G3, 2021, 11(12), jkab278
DOI: 10.1093/g3journal/jkab278
Advance Access Publication Date: 6 August 2021
Neurogenetics

Smooth-threshold multivariate genetic prediction incorporating gene–environment interactions

Masao Ueki[1,*] and Gen Tamiya[2,3,4]; for Alzheimer's Disease Neuroimaging Initiative[†]

[1]School of Information and Data Sciences, Nagasaki University, Nagasaki 852-8521, Japan,
[2]Tohoku University Graduate School of Medicine, Sendai, Miyagi 980-8575, Japan,
[3]Statistical Genetics Team, RIKEN Center for Advanced Intelligence Project, Chuo-ku, Tokyo 103-0027, Japan, and
[4]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Miyagi 980-8573, Japan

*Corresponding author: School of Information and Data Sciences, Nagasaki University, 1-14 Bunkyo-Machi, Nagasaki 852-8521, Japan. Email: uekimrsd@nifty.com
[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf (Accessed: 2021 August 31)

Abstract

We propose a genetic prediction modeling approach for genome-wide association study (GWAS) data that can include not only marginal gene effects but also gene–environment (GxE) interaction effects—*i.e.*, multiplicative effects of environmental factors with genes rather than merely additive effects of each. The proposed approach is a straightforward extension of our previous multiple regression-based method, STMGP (smooth-threshold multivariate genetic prediction), with the new feature being that genome-wide test statistics from a GxE interaction analysis are used to weight the corresponding variants. We develop a simple univariate regression approximation to the GxE interaction effect that allows a direct fit of the STMGP framework without modification. The sparse nature of our model automatically removes irrelevant predictors (including variants and GxE combinations), and the model is able to simultaneously incorporate multiple environmental variables. Simulation studies to evaluate the proposed method in comparison with other modeling approaches demonstrate its superior performance under the presence of GxE interaction effects. We illustrate the usefulness of our prediction model through application to real GWAS data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Keywords: genetic prediction; gene–environment interaction; smooth thresholding

Introduction

Although discovery of genetic risk factors for disease is an important goal of genome-wide association studies (GWAS), predicting disease development or related traits is an important task for applying GWAS results in precision medicine. Many researchers have explored algorithms for accurate genetic prediction based on GWAS data with a large number of single-nucleotide polymorphisms (SNPs) (Evans *et al.* 2009; Purcell *et al.* 2009; Yang *et al.* 2011; Chatterjee *et al.* 2013; de Los Campos *et al.* 2013; Dudbridge 2013; Makowsky *et al.* 2013; Maier *et al.* 2015; Moser *et al.* 2015; Vilhjálmsson *et al.* 2015; Privé *et al.* 2019), but no model has been found that performs universally well with all data, and performance is highly dependent on the data-generating mechanism (Cherlin *et al.* 2018). Popular models are linear in the variants (or SNPs), such as Purcell's gene score (Purcell *et al.* 2009) and genomic best linear unbiased prediction (BLUP) (Yang *et al.* 2011). As an alternative, we developed a statistical method for genetic prediction modeling called smooth-threshold multivariate genetic prediction (STMGP) (Ueki and Tamiya 2016), and Takahashi *et al.* (2020) recently demonstrated that the performance of STMGP was superior to that of other genetic prediction methods for predicting status of depression with actual GWAS data. STMGP is a sparse modeling method based on a multiple linear regression model such as the lasso (Tibshirani 1996) or the elastic net (Zou and Hastie 2005), and it is able to account for the ultrahigh dimensionality of the $p \gg n$ situation by filtering variants based on the corresponding marginal-effect P-values calculated from univariate regressions arising from a genome-wide scan. Sparseness is achieved by ignoring irrelevant variants; the corresponding regression coefficient estimates are set to zero as a result of shrinkage based on the strength of the marginal effect through the smooth-threshold estimating equations developed by Ueki (2009). STMGP also automatically tunes the prediction model by a C_p-type model selection criterion (as with the Akaike information criterion (Akaike 1973)), where the tuning parameter corresponds to the cutoff or threshold value for the marginal P-values that determines which effects to filter. The proposed C_p-type criterion based on Stein's unbiased risk estimation (SURE, Stein 1981; Ye 1998; Efron 2004) has a closed-form expression and is a computationally efficient alternative to cross-validation that is often used to choose a P-value cutoff in the genetic prediction context (Purcell *et al.* 2009; Warren *et al.* 2013).

Received: June 20, 2021. **Accepted:** July 12, 2021

Recent advances in data platforms now make it possible to integrate feature variables other than variants, such as those associated with lifestyle, clinical variables, imaging, etc. The simplest integration is to enter everything as an additive term in a multiple linear regression model as implemented in Ueki and Tamiya (2016) and Takahashi *et al.* (2020). While such an additive modeling approach is simple and straightforward, there may be cases where other approaches are more appropriate. One example is gene–environment (GxE) interaction, which has received attention recently as one potential candidate to unveil the missing heritability problem (Maher 2008; Manolio *et al.* 2009; Manolio 2013). With GxE interaction, the model to be estimated is no longer simply additive; rather, it involves terms that are multiplicative in the covariates. Many investigations have aimed at discovering genetic factors that contribute to GxE interactions in disease risk (Kraft *et al.* 2007; Kooperberg and LeBlanc 2008; Hamza *et al.* 2011; Ober and Vercelli 2011; Aschard *et al.* 2012; Sung *et al.* 2014; Kraft and Aschard 2015, 2015; Sung *et al.* 2016; Gauderman *et al.* 2017; Khoury 2017; McAllister *et al.* 2017; Ritchie *et al.* 2017; Moore *et al.* 2018; Osazuwa-Peters *et al.* 2020): the approach using GWAS data is sometimes called a genome-wide environment interaction study (GWEIS) (Meijsen *et al.* 2018; Arnau-Soler *et al.* 2019; Ueki *et al.* 2019). The need for GxE interactions depends on the data and target traits, but as with variant discovery, it would be beneficial to have a model for genetic prediction also that can incorporate GxE interactions (Aschard 2016). However, currently the number of such studies is very limited, especially with respect to human disease prediction.

