Hum Genet
DOI 10.1007/s00439-013-1350-z

ORIGINAL INVESTIGATION

A novel method for identifying nonlinear gene–environment interactions in case–control association studies

Cen Wu · Yuehua Cui

Received: 13 May 2013 / Accepted: 5 August 2013

Abstract The genetic influences on complex disease traits generally depend on the joint effects of multiple genetic variants, environmental factors, as well as their interplays. Gene × environment (G × E) interactions play vital roles in determining an individual's disease risk, but the underlying genetic machinery is poorly understood. Traditional analysis assuming linear relationship between genetic and environmental factors, along with their interactions, is commonly pursued under the regression-based framework to examine G × E interactions. This assumption, however, could be violated due to nonlinear responses of genetic variants to environmental stimuli. As an extension to our previous work on continuous traits, we proposed a flexible varying-coefficient model for the detection of nonlinear G × E interaction with binary disease traits. Varying coefficients were approximated by a non-parametric regression function through which one can assess the nonlinear response of genetic factors to environmental changes. A group of statistical tests were proposed to elucidate various mechanisms of G × E interaction. The utility of the proposed method was illustrated via simulation and real data analysis with application to type 2 diabetes.

Keywords B-spline · Nonlinear genetic penetrance · Phenotypic plasticity · Type 2 diabetes

Abbreviations

BIC	Bayesian information criterion
BMI	Body mass index
G × E	Gene–environment interaction
GENVEA	Gene, Environment Association Studies Consortium
GWAS	Genome-wide association study
HPFS	Health Professionals Follow-up Study
LM	Linear predictor model
LM-I	Linear predictor model with interaction
MAF	Minor allele frequency
NHS	Nurses' Health Study
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes mellitus
VC	Varying-coefficient

Introduction

It has been increasingly recognized that the predisposition of many complex diseases is not purely triggered by genetic factors. They are also influenced by environmental exposures, due to potential gene–environment interactions. For example, type 2 diabetes mellitus is a typical complex human disease whose incidence is heavily contingent on the environmental exposures such as behavioral and dietary factors, in addition to genetic susceptibility (Zimmet et al. 2001; Patel et al. 2013). Studies on gene × environment (G × E) interactions will shed novel light on the genetic responses to environment dynamics and how environment changes mediate gene expression to increase disease risks.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-013-1350-z) contains supplementary material, which is available to authorized users.

C. Wu · Y. Cui
Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

Y. Cui (✉)
Division of Medical Statistics, School of Public Health, Shanxi Medical University, Taiyuan 030001, Shanxi, China
e-mail: cui@stt.msu.edu

Published online: 24 August 2013

Such phenomenon that disease risk or genetic expression varies under different environment conditions is also termed phenotypic plasticity (Feinberg 2004).

G × E interactions were historically pursued by evaluating the gene effect under different environment conditions. Figure 1a shows the case of G × E interaction under two discrete environment conditions, protective and predisposing such as non-smoking and smoking. When environment conditions are measured in a continuous scale, more information is available to assess the gradient/dynamic change of genetic effect under subtle environment changes. For example, adult bone mineral density changes with age and vitamin D intake (Peacock et al. 2002). Figure 1b–d displays several scenarios where the environment mediator is measured in a continuous scale. Example of continuous environment could be age for age-related diseases such as Alzheimer, or body mass index for type 2 diabetes or hypertension. In Fig. 1b, no G × E interaction is observed since the genetic effects of the three genotypes are parallel to each other. Figure 1c shows a typical example of linear G × E interaction, while Fig. 1d displays a non-linear G × E interaction pattern assuming the *Aa* genotype is the baseline. As seen in the following section, most current G × E interaction model assumes the case displayed in Fig. 1c. Few statistical analysis has considered the case shown in Fig. 1d.

In fact, much literature work supports the view of non-linear G × E interaction. Sparrow et al. (2012) found that mutations in gene HES7 and MESP2 caused congenital scoliosis, and the risk was highly related to transient hypoxia during mice pregnancy. The rate of risk increase was non-linearly correlated with increasing hypoxic levels. Laitala et al. (2008) reported that the reaction of personal genetic effects on coffee consumption showed a non-linear relationship with age. Martinez et al. (2003) found that women carrying Gln27Glu genotype in ADRB2 gene had higher probability for obese and the obesity rate was non-linearly correlated with the amount of carbohydrate intake. Even though these empirical evidences are limited to small-scale observational studies, they underscore the importance of further exploration on non-linear G × E interaction when searching for genetic roots of complex diseases.

Within the statistical framework, G × E interactions in human diseases have been investigated mainly through model-based approaches, ranging from the standard linear model with interaction in diverse design settings, such as the case–control design, the case only design and the two-stage screening design, to more sophisticated models, such as profile likelihood-based semi-parametric models, empirical Bayesian models and Bayesian model average (reviewed in Mukherjee et al. 2012). However, as pointed out in Ma et al. (2011), the model-based regression framework generally needs strong model assumptions between genetic effects and environmental influences, which cannot be directly applied to the above mentioned empirical studies in which non-linear interaction exists.

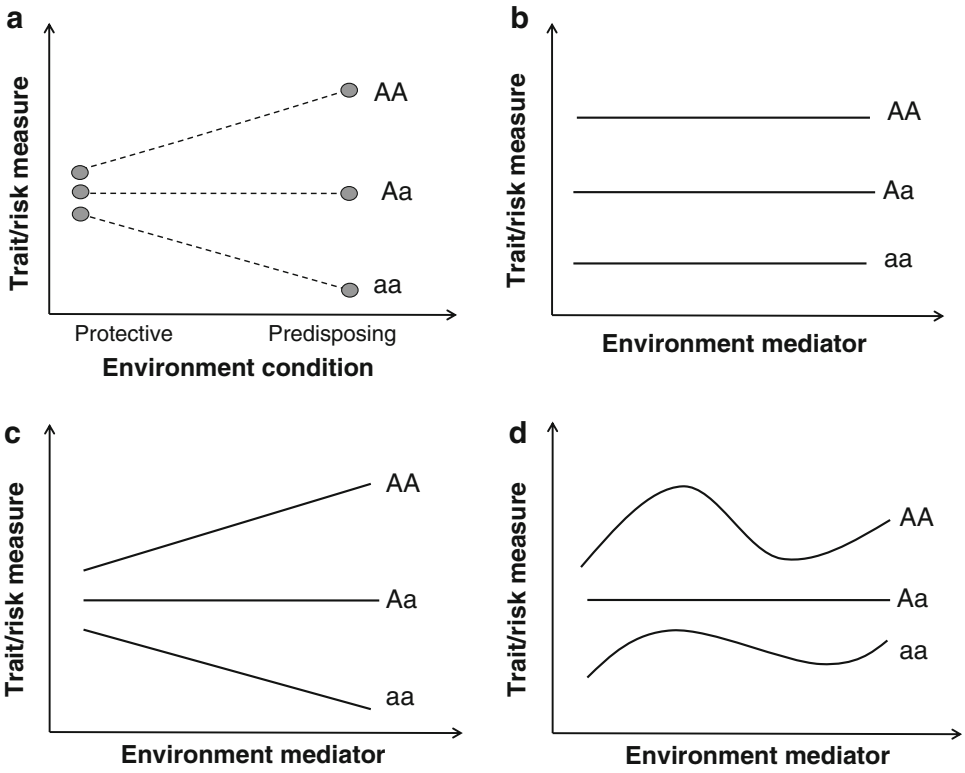

Fig. 1 Different models of gene–environment interaction: **a** the interaction of gene and environment in discrete environmental conditions; cases with **b** no G × E interaction; and **c** linear and DS non-linear G × E interactions. *AA*, *Aa* and *aa* represent three different genotypes in a gene, and environment mediator represents a continuous environment variable

In this paper, we extended the varying-coefficient (VC) model proposed in Ma et al. (2011) for continuous quantitative responses to binary disease responses. We first laid out the VC modeling framework for binary responses, with details on parameter estimation and hypothesis testing. The utility of our approach was demonstrated through extensive simulations. Finally, we applied our method to two case–control type 2 diabetes cohorts data sets, followed by discussions.

