and standard deviation 0.05. mRNA expression of *GRB10* G is generated by a model: $G_i = -10 + S_{causal,i} \times \alpha_S + M_i \times \alpha_M + \epsilon_{G|M,i}$, where $\epsilon_{G|M,i}$ follow standard normal distribution. Survival time T is generated by a model: $\log T_i = S_{causal,i}\beta_S + M_i\beta_M + G_i\beta_G + M_iS_{causal,i}\beta_{SM} + G_iS_{causal,i}\beta_{SG} + M_iG_i\beta_{MG} + M_iG_iS_{causal,i}\beta_{SMG} + \epsilon_{Ti}$ where ϵ_{Ti} follow standard normal. Censored time C is selected to control the censoring proportion at 70%. Observed follow up time T^* is the minimum of T and C, and survival status is death if $T \leq C$ or censored if $T > C$. For $\mathcal{G}(\cdot)$ transformation in analyses, we consider Box-Cox transformation $\mathcal{G}(x) = \frac{(1+x)^\rho - 1}{\rho}$ with $\rho = 1.2$. We also conduct simulation studies where data are generated with Box-Cox transformation with $\rho = 1.2$ or 1.0 and analyses is performed with correctly specified model (see Supplementary Materials, Tables S2-S7).

By setting different configurations of δ's and α's, we are able to generate data according to different DNA methylation-microRNA-mRNA expression relationships illustrated in Table S1. But here we will focus on the first condition in Table S1: $\delta_S = 0.04$, $\alpha_S = 2.5$ and $\alpha_M = 2.0$ since the testing procedures under other conditions are the same or just special cases. We study the performance of tests under various configurations of β's. Empirical size and power are estimated as percentage of p-value < 0.05 in 2000 simulations.

4.1 Size and power of $\Delta_{S\to Y}$, $\Delta_{S\to G\to Y}$ and $\Delta_{S\to MY}$

Empirical size and power of testing $H_0 : \Delta_{S\to Y} = 0$ are presented in Table 1. Empirical sizes are correct under different null models: all β's are zero, all β's are zero except $\beta_M (= 0.3)$, all β's except $\beta_G (= 0.3)$ are zero. For settings under the alternatives, the test with correct model specification has optimal power, and the omnibus test can almost reach the optimal power across different settings. For example, under the setting with only main effects ($\beta_S = 0.4$, $\beta_M = \beta_G = 0.3$), the proposed test focusing on main effects has the optimal power 86.5%; under the setting with main effects and two-way interactions ($\beta_S = 0.1$, $\beta_M = \beta_G = \beta_{MG} = \beta_{SM} = \beta_{SG} = 0.3$, $\beta_{SMG} = 0$), the test under the correct model have the optimal power 67.2%; and omnibus tests are very close to the two optimal tests with power 80.4% and 55.1%, respectively (Table 1). Type I error of standard score test with $4p = 48$ DF is largely inflated probably due to the DF and the high correlation among the markers.

Empirical size and power of testing $H_0 : \Delta_{S\to G\to Y} = 0$ are presented in Table 2. Empirical sizes are correct under different null: all β's are zero, all β's except $\beta_S (= 0.2)$ are zero, all β's except $\beta_M (= 0.2)$ are zero. Under the alternatives, tests assuming the correct models perform the best and the omnibus test can almost reach the optimal power with limited power loss, similar to the results for $\Delta_{S\to Y}$. For instance, under the setting with $\beta_S = 0.2$, $\beta_M = 0.2$, $\beta_G = 0.3$ and all other β's to be zero, the test for main effects performs optimally with power 86.9%, and the omnibus test has power 80.7%. Type I error of the conventional score test with $2p + 1 (= 25)$ DF is again largely inflated.

Similarly, type I error of our proposed methods for $H_0 : \Delta_{\mathbf{S}\to MY} = 0$ is protected under the null (Table 3). In contrast, type I error of the conventional score test with $3p + 3$ DF is inflated. Under the alternatives, tests assuming the correct models perform optimally, and the omnibus test approaches the optimal power across a wide range of settings.

The test size is also protected at type I error rate of 0.005 and 0.0005 (Table S8). Additional simulation studies with multiple causal methylation loci (Tables S9-S14) and different combinations of sample size, the number of methylation markers and censoring proportion (Tables S15-17) are presented and discussed in Supplementary Materials (Section 2).

5 Data Applications

We present two data application examples, both assessing the genomic contribution to overall survival of GBM. GBM is the most common malignant brain tumor that is rapidly fatal with median survival time of 15 months [29]. Due to its poor prognosis and lack of well-established environmental risk factors, it is important to identify genomic markers for outcome prognostication, which also help understand the progression mechanism of this fatal disease. Multiple sets of genomic data as well as survival information have been archived on TCGA. Here we exploit the multi-platform genomic data to investigate the mechanism of epigenetic effect on GBM mortality.