Statistical method

For a sample of n unrelated individuals collected from a population, let n_1 and n_2 be the number of affected (cases) and unaffected (controls) individuals, respectively, with $n = n_1 + n_2$. All individuals in the sample could be genotyped either based on candidate genes or on a whole genome-wide scale. Let $Y_i = 1$ if the ith individual is affected and 0 otherwise. Let G be the genetic variable which is coded as 0, 1, 2 corresponding to genotype aa, Aa and AA where allele A is the minor allele. This coding scheme assumes an additive disease model, although a genetic variant may show dominant or recessive action mode. In reality we can do a model selection to choose which one is the optimal one using AIC or BIC criterion.

Suppose in addition to the genetic variables, the disease risk is also affected by environmental factors as well as the interaction between them. Let X be the environmental variable which is measured in a continuous scale. Throughout this work, we are only interested in environment changes that display in a continuous scale (e.g., geographical location or temporal changes). Traditional analysis for G × E was commonly pursued by discretizing an environmental variable into different groups (e.g., old vs young), as shown in Fig. 1. However, we can have more information to assess the G × E relationship when a continuously measured environment factor is treated in a continuous scale. Thus, the purpose of the work is to model the genetic responses under different environmental stimuli, and further assess in what form genes respond to these changes.

For a continuous phenotype Y, the general form of an additive VC model to investigate the non-linear G × E interaction between X and G can be expressed as,

$$Y = \alpha(X) + \beta(X)G + \sigma(X)\varepsilon \tag{1}$$

where the error term ε satisfies $E(\varepsilon|X,G) = 0$ and $\mathrm{Var}(\varepsilon|X,G) = 1$. Ma et al. (2011) evaluated the performance of the model by assuming $\alpha(X) = \alpha_0 + \alpha_1 X$. The key components of the VC model lie in proper estimation of the smoothing function $\beta(X)$ and the variance function $\sigma^2(X)$, through which the effect of the genetic variant can be evaluated as a function of environment exposures. Various tests have been proposed to assess the linear or non-linear mechanisms via likelihood ratio test. When inhomogeneous variance (i.e., $\sigma(X)$ varies with X) and no parametric distribution are assumed for the error term, wild bootstrap is a common choice to assess the significance of the likelihood ratio statistic.

In human genetics, many diseases are displayed as discrete qualitative traits. The focus of this work is to extend the above model to responses that do not follow continuous distribution. In a generalized linear model setup, the relationship between the mean of a response variable Y and the independent variables (X, G) under the varying-coefficient model can be expressed as

$$E(Y|X,G) = \mu = g^{-1}\{\alpha(X) + \beta(X)G\}$$

where g is a link function. When Y is measured as counts (e.g., tumor numbers), a *log* link function can be assumed. When Y is a binary variable (i.e., affected vs unaffected), then a logit link function is commonly applied. In the later case, the logit varying-coefficient model is given by:

$$\mathrm{logit}(p) = \alpha(X) + \beta(X)G \tag{2}$$

where $p = \Pr(Y = 1|X,\ G)$. In this work, we allow the intercept function $\alpha(X)$ varies with X instead of assuming a linear structure, to make it more flexible to capture the underlying mean function when there is no genetic contribution (i.e., $\beta(X) = 0$).

If we allow $\beta(X) = \beta_1$, the logistic VC model is reduced to a logistic linear predictor model without G × E interaction (denoted as LM). If we allow $\beta(X) = \beta_1 + \beta_2 X$, the logistic VC model is reduced to a logistic linear predictor model with linear G × E interaction (denoted as LM-I), i.e.,

$$\begin{aligned}\mathrm{logit}(p) &= \alpha(X) + (\beta_1 + \beta_2 X)G \\ &= \alpha(X) + \beta_1 G + \beta_2 XG\end{aligned} \tag{3}$$

One can also put structures on the function of $\alpha(X)$. For example, we can let $\alpha(X) = \alpha_0 + \alpha_1 X$. Such a model like $\mathrm{logit}(p) = \alpha_0 + \alpha_1 X + \beta_1 G + \beta_2 XG$ is often applied in assessing G × E interactions in a typical logistic regression analysis by testing $H_0 : \beta_2 = 0$. It can also be seen that this model assumes a linear G × E interaction structure, that is, the function $\beta(X)$ is linear in X. Thus, without assuming specific structure on the linear predictors, the VC model has much flexibility to capture the underlying interaction mechanism via fitting $\beta(X)$ using smoothed nonparametric functions. The VC interaction model can be considered as a generalization to the linear interaction model.

Estimating $\beta(X)$ function

The nonparametric estimation of varying coefficients has undergone intensive investigations in the last two decades

and falls generally into three categories: the local kernel polynomial smoothing, polynomial spline, and smoothing spline (Fan and Zhang 2008; Huang et al. 2004). Huang et al. (2002) approximated the varying-coefficient functions via B-spline basis expansion. Using the B-spline technique, the authors further established the relevant asymptotic properties of the estimators, such as consistency, convergence rates and asymptotic normality. In addition, the estimation of B-spline estimators is computationally fast and numerically stable. These merits are especially important in the context of high-dimensional genetic data analysis, which make it a natural choice for us to choose when estimating the varying-coefficient functions $\alpha(X)$ and $\beta(X)$.

Let h be the degree of B-splines and N be the corresponding interior knots. Further assume that the knots are equally distributed for the B-spline basis matrix $\{\mathbf{B}_s{:}1 \le s \le (N+h+1)\}$. Ideally we can select h and N for $\alpha(X)$ and $\beta(X)$ separately using the B-spline technique when fitting each SNP variant. This process involves a search of optimal degree and knots through a list of possible combinations for both functions. This, however, could incur heavy computation burden when the estimation is done for each SNP given that the number of SNP variants to be tested could be huge. Thus, the degree h_0 and knots N_0 for $\alpha(X)$ are selected first by fitting a logistic VC model without the genetic components. Once the degree and knots for function $\alpha(X)$ are selected, they will be fixed when estimating degree and knots for function $\beta(X)$ for each SNP. The selection is done by using the Bayesian Information Criterion (BIC) criteria defined as,

$$\arg\min_{N,h} BIC(N,h) = \arg\min_{N,h} \ell(\gamma_1) + (N+h)\log(n)/n,$$

where $\ell(\gamma_1)$ refers to the log-likelihood function. A grid search for possible combinations of N and h can be done and the values corresponding to the minimum BIC are the "optimal" ones.

Once the degree and knots for $\alpha(X)$ are determined, the function $\alpha(X)$ can be estimated by $\hat{\alpha}(X) = \hat{\gamma_I}^{\mathrm{T}}\mathbf{B}_1(X) = \sum_{k=1}^{N_0+h_0+1} \hat{\gamma}_{1k}B_{1k}(x)$. The degree h_1 and the number of knots N_1 for $\beta(X)$ are also selected using the same BIC criterion defined above. The estimator for $\beta(X)$ is given by $\hat{\beta}(X) = \gamma_2{}^{\mathrm{T}}\mathbf{B}_2(X) = \sum_{k=1}^{N_1+h_1+1} \gamma_{2k}B_{2k}(x)$. Regular Newton–Raphson or Fisher scoring algorithm can be applied to estimate the parameters.

Assessing G × E interaction

Our goal is to assess if a genetic variant is sensitive to environment changes. If it does, then in what form, linear or nonlinear. For this purpose, we first propose to assess if the genetic effect is a constant by testing

$$\begin{cases} H_0^C: & \beta(\cdot) = \beta \\ H_a^C: & \beta(\cdot) \neq \beta \end{cases} \tag{4}$$

where β is an unknown constant and $\text{logit}(p) = \alpha(X) + \beta G$ is the corresponding reduced model under the null hypothesis. Under the H_0, the genetic effect is a constant and its contribution to disease risk has nothing to do with environmental changes. If we fail to reject the null, then association can be assessed via testing H_0: $\beta = 0$ by fitting the reduced model. Rejecting the null hypothesis leads to the conclusion that the G × E interaction exists. We next test the linear effect of G × E interaction by formulating,

$$\begin{cases} H_0^L: & \beta(\cdot) = \beta_1 + \beta_2 X \\ H_a^L: & \beta(\cdot) \neq \beta_1 + \beta_2 X \end{cases} \tag{5}$$

where β_1 and β_2 are unknown constants. Under the H_0, the reduced model is given by $\text{logit}(p) = \alpha(X) + \beta_1 G + \beta_2 XG$. If we fail to reject the null, then association can be assessed via testing H_0: $\beta_1 = \beta_2 = 0$ by fitting the reduced model. If the null is rejected, it indicates nonlinear G × E interaction effect and next we fit model 2 to assess genetic association.

The above tests are sequential. At each step if we fail to reject the null, we stop and fit the null model and assess the genetic effect by a likelihood ratio test or using a conditional bootstrap approach (Cai et al. 2000). When H_0^L is rejected, a nonlinear G × E interaction effect is implied and we allow the data tell the shape of the effect by fitting the above described nonparametric B-spline functions. The nonlinear effect is then assessed by testing H_0: $\beta(X) = 0$ using a likelihood ratio test which asymptotically follows a Chi-square distribution with the degrees of freedom equal the number of fitted B-spline coefficients of function $\beta(X)$, or using a conditional bootstrap method proposed in Cai et al. (2000). The bootstrap method may give more accurate result, but certainly is more time-consuming.

Simulation

The statistical behavior of the proposed approach was evaluated through extensive Monte Carlo simulations. When using B-spline functions to estimate the varying-coefficients, a uniform distribution on X is generally assumed. In real application, the environment measure (X) may not be uniformly distributed as in the type 2 diabetes data analyzed later in the paper. Instead, it is often normally distributed. To mimic real situations, we generated a continuous environment measure X^* from a normal distribution, and subsequently transformed it by $X = \Phi(\frac{X^* - \bar{X^*}}{S_{X^*}})$, to make X^* evenly distributed on the B-spline subintervals, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $\bar{X^*}$ and S_{X^*} are the sample mean

and sample standard deviation of X^*, respectively. The B-spline basis matrix was constructed on the transformed values. For a given minor allele frequency (MAF) p_A and assuming Hardy–Weinberg equilibrium, SNP genotypes AA, Aa and aa were simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$ for the three genotypes, respectively. We coded the genetic variable G_i as (2, 1, 0) corresponding to genotypes (AA, Aa, aa).

False positive control

We first evaluated the false positive control for the VC model at the nominal 0.05 level. For comparison purpose, we also reported the error rate for the linear predictor model with and without interaction. Under the null of no genetic effects, the disease phenotypes were simulated with $\text{logit}(p) = \alpha_0 + \alpha(x)$, where $\alpha(x)$ was generated via the B-spline basis function, i.e., $\alpha(x) = \sum_{k=1}^{4} \gamma_k B(x)$ for given spline coefficients $\gamma_1 = 6.162$, $\gamma_2 = 5.948$, $\gamma_3 = 3.858$, $\gamma_4 = 3.640$. The spline coefficients were obtained by fitting the real data (described later) without fitting the genetic effect. We added a constant α_0 in order to control the simulated proportion of case:control ratio to approximately 1:1 (by varying the size of α_0). A total of 10,000 simulation replicates were taken under all the combinations of sample size ($n = 500$, 1,000, 2,000) and MAF ($p_A = 0.1$, 0.3, 0.5).

The results were summarized in Fig. 2. As we can observe, the false positive rates were estimated sensibly from the simulated data. The VC model slightly overestimated the false positive rate under low allele frequency ($p_A = 0.1$). But the performance improved as MAF increases for a fixed sample size. In addition, the performance improved as sample size increased under a fixed MAF. In general, there were no significant deviations from the nominal 0.05 level for all the 3 models, except in some cases under low MAF and small sample size.

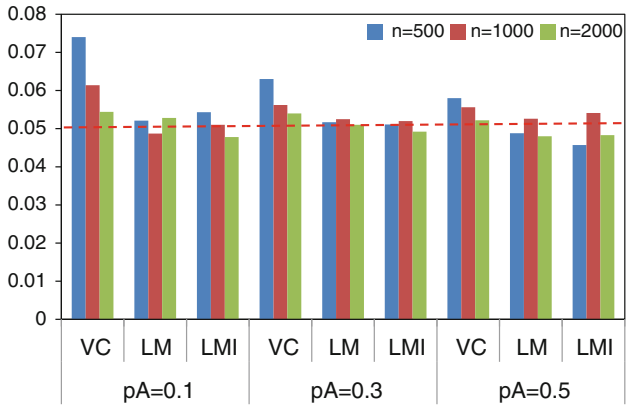

Fig. 2 The false positive rate of different models at the 0.05 level (color figure online)

Power evaluation

For given genetic effects, the disease status was simulated from a Bernoulli trial. The varying-coefficient function $\beta(\cdot)$ was estimated through $\hat{\beta}(x) \equiv \sum_{s=1}^{N_1+h_1+1} \hat{\gamma}_s B(x)$. In a typical simulation study with VC models, people generally simulate data assuming a nonlinear function such as a sin or exponential function. As SNPs do not function in such form, we simulated data according to the fit calculated from the real data to make it more realistic. Three scenarios were considered. Scenario 1 assumed that the true G × E interaction was nonlinear and the data were generated with the VC model. In scenario 2, we assumed there was no G × E interaction, while in scenario 3 we assumed linear G × E interaction. The simulated data were then analyzed using the VC, LM-I and LM models, to compare the performance of detecting significant SNP effect under model miss-specification .

For a given MAF, the data assuming nonlinear G × E interaction were generated with the following VC model,

$$\text{logit}(p_i) = \alpha_0 + \alpha(X_i) + \beta(X_i)G_i$$

where $p_i = p(Y = 1|X, G)$, and α_0 was a constant used to control the case:control ratio to make it close to 1. The varying coefficient functions $\alpha(X)$ and $\beta(X)$ were computed based upon the quadratic B-spline basis matrix with $\alpha(X) = \gamma_1{}'\mathbf{B}_1(X)$ and $\beta(X) = \gamma_2{}'\mathbf{B}_2(X)$, where $\gamma_1 = (7.287, 7.146, 3.917, 3.413)^{\mathrm{T}}$ and $\gamma_2 = (0.080, -0.460, -0.201, 0.465)^{\mathrm{T}}$ were obtained from real data fit, namely SNP rs4506565 on chromosome 10 of the Nurses' Health Study (NHS) data in GENEVA consortium (described later). The binary responses were then generated from a Bernoulli trial with case probability p_i.

The likelihood ratio test was applied to assess the significance of each test illustrated in previous section. The comparison results are shown in Fig. 3. As we expected, a common trend for the three models is that the power increases as MAF and sample size increase. Under the same sample size or MAF, the VC model always has the best power among the three, which is not surprising since the phenotypes were generated from a VC model. In addition, the LM-I model performs better than the LM model since structurally it is more close to the VC model.