5.1 *GRB10* gene and GBM survival

We integrate epigenetic DNA methylation of *GRB10*, expression of microRNA miR-633 and gene expression of *GRB10* to jointly model overall survival of GBM. There are 271 patients with complete level 3 data on methylation, microRNA and gene expression arrays. We combine 12 methylation loci at *GRB10* from Illumina 27K array and its expression value on Agilent G4502A expression array as well as the expression of microRNA, miR-633, to perform a gene-based integrated analysis. We have shown that DNA methylation of *GRB10* gene is significantly associated with overall survival of GBM, and that *GRB10* expression is regulated by its methylation [3], which is also supported by the existing literature [5]. We have found that two methylation sites of *GRB10* are associated with the expression of miR-633 with p-value = 0.017 and 0.012, and the expression of miR-633 is also highly associated with expression of *GRB10* with p-value = 0.0031 from Wald-type univariate hypothesis tests for least square estimators. Furthermore, literature has shown that *GRB10* gene is the target of miR-633 [4] and microRNA expression can be regulated by methylation [30]. Therefore, based on the evidence from literature and statistical analyses, we set up a model as Figure 1, with $\mathbf{S}$, M and G being 12 DNA methylation loci of *GRB10*, miR-633 and *GRB10* expressions, respectively.

The results of the proposed integrated analyses for *GRB10* are provided in Table 4. The effects of DNA methylation of *GRB10* mediated through *GRB10* expression ($\Delta_{\mathbf{S}\to G\to Y}$: omnibus p-value=0.0045) or miR-633 ($\Delta_{\mathbf{S}\to MY}$: omnibus p-value=0.0081) expression are prominent, compared to the effect independent of the two expression values ($\Delta_{\mathbf{S}\to Y}$: omnibus p-value=0.14). The overall effect of methylation on survival is also significant (omnibus p-value=0.012). In contrast, likelihood ratio test (LRT) can not be performed due to failure

in convergence when fitting model (1), and score test does not protect the type I error, as shown in simulation studies. We conclude that *GRB10* methylation has a significant effect on overall survival of GBM, which is mostly mediated by miR-633 expression or *GRB10* expression.

5.2 miR-223 and GBM survival

In the second example, we apply our proposed procedures to examine the effect between miR-223 and GBM survival, accounting for expression values of 16 mediation genes. Our previous work suggests that the prognostic effect of miR-223 expression is mediated by expression levels of the 16 genes [31]. We set up a integrated analysis illustrated in Figure 2. It can be viewed as a simplified case of Figure 1, with $\mathbf{S}$ being the scalar expression value of miR-223, $M = G$ being the expression values of the 16 mediation genes. It follows that there are only two path-specific effects: $\Delta_{S\to \mathbf{G}\to Y}$, the effect of miR-223 expression on the GBM survival, mediated through expressions of the 16 mediation genes, and $\Delta_{S\to Y}$, the effect of miR-223 expression independent of the 16 mediation genes.

There are 504 GBM patients with complete level 3 data on microRNA and gene expression arrays. Both path-specific effects of miR-223 are highly significant, as shown in Table 5. The omnibus p-value for the effect of miR-223 mediated through the 16 genes is $< 10^{-6}$, and the p-value for the effect of miR-223 independent of the 16 genes is 0.0009. The p-value of the overall effect is 0.0008. We conclude that miR-223 may be a promising prognostic marker for GBM patients, and the mechanisms mediated through gene expression or other pathways are both highly significant and deserve further research.

6 Discussion

In this paper, we propose a testing procedure for path-specific effects of genomic markers on survival outcome through a semiparametric linear transformation modeling framework. We are able to decompose the genomic effect into molecule-specific components using the path-specific effect approach. In addition to shedding light on the mechanism of disease etiology, the path-specific effect may have translational utility. Epigenetic alterations such as microRNA expression and DNA methylation are potentially reversible [32, 33, 34], and microRNA regulation has specificity in target genes. The findings from our path-specific effect analyses provide more specific hypotheses and mechanisms for biologists to validate, compared to conventional epigenome-wide association studies. Furthermore, the path-specific effect can also highlight biomarkers where therapeutic devices may be developed. For example, we observe a significant effect of DNA methylation of *GRB10* mediated through miR-633 $\Delta_{S\to MY}$ and its mRNA expression $\Delta_{S\to G\to Y}$ (Table 4); one may thus design a gene-specific intervention on mRNA expression of *GRB10* through miR-633 or other small RNA to improve GBM survival even though there is little gene- or loci-specific intervention is available on DNA methylation.

We note that carrying out the NPMLE and the resampling perturbation procedures is computationally intensive but not prohibitive. For the analyses of GBM survival data in Section 5.1 performed on a laptop with Intel i5-3380M 2.90 GHz CPU and 8.00 RAM,

the proposed testing procedure with 1000 resampling perturbation takes 3.95 seconds if the tuning parameter λ is pre-specified and 30.30 seconds if λ is selected via GCV. All simulation studies (n=271 and p=12; 1000 resampling perturbation and 2000 replicated) are performed using a computer cluster with 2 - 8core Intel Xeon CPUs running at 2.53 GHz, 24.00 RAM and a Linux environment. The total time for completing each simulation is 2.58 hours with pre-specified λ and 15.50 hours with GCV selected λ. The Matlab codes are available in Supplementary Materials.