We also simulated data assuming no G × E interaction using the following model,

$$\text{logit}(p_i) = \alpha_0 + \alpha(X_i) + \beta_1 G_i$$

where $\alpha(x)$ was generated from the B-spline basis function with $\alpha(X) = \gamma_0{}'\mathbf{B}_0(X)$. The spline coefficient vector was given by $\hat{\gamma}_0 = (5.977, 6.011, 3.843, 3.668)^{\mathrm{T}}$, and the genetic coefficient was set as $\beta_1 = 0.271$ (corresponding to an odds ratio of 1.3). These coefficients were obtained by

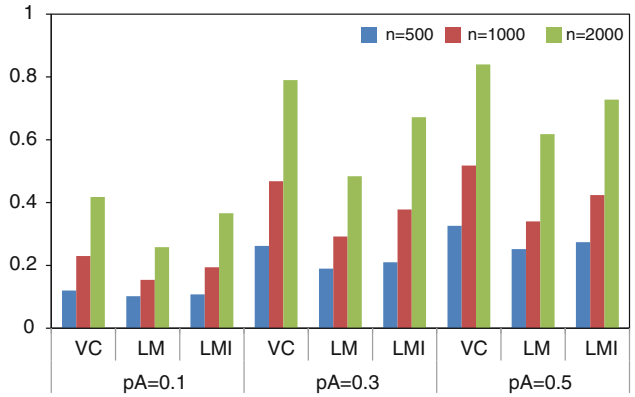

Fig. 3 The power of different models under different MAFs and sample sizes when data were generated with the VC model (color figure online)

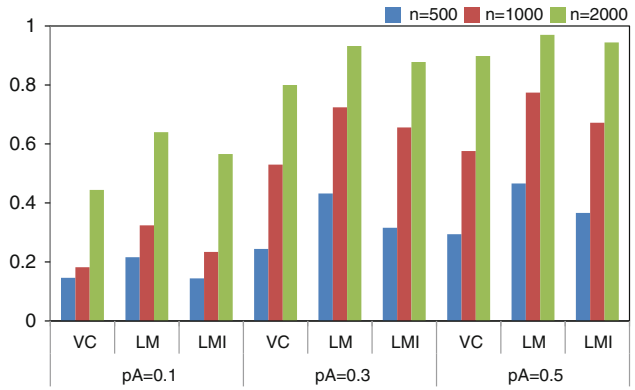

Fig. 4 The power of different models under different MAFs and sample sizes when data were generated with the LM model (color figure online)

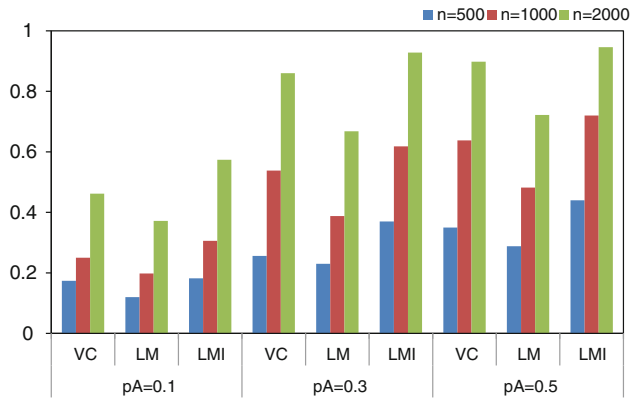

Fig. 5 The power of different models under different MAFs and sample sizes when data were generated with the LM-I model (color figure online)

fitting the real data with a linear predictor without interaction for SNP rs12255372 on chromosome 10. α_0 was used to adjust the case:control ratio as described before under different sample sizes and MAFs. The results shown in Fig. 4 demonstrate that the LM model outperforms the other two models in all the scenarios, since data were analyzed with the true data generating model. As MAF increases, the power differences among the three models diminishes for larger sample sizes. For example, the power difference among the three models is very small under $p_A = 0.5$ and $n = 2{,}000$.

The following linear interaction model (LM-I) was assumed to generate the linear G × E interaction data,

$$\text{logit}(\text{p}_i) = \alpha_0 + \alpha(X_i) + \beta_1 G_i + \beta_2 X_i G_i$$

where $\alpha(X) = \gamma_{\boldsymbol{0}}^{'} \mathbf{B}_0(X)$ with spline coefficients $\hat{\gamma}_0 = (6.358, 6.481, 4.232, 4.113)^{\text{T}}$. The genetic coefficient $\beta_1 = 0.226$ and interaction coefficient $\beta_2 = -0.787$. All the coefficients were obtained by fitting model 3 to SNP rs17537178 on chromosome 10. Figure 5 shows that the linear interaction model has the best performance among the three. In addition, the power of the VC model is more close to the linear interaction model since it is more structurally close to the linear interaction model. As sample size and MAF increase, the power difference between the VC and LM-I model vanishes quickly.

In summary, when the true G × E interaction is linear or when there is no interaction at all, the model assuming linear or constant coefficient outperforms the VC model. However, the VC model outperforms the other two when the true interaction in nonlinear. In addition, the LM or LM-I models suffer more from power loss when the underlying true interaction is nonlinear in comparison to the case when the underlying truth is linear or no interaction. This is not surprising since the B-spline estimator is consistent for large samples. Under large sample sizes, the VC model should perform similar to the LM and LM-I model. However, one has to be careful in finite samples. The simulation results suggest that one should assess the function $\beta(X)$ first before testing $\beta(X) = 0$. In practice, one can test if $\beta(X) = \beta$ or $\beta(X) = \beta_1 + \beta_2 X$, then fit the appropriate model depending on the test result.

Real data analysis

The fast increase in global prevalence of type 2 diabetes draws worldwide attentions for the disease. About 50 novel loci have been reported in association with type 2 diabetes so far (Perry et al. 2012). However, only a small proportion of disease heritability has been explained by these loci, leaving the question of how to effectively account for gene–environment interaction in the search of T2D susceptibility variants with the hope to capture the missing heritability. We applied our model to two nested case–control cohort studies of type 2 diabetes, the Nurses' Health Study (NHS) and the Health Professionals Follow-

up Study (HPFS), from the Gene, Environment Association Studies Consortium (GENVEA) (Cornelis et al. 2010). The two data sets are well-characterized cohorts of genome-wide association studies investigating a set of hypotheses about the dietary and lifestyle factors to the triggering of a series of diseases, including type 2 diabetes, for women and men. Details of the two cohorts can be found from Colditz et al. (2005) and Rimm et al. (1991). The data sets from the two cohort studies originally contain 3,391 females (NHS) and 2,599 males (HPFS) with European ancestry. After data cleaning by removing subjects with unmatched phenotypes and genotypes, excluding SNPs with MAF <0.05 and deviation from Hardy–Weinberg equilibrium, the final data contain 3,391 females (1,646 cases and 1,745 controls) with 635,748 SNPs in the NHS set and 2,570 males (1,300 cases and 1,270 controls) with 636,764 SNPs in the HPFS set.

Body mass index (BMI), calculated as the quotient between an individual's mass (kg) and the square of height (m^2), is an indicator of human obesity. It is widely recognized that the risk of type 2 diabetes could be potentially influenced by obesity condition evidenced by strong association between them for both women and men (Holbrook et al. 1989; Carey et al. 1997; Chan et al. 1994). Therefore, individual's BMI can be regarded as a type of environmental condition pivotal in evaluating the incidence of type 2 diabetes. Individuals carrying the same gene may have different risks of type 2 diabetes under different obese conditions. The phenomenon could be elucidated, at least partially, by the complicated interaction mechanism between the carrier's gene and the environment (measured by BMI). Thus, we can treat the genetic sensitivity to obese as a dynamic process which can be captured by the proposed VC model, if any.

Qi et al. (2010) assessed population stratification for the two data sets and found that the genomic inflation factors were 1.02 in NHS and 1.01 in HPFS, indicating no issue of population stratification. Therefore, we analyzed the two data sets separately without considering population structure, in order to find sex-specific genes responsible for T2D risk. Figure 6 shows the Manhattan plot of the $-\log10(p$ values) for the male data. To compare the performance of the three models (VC, LM, LM-I), we plotted all the signals at each SNP locus. It can be seen that the overall signals for the three models are quite consistent. The dashed red line corresponds to the genome-wide Bonferroni threshold (7.9E−8) and the dotted blue line corresponds to the suggestive threshold (5E−6). Table 1 tabulated SNPs that passed both threshold. Seven SNPs passed the Bonferroni threshold are marked by *. Testing constant coefficient showed that the majority of the SNPs has constant coefficients, which indicated they are not sensitive to obese condition. This also explained why the LM model gives relatively stronger signals than the other two models. Columns with P_{const} and P_{linear} showed the p values for testing H_0: $\beta(X) = \beta$ and H_0: $\beta(X) = \beta_1 + \beta_2 X$. The smaller p values for testing constant and linear coefficients in the top panel showed that the effects of those SNPs were neither constant nor linear, thus the VC model gave the strongest signals evidenced by smaller P_{vc} than P_{LM} and P_{LMI}. For example, SNP rs4635456 had $P_{\text{const}} = 9.5\text{E}{-}07$ and $P_{\text{linear}} = 0.0117$ which indicated the coefficient of this SNP is varying over BMI. Thus, fitting a VC model gave the strongest signal ($P_{\text{vc}} = 3.05\text{E}{-}06$ vs $P_{\text{LM}} = 0.6299$ and $P_{\text{LMI}} = 1.58\text{E}{-}05$).

The mid-panel in the table listed SNPs with the strongest signals fitted with the LM model. The P_{const} values for the SNPs were all large (>0.05), which suggests that $\beta(X)$ was a constant and there was no G × E interaction for these SNPs. Hence the LM model assuming no interaction gave the strongest signals. The bottom SNP in the table had the strongest signal when data were fitted with the LM-I model since we rejected constant coefficient ($P_{\text{const}} = 0.0108$) but failed to reject linear coefficient ($P_{\text{linear}} = 0.6792$).

Among the SNPs listed in the table, some have been reported in other studies. For example, transcription factor 7-like 2 (TCF7L2) is an intensively examined gene associated with a broad categories of diseases, including type 2 diabetes. The causal genetic association between SNPs of the gene and the type 2 diabetes was first reported in Grant et al. (2006) and was subsequently replicated in many ethnic groups (Jin and Liu 2008). As the SNPs in this gene are not sensitive to obese, it is not surprise that they can be identified in other studies by using methods assuming a linear relationship. But our method identified three more that show nonlinear G × E relationship, even though they did not pass the genome-wide Bonferroni threshold. We also did QQ plot and histogram of the p values for data fitted with the three models (see supplemental files for details). The p values are quite uniformly distributed and only a few show departure from the expected values (see the QQ plot). This indicates that the models have no serious inflation of false positives and the strong signals are likely to be true.

Figure 7 showed the Manhattan plot of the $-\log10(p$ values) for the female data. Even though no SNPs passed the genome-wide Bonferroni threshold, we did see stronger signals fitted by the VC model. Those SNPs that passed the suggestive threshold were listed in Table 2. Again, gene TCF7L2 does not show sign of sensitivity to obese to affect T2D risk. Gene GLI2 show sign of interaction with obese to affect T2D risk. Two SNPs in gene NRIP1 located on chromosome 21 show sign of nonlinear interaction with obese to affect T2D risk. In comparison to the male data, it is clear that SNP effects are stronger in the male population than in the female population. Moreover, the genetic effects in females are relatively more sensitive

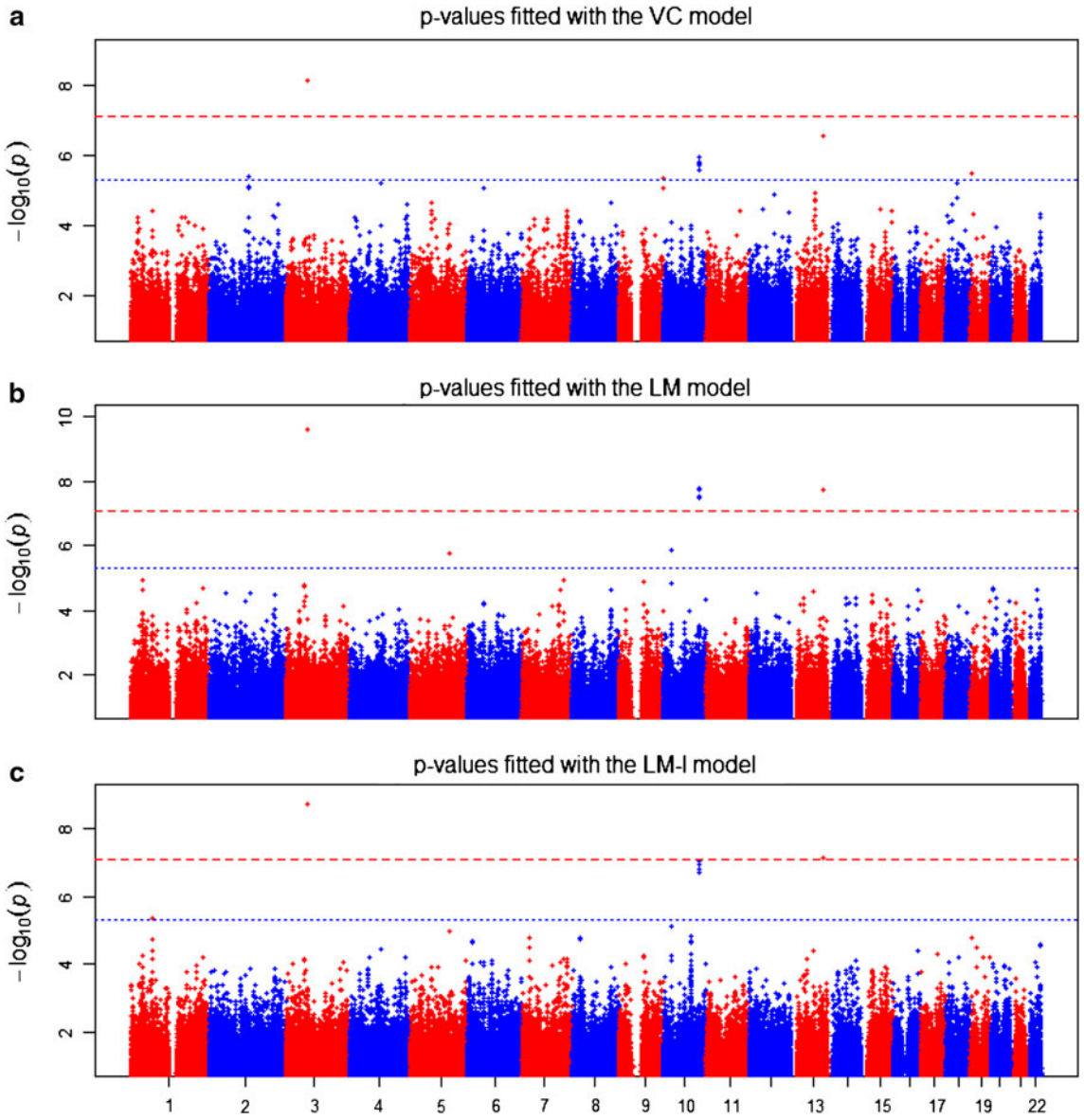

Fig. 6 The Manhattan plot of $-\log10(p$ values) for testing: **a** H_0: $\beta(X) = 0$; **b** H_0: $\beta = 0$; and **c** H_0: $\beta_1 = \beta_2 = 0$ when fitting the VC, LM and LM-I model, respectively, for the male data set (color figure online)

to obese to affect T2D risk. In summary, strong sex-specific genetic effects were observed, for example, those SNPs on chromosome 2, 3, 4, and 21.