The proposed test is a score test for the variance component of the parameters of interest. Instead of fitting a large model as shown in (1), one only needs to fit a model under the null, which makes the method numerically stable. The non-parametric maximum likelihood estimator, proposed by Zeng and Lin [22] for the null model using Newton-Raphson or EM algorithm requires iteration where we use $\boldsymbol{\alpha} = 0$ and Λ' being the inverse of the number of events as initial values. In our simulation studies, the convergence rates are extremely high with 99.8% for $\Delta_{S\to G\to Y}$ and 100% for $\Delta_{S\to Y}$ and $\Delta_{S\to MY}$. One alternative would be to obtain initial parameters from a consistent estimator [20] to assure a better convergence and to stablize the estimating procedure. On the other hand, as the proposed method relies on a resampling-based perturbation procedure to approximate the tail probability, it remains difficult to precisely approximate a very small p-value in practice.

Our approach extends the previous work for genetic analyses [7, 8] to facilitate integrated genomic analyses, and the proposed omnibus test synthesizes information from various candidate models to boost statistical power as well as to preserve the robustness to model misspecification. The linear transformation model has also been extended to incorporate dependent failure time, repeated measurement as well as time-varying covariates [22]. Based on our current work, its flexibility may facilitate future directions for big data sciences. For instance, the model (1) can be easily extended to incorporate time-varying genomic markers. As the genomic profile is dynamic during cancer development, 'time-varying integrative genomics' may better reveal the biological mechanisms behind this fatal disease.

The estimate of α in (6) is biased using an $L2$ ridge regression. The bias is a function of the tuning parameter λ. We address this in our theoretical development as well as in numerical studies. It should be noted that here we focus on hypothesis testing rather than estimation, and our testing procedure is developed under the null. To ensure its validity, one has to derive the distribution of test statistic $Q(\hat{\psi})$ that incorporates λ under the null. We show in Appendix 7.2 and Section 2.2.2 that with a bounded tuning parameter $\lambda = o(\sqrt{n})$, the asymptotic distribution of $\hat{\psi}$ is a function of score U_ζ and λ in (8). In real application, one still has to approximate $A(\psi)$ and $U_\zeta(\psi)$ in (8) by plugging in $\hat{\psi} = (\hat{\alpha}^\mathsf{T}, \hat{\Lambda}'^\mathsf{T})^\mathsf{T}$. Therefore, we also evaluate the validity of our testing procedure in simulation studies with empirical estimates under finite sample. As shown in the first three columns of Table 2 (Null), our proposed testing procedures Q_1-Q_4 and the omnibus test protect Type I Error at 5%.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the editor, the associate editor and two anonymous referees for their insightful comments that improved the presentation of the paper. This study is supported by National Institutes of Health grants CA182937 and AG048825.

7: Appendix

7.1 Estimating equation of model (1)

The log-likelihood can be written as $l_n = -\sum_i \{\delta_i \log d\Lambda(T_i^* \mid \mathbf{Z}_i) - \Lambda(T_i^* \mid \mathbf{Z}_i)\}$, where $\delta_i = 1$ if subject i is death and 0 otherwise and $\Lambda(T_i^*) = \sum_j I(T_j^* \le T_i^*)\Lambda_j^{'}$. It follows that the score for $\boldsymbol{\gamma}$ and $\Lambda_j^{'}$ are:

$$U_{\boldsymbol{\gamma}} = \sum_i \left[\left\{ \delta_i \frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))} - \mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)) \right\} e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)\mathbf{Z}_i + \delta_i \mathbf{Z}_i \right]$$

$$U_{\Lambda_j^{'}} = \frac{1}{\Lambda_j^{'}} + \sum_i \left[\left\{ \delta_i \frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))} - \mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)) \right\} e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i} I(T_j^* \le T_i^*) \right].$$

The scores for $\boldsymbol{\gamma}$ and $\Lambda_j^{'}$ can be re-expressed as a set of estimating equations:

$$U_{\boldsymbol{\gamma}} = \sum_i U_{\boldsymbol{\gamma} i}, \qquad U_{\Lambda_j} = \sum_i U_{\Lambda_j i}, \qquad j = 1, \ldots, m,$$

and

$$U_{\boldsymbol{\gamma} i} = \begin{bmatrix} U_{\boldsymbol{\beta} i} \\ U_{\boldsymbol{\alpha} i} \end{bmatrix} = \left[\left\{ \delta_i \frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))} - \mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)) \right\} e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*) + \delta_i \right] \mathbf{Z}_i$$

$$U_{\Lambda_j i} = \left[\left\{ \delta_i \frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))} - \mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)) \right\} e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda^{*(j)}(T_i^*) + \delta_i I(T_j^* \le T_i^*) \right].$$

where $\Lambda^{*(j)}(T_i^*) = \Lambda(T_i^*)I(T_i^* \le T_j^*) + \Lambda(T_j^*)I(T_i^* > T_j^*)$. We can denote $U_\xi^\top = (U_{\boldsymbol{\gamma}}^\top, U_\Lambda^\top) = (U_{\boldsymbol{\beta}}^\top, U_{\boldsymbol{\psi}}^\top)$ and $U_\Lambda^\top = (U_{\Lambda_1}, \ldots, U_{\Lambda_m})$.