To further demonstrate the utility of the method, we plotted the dynamic effect of SNP rs13050325 on chromosome 21 from the female data (upper panel) and SNP rs4635456 on chromosome 19 from the male data (lower panel). The two curves in the left side of Fig. 8 show the estimated dynamic genetic effect as a function of BMI fitted with the B-spline function. We can see clear nonlinear genetic effects over BMI, which indicates nonlinear interaction between BMI and the variants. The right side figures show the plot of fitted probabilities against individual BMI values corresponding to different genotypes. We coded the heterozygote as 0 in our model. This implies that the green curves in the two plots correspond to the mean fitted probability when $G = 0$. In general, the risk of T2D increases as BMI increases. This is consistent with our prior knowledge that the disease prevalence is strongly associated with body weight (McCarthy 2010).

For SNP rs13050325 on chromosome 21, the allele frequency for the minor allele G is 0.2587. For SNP rs4635456 on chromosome 19, the allele frequency for the minor allele G is 0.3771. In both cases, the overall trend for T2D risk for the baseline (corresponding to genotype AG) increased as BMI level increases (green curve). However, individuals carrying AA genotype had much higher chance to develop T2D than those carrying AG or GG genotype.

Table 1 List of SNPs with p value <5E−06 in the HPFS (male) data set

SNP ID	Gene name	Chr	P_{vc}	P_{const}	P_{linear}	P_{LM}	P_{LMI}	P_i
Fitted with VC model								
rs4635456	SEMA6B	19	**3.05E−06**	9.49E−07	0.0117	0.6299	1.58E−05	2.91E−06
rs4972250	Unknown	2	**3.99E−06**	2.21E−06	1.65E−06	0.2772	0.1982	0.1516
rs4842244	RXRA	9	**4.18E−06**	1.25E−06	2.91E−06	0.7146	0.0886	0.0299
Fitted with LM model								
rs2371765	ADAMTS9-AS2	3	6.82E−09	0.2909	–	**2.38E−10***	1.88E−09	0.8140
rs7901695	TCF7L2	10	1.49E−06	0.8638	–	**1.72E−08***	1.06E−07	0.5633
rs7991210	PCCA	13	2.80E−07	0.2234	–	**1.81E−08***	7.07E−08	0.2655
rs12243326	TCF7L2	10	1.14E−06	0.6896	–	**1.87E−08***	8.88E−08	0.3570
rs4132670	TCF7L2	10	1.64E−06	0.8560	–	**1.94E−08***	1.07E−07	0.4632
rs12255372	TCF7L2	10	1.89E−06	0.7372	–	**2.93E−08***	1.49E−07	0.4076
rs4506565	TCF7L2	10	2.66E−06	0.8546	–	**3.31E−08***	1.87E−07	0.4967
rs11013381	C10orf67	10	7.83E−05	0.8865	–	**1.32E−06**	7.74E−06	0.7106
rs6893115	Unknown	5	9.19E−05	0.8287	–	**1.79E−06**	1.11E−05	0.9266
Fitted with LMI model								
rs699253	PDE4B	1	3.93E−05	0.0108	0.6792	1.5E−04	**4.21E-06**	0.00125

Man with genotype AA had the lowest risk of conferring T2D susceptibility when BMI level was below 28 in male and below 33 in female. After the transition points, the AA genotype triggers larger effect, resulting in higher risk of T2D. The association signals for both LM and LM-I model are weaker than the one fitted with the VC model, leading to potential miss-identification of these variants. The results offered personalized preventive suggestions based upon our findings fitted with the VC model. For example, man carrying genotype AA at this SNP locus should pay more attention to control their body weight if their BMI level is above 28 to avoid the risk of T2D.

Discussion

It is broadly recognized that naturally occurring variations in most complex disease traits have a genetic basis. However, the degree of variability is believed to have a strong environmental component in addition to genetic causes for many disease traits such as obesity and type 2 diabetes (Qi and Cho 2008). Recent effort on epigenetics study reveals the importance of epigenetic modification on complex diseases (Liu et al. 2008). These epigenetic changes involve major chromatin remodeling processes such as DNA methylation and histone modification. These structural changes represent environmentally driven plasticity at the DNA level which reveals the interplay of gene-environment interaction in the regulation of phenotype and is increasingly recognized as the epigenetic basis of many complex diseases (Liu et al. 2008). Large efforts have been devoted to the exploration of epigenetic mechanisms for better understandings of the molecular machinery underlying complex diseases (Feinberg and Irizarry 2010). However, how the environment mediates epigenetic changes to affect phenotypic plasticity is still poorly understood, largely due to the lack of powerful statistical methods to dissect this complicated process.

In this article, we proposed a novel statistical method by modeling the genotypic effect on disease risk as a dynamic function of environment mediators. Our model is built upon well-studied statistical varying-coefficient model implemented with the nonparametric spline technique to estimate the varying coefficients. The model extends out previously developed method on continuous traits to a case–control population-based design. Simulation studies show dramatically improved power when the underlying genetic penetrance behaves nonlinearly under certain environmental stimulus. Our model can capture the dynamic changes of the gene functions over environmental changes, hence has particular power to tackle long-standing genetic questions regarding gene action and phenotypic plasticity (Feinberg 2004).

Our simulation studies indicate that model miss-specification is a big issue in G × E study. The power to detect genetic signals is heavily dependent upon the models to fit the data. Simple models are always the first choice due to their simplicity in interpretation. However, if they cannot capture the underlying functional mechanism, they suffer tremendously from power loss. For example, if the true genetic effect does vary nonlinearly across environmental changes, fitting a simple linear model would result in loss of power (Fig. 3). On the other hand, complex models always suffer from large degrees of freedom for testing.

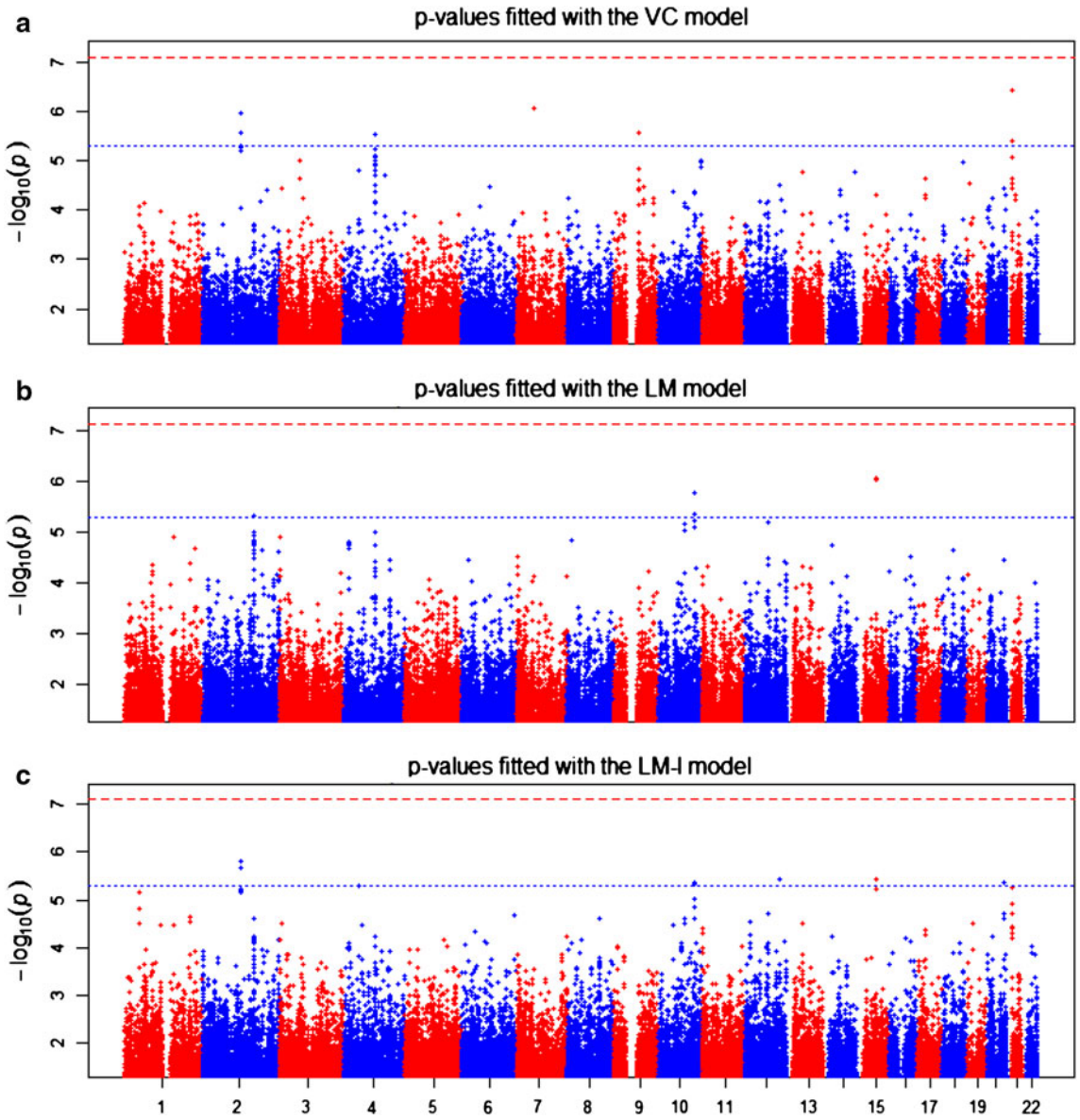

Fig. 7 The Manhattan plot of $-\log10(p$ values) for testing **a** H_0: $\beta(X) = 0$; **b** H_0: $\beta = 0$; and **c** H_0: $\beta_1 = \beta_2 = 0$ when fitting the VC, LM and LM-I model, respectively, for the female data set (color figure online)

We proposed a sequential testing procedure to assess if a simpler model fits the data better. The real data analysis confirms that this strategy works. For example, when testing constant shows that there is no G × E interaction, the model with linear predictor and without interaction term gives the smallest testing p values (see Tables 1, 2). In real data analysis, one should always start by assessing constant coefficient first, then move to test linear or varying coefficients.

We applied our model to two type 2 diabetes data sets. Cornelis et al. (2011) evaluated seven statistical models to dissect G × E interactions using the same data sets. Both Cornelis et al. (2011) and our work treated BMI as the environmental factor. Cornelis et al. (2011) claimed that specifying BMI as a continuous covariate will lead to inflated type 1 error, which has consequence in detecting increased number of false positives as the true signal. They converted the continuous environment factor BMI into a binary variable prior to further comparisons of all the seven models. However, this conversion will result in information loss, which might be the reason that there are no G × E interaction signals passing the genome-wide significance levels for all the seven models in both data sets in their analysis (Cornelis et al. 2011). In our approach, we allowed the nonlinear effect of BMI on type 2 disease [modeled by function $\alpha(X)$] rather than treated it as a linear

Table 2 List of SNPs with p value <5E−06 in the NHS (female) data set

SNP ID	Gene name	Chr	P_{vc}	P_{const}	P_{linear}	P_{LM}	P_{LMI}	P_i
Fitted with VC model								
rs13050325	NRIP1	21	**3.79E−07**	3.77E−06	0.0016	0.0062	1.23E−05	1.0E−04
rs2331061	LANCL2	7	**8.60E−07**	1.30E−06	5.26E−07	0.0703	0.1456	0.4471
rs1466042	GLI2	2	**1.10E−06**	3.48E−06	0.0389	0.0241	1.61E−06	3.37E−06
rs11145373	VPS13A	9	**2.63E−06**	8.41E−04	2.56E−04	1.27E−04	6.2E−04	0.7679
rs3775043	UNC5C	4	**2.95E−06**	0.0018	6.96E−04	6.11E−05	2.56E−04	0.4929
rs12627409	NRIP1	21	**4.00E−06**	2.38E−05	0.0441	0.0119	5.60E−06	2.38E−05
Fitted with LM model								
rs10519107	RORA	15	4.84E−05	0.8381	–	**8.52E−07**	3.72E−06	0.3802
rs809736	RORA	15	4.96E−05	0.8145	–	**9.22E−07**	5.84E−06	0.8961
rs4506565	TCF7L2	10	4.35E−05	0.4953	–	**1.69E−06**	4.66E−06	0.2018
rs7901695	TCF7L2	10	4.42E−05	0.4895	–	**1.75E−06**	4.30E−06	0.1729
rs12255372	TCF7L2	10	1.2E−04	0.5576	–	**4.47E−06**	9.73E−06	0.1543
rs4368343	Unknown	2	1.88E−04	0.7537	–	**4.75E−06**	2.53E−05	0.6320
Fitted with LMI model								
rs2677528	GLI2	2	2.63E−06	2.62E−05	0.0732	0.0064	**2.16E−06**	1.55E−05
rs7978946	Unknown	12	3.09E−05	0.0117	0.6078	1.04E−04	**3.62E−06**	0.0016
rs887370	TSHZ2	20	3.63E−05	1.45E−05	0.5868	0.4492	**4.46E−06**	9.30E−07

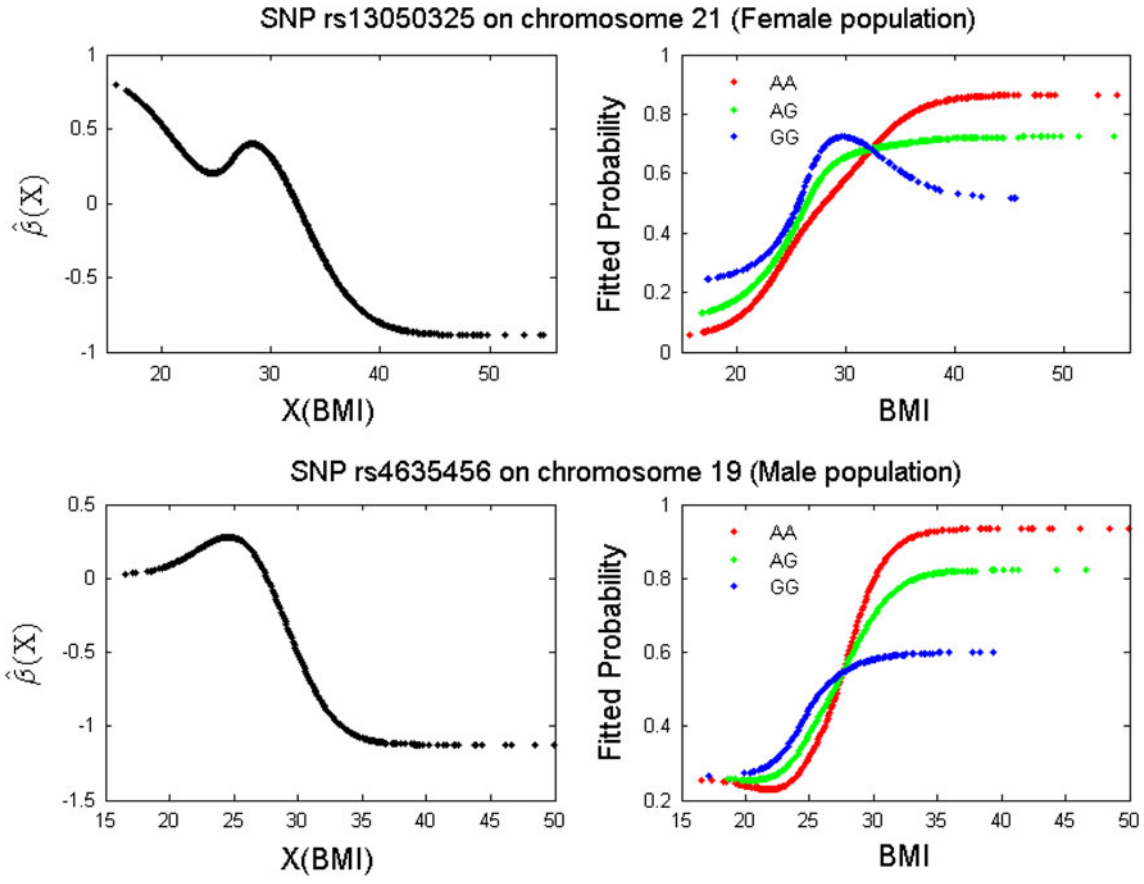

Fig. 8 The estimated varying-coefficient function and fitted probability of SNP rs13050325 and SNP rs4635456 (color figure online)