And the derivatives of the estimating equations are:

$$\frac{\partial U_\xi}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial U_{\boldsymbol{\gamma}}}{\partial \boldsymbol{\gamma}} & \frac{\partial U_{\boldsymbol{\gamma}}}{\partial \Lambda^{'}} \\ \frac{\partial U_\Lambda}{\partial \boldsymbol{\gamma}} & \frac{\partial U_\Lambda}{\partial \Lambda^{'}} \end{bmatrix} = \begin{bmatrix} \frac{\partial U_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} & \frac{\partial U_{\boldsymbol{\beta}}}{\partial \boldsymbol{\psi}} \\ \frac{\partial U_{\boldsymbol{\psi}}}{\partial \boldsymbol{\beta}} & \frac{\partial U_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}} \end{bmatrix}.$$

The element of $\frac{\partial U_\xi}{\partial \boldsymbol{\theta}}$ can be expressed as follows:

$$\frac{\partial U_\gamma}{\partial \boldsymbol{\gamma}} = \sum_i (d_{1i}\Lambda^2(T_i^*) + d_{0i}\Lambda(T_i^*))\,\mathbf{Z}_i\mathbf{Z}_i^\top, \qquad \frac{\partial U_{\Lambda_j}}{\partial \Lambda_k'} = \sum_i (d_{1i}\Lambda^{*(j)}(T_i^*) + d_{0i}I(T_k^* \le T_j^*))\, I(T_k^* \le T_i^*)$$

$$\frac{\partial U_\gamma}{\partial \Lambda_j'} = \sum_i (d_{1i}\Lambda(T_i^*) + d_{0i})\,\mathbf{Z}_i I(T_j^* \le T_i^*), \qquad \frac{\partial U_{\Lambda j}}{\partial \boldsymbol{\gamma}} = \sum_i (d_{1i}\Lambda(T_i^*) + d_{0i})\,\Lambda^{*(j)}(T_i^*)\mathbf{Z}_i^\top,$$

where $d_{1i} = \left[\left\{\frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)} - \left(\frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*)}\right)^2\right\}\delta_i - \mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))\right] e^{2\boldsymbol{\gamma}^\top \mathbf{Z}_i}$ and

$d_{0i} = \left\{\delta_i \frac{\mathcal{G}''(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))}{\mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))} - \mathcal{G}'(e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}\Lambda(T_i^*))\right\} e^{\boldsymbol{\gamma}^\top \mathbf{Z}_i}$, and the (j, k)-th element of $\frac{\partial U_\Lambda}{\partial \Lambda'}$ is $\frac{\partial U_{\Lambda_j}}{\partial \Lambda_k'}$.

7.2 Distribution of $Q(\widehat{\boldsymbol{\psi}})$

Denote $\boldsymbol{\alpha}_0$, Λ_0', $\boldsymbol{\psi}_0$, $\boldsymbol{\beta}_0 (= \mathbf{0})$ and $\boldsymbol{\theta}_0$ are the true parameters under the null (4) for their counterparts $\boldsymbol{\alpha}$, Λ' , $\boldsymbol{\psi}$ $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. A simple Taylor series expansion shows

$$n^{-1/2}\widehat{U}_\beta(\boldsymbol{\beta}_0) = \quad n^{-1/2}U_\beta(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}}) = n^{-1/2}U_\beta(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + n^{-1/2}\frac{\partial U_\beta}{\partial \boldsymbol{\psi}}\Big|_{\psi^*}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 \quad \text{(A. 1)}$$

),

where $\boldsymbol{\psi}^*$ is between $\widehat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}_0$. Another Taylor expansion can show that

$$\begin{aligned} 0 = n^{-1/2}\widehat{U}_\psi^\lambda(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}}) &= n^{-1/2}U_\psi(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}}) - n^{-1/2}\lambda \mathbf{I}_2\boldsymbol{\psi} \\ &= \quad (n^{-1/2}U_\psi(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + n^{-1/2}\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}\Big|_{\psi^*}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) - n^{-1/2}\lambda \mathbf{I}_2\widehat{\boldsymbol{\psi}} \\ &= n^{-1/2}U_\psi(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + n^{-1/2}(\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}\Big|_{\psi^*} - \lambda \mathbf{I}_2)(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) - n^{-1/2}\lambda \mathbf{I}_2\boldsymbol{\psi}_0, \end{aligned}$$

where $\mathbf{I}_2$ is $(q+m)\times(q+m)$ block diagonal matrix with the top $q\times q$ block diagonal matrix being $\mathbf{I}_{q\times q}$ and the bottom $m\times m$ block diagonal matrix being 0. Since $\lambda = o(\sqrt{n})$, it follows that $\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) = \left[n^{-1}\left(-\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}\Big|_{\psi^*} + \lambda \mathbf{I}_2\right)\right]^{-1} n^{-1/2}U_\psi(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + o_p(1)\cdot \mathbf{J}$, where $\mathbf{J}$ is a vector of 1's with length the same as $\boldsymbol{\beta}$. By plugging it in (A. 1), one can obtain

$$\begin{aligned} n^{-1/2}U_\beta(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}}) = \quad & n^{-1/2}U_\beta(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + n^{-1/2}\frac{\partial U_\beta}{\partial \boldsymbol{\psi}}\Big|_{\psi^*}\left(\quad -\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}\Big|_{\psi^*} + \lambda \mathbf{I}_2\right)^{-1} U_\psi(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + o_p(1)\cdot \mathbf{J} \\ = n^{-1/2}\Big[\quad & U_\beta(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0) + \frac{\partial U_\beta}{\partial \boldsymbol{\psi}}\Big|_{\psi^*}\left(\quad -\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}\Big|_{\psi^*} + \lambda \mathbf{I}_2\right)^{-1} U_\psi(\boldsymbol{\beta}_0, \boldsymbol{\psi}_0)\Big] + o_p(1)\cdot \mathbf{J}. \end{aligned}$$

Thus (A. 1) becomes

$$n^{-1/2}U_\beta(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\psi}}) = n^{-1/2}\mathbf{A}U_\zeta(\boldsymbol{\theta}_0) + o_p(1)\cdot \mathbf{J}. \quad \text{(A. 2)}$$

Recall $\mathbf{A} = \left[\mathbf{I}_{2p+2\times 2p+2}, \quad \frac{\partial U_\beta}{\partial \boldsymbol{\psi}}\Big|_{\psi^*}\left(-\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}\Big|_{\psi^*} + \lambda \mathbf{I}_2\right)^{-1}\right]$, and $\frac{\partial U_\beta}{\partial \boldsymbol{\psi}}$, $\frac{\partial U_\psi}{\partial \boldsymbol{\psi}}$, U_ζ are provided in the above section

References

[1]. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA and Visscher PM. Finding the missing heritability of complex diseases. Nature 2009; 461(7265): 747–753. DOI: 10.1038/nature08494. [PubMed: 19812666]

[2]. Wang W, Baladandayuthapani V, Morris JS, Boom BM, Manyam G and Do KA. iBAG: integative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics 2013; 29(2):149–159. DOI: 10.1093/bioinformatics/bts655. [PubMed: 23142963]

[3]. Smith AA, Huang YT, Eliot M, Houseman EA, Marsit JK, Wiencke JK and Kelsey KT A novel approach to the discovery of survival biomarkers in glioblastoma using a joint analysis of DNA methylation and gene expression. Epigenetics 2014; 9(6): 873–883. DOI: 10.4161/epi.28571. [PubMed: 24670968]

[4]. Jia P, Sun J, Guo AY and Zhao Z. SZGR: a comprehensive schizophrenia gene resource. Molecular Psychiatry 2010; 15(5):453–462. DOI: 10.1038/mp.2009.93. [PubMed: 20424623]

[5]. Turan N, Ghalwash MF, Katari S, Coutifaris C, Obradovic Z and Sapienza C. DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? BMC Medical Genomics 2012; 5:10. DOI: 10.1186/1755-8794-5-10. [PubMed: 22498030]

[6]. Robins JM Semantics of causal DAG models and the identification of direct and indirect effects. Oxford University Press, New York, 2003.

[7]. Cai T, Tonini G and Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 2011; 67(3): 975–986. DOI: 10.1111/j.1541-0420.2010.01544.x. [PubMed: 21281275]

[8]. Tzeng JY, Lu W and Hsu FC. Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression. The Annals of Applied Statistics 2014; 8(2): 1232–1255. DOI:10.1214/14-AOAS735. [PubMed: 25018788]

[9]. Robins JM and Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992; 3(2): 143–155. [PubMed: 1576220]

[10]. Pearl J. Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. Morgan Kaufmann, San Francisco, 2001; 411–420.

[11]. VanderWeele TJ and Vansteelandt S. Conceptual issues concerning mediation, intervention and composition. Statistics and its Interface 2009; 2:457–468. DOI: 10.4310/SII.2009.v2.n4.a7.

[12]. Imai K, Keele L and Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science 2010; 25(1):51–71. DOI:10.1214/10-STS321.