function (i.e., $\alpha_0 + \alpha_1 X$). This greatly alleviated the type 1 error inflation compared to a model fitted with a linear function in BMI (data not shown). In our analysis, several signals reached the genome-wide significance level, which is a piece of convincing evidence for keeping the continuous BMI measure as an environmental variable.

In the real data analysis, we observed strong sex-specific variants associated with T2D. There was not much overlap between genes identified in both data sets except for SNPs in gene TCF7L2. Identification of SNPs in gene TCF7L2 on the pathogenesis of type 2 diabetes has been successfully replicated from different populations (Grant et al. 2006). This information indicates the robustness of our model. In addition, we observed stronger signals in the male data evidenced by 7 SNPs from 3 genes reaching the genome-wide significance threshold (cutoff = 7.9E−8), as shown in Table 2. However, we observed stronger BMI × G interaction to affect T2D in females than in males evidenced by more nonlinear G × E interaction in the female data set (Table 2). We could miss these signals if we only focused on linear predictor models. In a recent investigation of an Italian population, Vaccaro et al. (2008) found a significantly higher average BMI levels in diabetic women. So possibly certain genes may be sensitive to high BMI level to increase T2D risk. Our model provides a testable framework to identify the underlying genetic blueprint sensitive to obese changes to affect T2D risk. The results obtained by our model can be applied to pathway or gene-set enrichment analysis to identify potential sex-specific pathways for T2D.

In this work, we generalized the VC model for continuous quantitative response to the case–control binary response. There are several ongoing work worthy of further investigation. First, the model can be easily extended to other types of phenotype data, such as count data or survival data by applying different link functions. Second, more replication studies are needed by applying our approach to type 2 diabetes of different ethnic groups to further confirm the robustness of the method. Third, it is worth noting the interesting result reported by Perry et al. (2012) that stratification on the type 2 diabetes patients based on BMI might help enrich the significance of potential susceptibility loci. We could also try to carry out analysis to test if this hypothesis leads to any new discoveries based on the VC model. Finally, our model can easily incorporate population stratification (PS) effect by first doing a principal component analysis using software such as EIGENSTRAT (Price et al. 2006), then incorporate those PCs as covariates into the model to account for the effect of PS.

Acknowledgments The authors wish to thank three anonymous referees for their constructive comments that greatly improved the manuscript. This work was partially supported by NSF grant DMS-1209112 and by National Natural Science Foundation of China grant 31371336. Funding support for the GWAS of Gene and Environment Initiatives in Type 2 Diabetes was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1, through dbGaP accession number phs000091.v2.p1.

Conflict of interest The authors declare no conflict of interest.

References

Cai Z, Fan J, Li R (2000) Efficient estimation and inferences for varying-coefficient models. J Am Stat Assoc 95:888–902

Carey VJ, Walters EE, Colditz GA, Caren G, Solomon et al (1997) Body fat distribution and risk of non-insulin-dependent diabetes mellitus in women. The Nurses' Health Study. Am J Epidemiol 145:614–619

Chan JM, Rimm EB, Colditz GA, Stampfer MJ, Willett WC (1994) Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. Diabetes Care 17:961–969

Colditz GA, Hankinson SE (2005) The Nurse's Health Study: lifestyle and health among women. Nat Rev Cancer 5:388–396

Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC et al (2010) The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 34:364–372

Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P (2011) Gene–environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol 175:191–202. doi:10.1093/aje/kwr368

Fan J, Zhang W (2008) Statistical methods with varying coefficient models. Stat Interface 1:179–195

Feinberg AP (2004) Phenotypic plasticity and the epigenetics of human disease. Nature 447:433–440

Feinberg AP, Irizarry RA (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci USA 107:1757–1764

Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A et al (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38:320–323

Holbrook TL, Barrett-Connor E, Wingard DL (1989) The association of lifetime weight and weight control patterns with diabetes among men and women in an adult community. Int J Obes 13:723–729

Huang JZ, Wu CO, Zhou L (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89:111–128

Huang J, Wu C, Zhou L (2004) Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Stat Sin 14:763–788

Jin T, Liu L (2008) The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus. Mol Endocrinol 22:2383–2392

Liu L, Li Y, Tollefsbol TO (2008) Gene–environment interactions and epigenetic basis of human diseases. Curr Issues Mol Biol 10:25–36

Laitala VS, Kaprio J, Silventoinen K (2008) Genetics of coffee consumption and its stability. Addiction 103:2054–2061

Ma SJ, Yang LJ, Romero R, Cui YH (2011) Varying coefficient model for gene–environment interaction: a non-linear look. Bioinformatics 27(15):2119–2126

Gamboa-Melndez MA, Huerta-Chagoya A, Moreno-Macas H, Vzquez-Crdenas P et al (2012) Contribution of common genetic

variation to the risk of type 2 diabetes in the Mexican Mestizo population. Diabetes 61:3314–3321. doi:10.2337/db11-0550

Martinez JA, Corbalan MS, Sanchez-Villegas A et al (2003) Obesity risk is associated with carbohydrate intake in women carrying the Gln27Glu beta2-adrenoceptor polymorphism. J Nutr 133:2549–2554

McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. N Engl J Med 363:2339–2350

Mukherjee B, Ahn J, Gruber SB, Chatterjee N (2012) Testing gene–environment interaction in large-scale case–control association studies: possible choices and comparisons. Am J Epidemiol 175:177–190

Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ (2013) Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. Hum Genet 132:495–508. doi:10.1007/s00439-012-1258-z

Peacock M, Turner CH, Econs MJ, Foroud T (2002) Genetics of osteoporosis. Endocr Rev 23:303–326

Perry JRB, Voight BF, Yengo L, Amin N, Dupuis J et al (2012) Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. PLoS Genet 8(5):e1002741. doi:10.1371/journal.pgen.1002741

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association. Nat Genet 38:904–909

Qi L, Cho YA (2008) Gene–environment interaction and obesity. Nutr Rev 66:684–694

Qi L, Cornelis MC, Kraft P et al (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 19:2706–2715

Rimm EB, Giovannucci EL, Willett WC, Colditz GA, Ascherio A, Rosner B, Stampfer MJ (1991) Prospective study of alcohol consumption and risk of coronary disease in men. Lancet 338:464–468

Sparrow DB et al (2012) A mechanism for gene–environment interaction in the etiology of congenital scoliosis. Cell 149:295–306

Vaccaro O, Boemi M, Cavalot F, De Feo P, Miccoli R, Patti L, Rivellese AA, Trovati M, Ardigo D, Zavaroni I (2008) The clinical reality of guidelines for primary prevention of cardiovascular disease in type 2 diabetes in Italy. Atherosclerosis 198:396–402

Zimmet P, Alberti KGMM, Shaw J (2001) Global and societal implications of the diabetes epidemic. Nature 414:782–787. doi:10.1038/414782a