[13]. Huang YT, VanderWeele TJ and Lin X. Joint analysis of SNP and expression data in genetic association studies of complex diseases. Annals of Applied Statistics 2014; 8(1):352–376. DOI: 10.1214/13-AOAS690. [PubMed: 24729824]

[14]. Zhao SD, Cai TT and Li H. More powerful genetic association testing via a new statistical framework for integrative genomics. Biometrics 2014; 70(4): 881–890. DOI: 10.1111/biom.12206. [PubMed: 24975802]

[15]. Huang YT. Integrative modeling of multiplatform genomic data under the framework of mediation analysis. Statistics in Medicine 2015; 34(1): 162–178. DOI: 10.1002/sim.6326. [PubMed: 25316269]

[16]. Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 1972; 34(2): 187–220.

[17]. Anderson PK, and Gill RD. Cox's regression model for counting process: a large sample study. Annals of Statistics 1982; 10(4): 1100–1120. DOI: 10.1214/aos/1176345976.

[18]. Kalbfleisch JD and Prentice RL. The Statistical Analysis of Failure Time Data, 2nd Edition, Hoboken: Wiley, 2002.

[19]. Bennett S. Analysis of survival data by the proportional odds model. Statistics in Medicine 1983; 2(2): 273–277. DOI: 10.1002/sim.4780020223. [PubMed: 6648142]

[20]. Cheng SC, Wei LJ and Ying Z. Analysis of transformation models with censored data. Biometrika 1995; 82(4): 835–845. DOI: 10.1093/biomet/82.4.835.

[21]. Cai T, Cheng SC and Wei LJ. Semiparametric mixed-effects models for clustered failure time data. Journal of the American Statistical Association 2002; 97(458): 514–522. DOI:10.1198/016214502760047041.

[22]. Zeng D and Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. Journal of the Royal Statistical Society, Series B 2007; 69(4): 507–564. DOI: 10.1111/j.1369-7412.2007.00606.x.

[23]. Lin X. Variance component test in generalised linear models with random effects. Biometrika 1997; 84(2):309–326. DOI: 10.1093/biomet/84.2.309.

[24]. Craven P and Wahba G. Smoothing noisy data with spline functions. Numerische Mathematik 1979; 31(4): 377–403. DOI: 10.1007/BF01404567.

[25]. O'Sullivan F, Yandell BS and Raynor WJ Jr. Automatic smoothing of regression functions in generalized linear models. Journal of the American Statistical Association 1986; 81(393):96–103. DOI: 10.1080/01621459.1986.10478243.

[26]. Parzen M, Wei LJ, and Ying Z. A resampling method based on pivotal estimating functions. Biometrika 1994; 81(2):341–350. DOI: 10.2307/2336964.

[27]. Cai T, Wei LJ, and Wilcox M. Semiparametric regression analysis for clustered failure time data. Biometrika 2000; 87(4): 867–878. DOI: 10.1093/biomet/87.4.867.

[28]. Huang YT and Cai T. Mediation analysis for survival data using semiparametric probit models. Biometrics 2016; DOI: 10.1111/biom.12445.

[29]. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J, Janzer RC, Ludwin SK, Gorlia T, Allgeier A, Lacombe D, Cairncross JG, Eisenhauer E, Mirimanoff RO, European Organisation for Treatment of Cancer Brain Tumor and Radiotherapy Groups and National Cancer Institute of Canada Clinical Trials Group. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine 2005; 352(10):987–996. DOI: 10.1056/NEJMoa043330. [PubMed: 15758009]

[30]. Suzuki H, Maruyama R, Yamamoto E and Kai M. DNA methylation and microRNA dysregulation in cancer. Molecular Oncology 2012; 6(6):567–578. DOI: 10.1016/j.molonc.2012.07.007. [PubMed: 22902148]

[31]. Huang YT, Hsu T, Kelsey KT and Lin CL. Integrative analysis of micro-RNA, gene expression and survival of glioblastoma multiforme. Genetic Epidemiology 2015; 39(2): 134–143. DOI: 10.1002/gepi.21875. [PubMed: 25537983]

[32]. Issa JP, Gharibyan V, Cortes J, Jelinek J, Morris G, Verstovsek S, Talpaz M, Garcia-Manero G and Kantarjian HM. Phase II study of low-dose decitabine in patients with chronic myelogenous leukemia resistant to imatinib mesylate. Journal of Clinical Oncology 2005; 23(17):3948–3956. DOI: 10.1200/JCO.2005.11.981. [PubMed: 15883410]

[33]. Kaminskas E, Farrell A, Abraham S, Baird A, Hsieh LS, Lee SL, Leighton JK, Patel H, Rahman A, Sridhara R, Wang YC and Pazdur R. Approval summary: azacitidine for treatment of myelodysplastic syndrome subtypes. Clinical Cancer Research 2005; 11(10):3604–3608. DOI: 10.1158/1078-0432.CCR-04-2135. [PubMed: 15897554]

[34]. Garcia-Manero G, Kantarjian HM, Sanchez-Gonzalez B, Yang H, Rosner G, Verstovsek S, Rytting M, Wierda WG, Ravandi F, Koller C, Xiao L, Faderl S, Estrov Z, Cortes J, O'Brien S, Estey E, Bueso-Ramos C, Fiorentino J, Jabbour E and Issa JP. Phase 1/2 study of the combination of 5-aza-2'-deoxycytidine with valproic acid in patients with leukemia. Blood 2006; 108(10):3271–3279. DOI: 10.1182/blood-2006-03-009142. [PubMed: 16882711]

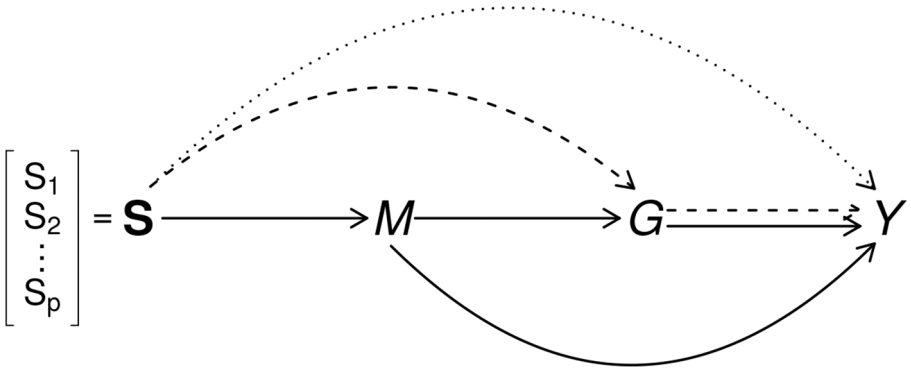

Figure 1:

Causal diagram of a set of DNA methylations (**S**), microRNA expression (M), gene expression (G) and outcome of interest ($Y = H^*(T)$). Three path-specific effects are in different line styles: $\Delta_{\mathbf{s}\to Y}$, effect of methylation on outcome independent of microRNA and mRNA gene expression is in dotted line; $\Delta_{\mathbf{s}\to G\to Y}$, effect of methylation mediated through gene expression but not through microRNA is in dashed lines; $\Delta_{\mathbf{s}\to MY}$, effect mediated through microRNA is in solid lines.

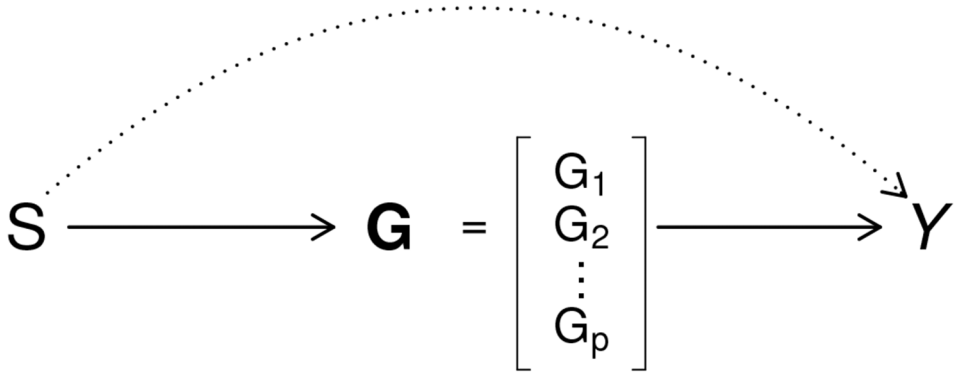

Figure 2:

Causal diagram of microRNA miR-223 expression (S), 16 gene expression values (**G**) and outcome of interest ($Y = H^*(T)$). Two path-specific effects are in different line styles: $\Delta_{S \to Y}$, effect of miR-223 on outcome independent of 16 mRNA gene expression is in dotted line; $\Delta_{S \to G \to Y}$, effect of miR-223 expression mediated through mRNA expression of the $p\,(=16)$ genes is in solid lines.

Table 1:

Empirical size and power (%) of testing $\Delta_{S \to Y}$. Q_1: model with only main effects; Q_2: model with main effects and microRNA-by-expression interaction; Q_3: model with main effects and two-way interactions; Q_4: model with main effects, two-way and three-way interactions; Omnibus: the omnibus test for Q_1-Q_4; Score test: the classic score test for β's.

	Null			**Alternative**							
β_S	0	0	0	0.4	0.2	0.4	0.6	0.1	0.1	0	0
β_M	0	0.3	0	0	0.3	0.3	0.3	0.3	0.3	0	0
β_G	0	0	0.3	0	0.3	0.3	0.3	0.3	0.3	0	0
β_{MG}	0	0	0	0	0	0	0	0.3	0.3	0	0
β_{SM}	0	0	0	0	0	0	0	0.2	0.3	0	0
β_{SG}	0	0	0	0	0	0	0	0.2	0.3	0	0
β_{SMG}	0	0	0	0	0	0	0	0	0	0.5	1.0
Q_1	4.10	3.90	3.85	89.3	25.8	85.0	99.9	4.25	3.85	5.90	7.00
Q_2	4.10	4.20	3.85	89.3	26.5	84.9	99.9	4.20	4.10	6.25	8.50
Q_3	3.65	3.95	4.35	70.5	14.3	59.6	95.8	19.7	55.8	7.10	19.8
Q_4	2.95	3.30	3.25	62.1	11.0	49.4	91.9	16.1	46.5	61.9	95.6
Omnibus	3.90	4.10	3.75	83.4	21.3	78.0	99.6	13.2	43.4	45.7	90.0
Score test	48.3	49.9	51.9								

Table 2:

Empirical size and power (%) of testing $\Delta_{S \to G \to Y}$. Q_1: model with only main effects; Q_2: model with main effects and microRNA-by-expression interaction; Q_3: model with main effects and two-way interactions; Q_4: model with main effects, two-way and three-way interactions; Omnibus: the omnibus test for Q_1-Q_4; Score test: the classic score test for β's. The tuning parameter λ was chosen using GCV.

	Null			**Alternative**							
β_S	0	0.2	0	0	0.2	0.2	0.2	0.1	0.1	0	0
β_M	0	0	0.2	0	0.2	0.2	0.2	0.1	0.1	0	0
β_G	0	0	0	0.3	0.2	0.3	0.4	0.1	0.1	0	0
β_{MG}	0	0	0	0	0	0	0	0.3	0.5	0	0
β_{SM}	0	0	0	0	0	0	0	0.3	0.5	0	0
β_{SG}	0	0	0	0	0	0	0	0.3	0.5	0	0
β_{SMG}	0	0	0	0	0	0	0	0	0	0.5	0.8
Q_1	4.70	4.40	4.65	90.2	56.2	88.5	98.8	6.30	3.35	7.55	11.3
Q_2	5.85	4.95	4.90	82.5	44.3	81.1	97.0	9.60	8.45	9.60	13.4
Q_3	5.00	5.20	5.00	73.5	36.8	72.8	94.1	60.5	93.5	10.9	20.7
Q_4	5.10	4.90	4.60	66.9	31.6	66.7	91.5	53.3	88.7	63.8	88.7
Omnibus	5.35	4.70	4.80	84.1	47.2	82.8	97.6	45.8	85.1	45.2	78.5
Score test	45.3	45.0	48.3								

Table 3:

Empirical size and power (%) of testing $\Delta_{S \to MY}$. Q_1: model with only main effects; Q_2: model with main effects and microRNA-by-expression interaction; Q_3: model with main effects and two-way interactions; Q_4: model with main effects, two-way and three-way interactions; Omnibus: the omnibus test for Q_1-Q_4; Score test: the classic score test for β's. The tuning parameter λ was chosen using GCV.

	Null		Alternative								
β_S	0	0.4	0	0	0.3	0.3	0.3	0.1	0.1	0	0
β_M	0	0	0.3	0	0.1	0.2	0.3	0	0	0	0
β_G	0	0	0	0.3	0.2	0.2	0.2	0	0	0	0
β_{MG}	0	0	0	0	0	0	0	0.3	0.5	0	0
β_{SM}	0	0	0	0	0	0	0	0.3	0.5	0	0
β_{SG}	0	0	0	0	0	0	0	0.3	0.5	0	0
β_{SMG}	0	0	0	0	0	0	0	0	0	0.5	1.0
Q_1	4.65	5.00	89.8	83.0	62.9	87.8	98.2	4.90	5.15	47.2	78.3
Q_2	5.25	5.35	86.0	75.8	57.0	83.5	96.6	7.45	9.65	43.4	73.6
Q_3	5.55	4.80	76.9	65.4	44.9	72.8	92.9	49.1	93.3	36.3	67.3
Q_4	4.30	4.60	73.9	59.0	41.7	69.6	91.3	41.0	87.3	80.6	99.0
Omnibus	5.30	4.65	86.4	76.6	55.9	83.2	96.6	33.9	85.5	69.6	97.4
Score test	51.0	49.0									

Table 4:

p-values for three path-specific effects of *GRB10* gene and miR-633 on GBM survival. Q_1-Q_4 correspond to the models mentioned in Section 2.2.3.

	$\Delta_{S \to Y}$	$\Delta_{S \to G \to Y}$	$\Delta_{S \to MY}$	$\Delta_{overall}$
Q_1	0.10	0.0047	0.0047	0.0142
Q_2	0.09	0.0047	0.0052	0.0086
Q_3	0.14	0.0035	0.0150	0.0163
Q_4	0.21	0.0045	0.0170	0.0159
Omnibus	0.14	0.0045	0.0081	0.0119

Table 5:

p-values for two path-specific effects of miR-223 (S) and 16 mediation genes ($\mathbf{G}$) on GBM survival. Q_1 corresponds to the main-effect model, and Q_2 corresponds to the model with both main and interactive effects.

	$\Delta_{S\to Y}$	$\Delta_{S\to G\to Y}$	$\Delta_{overall}$
Q_1	0.0007	$< 10^{-6}$	0.0045
Q_2	0.0052	$< 10^{-6}$	0.0009
Omnibus	0.0009	$< 10^{-6}$	0.0008