To address this issue, we present a straightforward extension of our STMGP method to allow incorporation of GxE interaction effects for building a genetic prediction model using large-scale genome-wide SNP data in conjunction with environmental variables. The proposed method can incorporate multiple environmental variables. The STMGP method requires as input the marginal association *P*-values from univariate regression models for each individual variant. This requirement implies that GxE interaction can be fit directly in the STMGP framework if it is expressed in a univariate regression model. The standard univariate GxE interaction model for variant j in n samples is

$$y_i = \mu_i + \epsilon_i = \beta_{0j} + \beta_{1j}E_i + \beta_{2j}G_{ij} + \beta_{3j}E_iG_{ij} + \epsilon_i,$$

where $i = 1, \ldots, n$. This model contains three terms: E_i, G_{ij}, and E_iG_{ij}. Here, y_i is the response variable, μ_i is the conditional mean of y_i, E_i is the environmental variable, G_{ij} is the jth variant ($j = 1, \ldots, p$), p is the number of all variants, ϵ_i is the error variable, and β_{0j}, β_{1j}, β_{2j}, and β_{3j} are the corresponding regression coefficients. In general, removing either E_i or G_j will change the regression coefficient estimate of the GxE interaction term (see Appendix for additional discussion). In this sense, the three terms—E_i, G_{ij}, and E_iG_{ij}—are considered one set, meaning that the GxE interaction effects cannot be represented by a univariate model. To overcome this issue, we propose a simple approximation by a univariate regression model (the rationale is given in the "Materials and Methods" section),

$$y_i = \beta_{0j} + \beta_{1j}\tilde{E}_i + \beta_{3j}\tilde{E}_iG_{ij} + \epsilon_i,$$

in which $\tilde{E}_i$ is the centered value of E_i, i.e. $\tilde{E}_i = E_i - \overline{E}$ with $\overline{E}$ the sample mean of $E_1, \ldots, E_n$. In words, $\beta_{2j}G_{ij}$ is simply removed from the standard model and $\tilde{E}_i$ is used instead of E_i. As a result of this approximation, a one-to-one correspondence is made between the regression coefficient β_{3j} and the single predictor variable E_iG_{ij}. Thus, the STMGP method can now incorporate the GxE interaction directly.

Materials and Methods

We use vector and matrix notation. Let $y = (y_1, \ldots, y_n)^T$, $\mu = (\mu_1, \ldots, \mu_n)^T$, $E = (E_1, \ldots, E_n)^T$, and $G_j = (G_{1j}, \ldots, G_{nj})^T$ ($j = 1, \ldots, p$). We first briefly explain the STMGP framework (Ueki and Tamiya 2016), then we present our proposed approach.

STMGP framework

Consider the linear multiple regression model, $y = \mu + \epsilon$, where $\mu = X\beta$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is the error vector, X is an nxp-dimensional design matrix, and β is the corresponding vector of p regression coefficients. In application to GWAS data without GxE interactions, we set $X = (G_1, \ldots, G_p)$. Note that p is much larger than n in typical GWAS data—*i.e.*, $p \gg n$. Sparse modeling in which some of the regression coefficients are set to zero is often used in GWAS (Hoggart *et al.* 2008; Ayers and Cordell 2010; Abraham *et al.* 2013; Lello *et al.* 2018; Privé *et al.* 2019). If disease-susceptibility SNPs show relatively large marginal signals, marginal association screening effectively reduces the dimensionality. The polygenic score, including the gene score method (Purcell *et al.* 2009) and its multivariate generalization (Warren *et al.* 2013), uses upper-ranked SNPs with marginal association as predictors to build the prediction model. The former uses independent SNPs after pruning on the basis of LD (linkage disequilibrium), which means that LD is not modeled.

The STMGP method (Ueki and Tamiya 2016) is a variant of the multivariate gene score method (Warren *et al.* 2013), which is essentially the multiple regression model for the upper-ranked SNPs, and it accounts for correlations among SNPs by not including LD-based pruning. Let $T_j(y, X)$ denote a test statistic for marginal association that takes a nonnegative value. Examples of $T_j(y, X)$ include the squared Pearson's correlation and the F statistic. Let $t > 0$ be a cutoff value for $T_j(y, X)$ defining inclusion of SNPs. The cutoff value t corresponds to a quantile of the null distribution of $T_j(y, X)$, as in hypothesis testing. The linear multiple regression after marginal association screening uses X_j satisfying $T_j(y, X) > t$ in the model. Without loss of generality, assume that a large value of $T_j(y, X)$ indicates stronger marginal association. Multiple regression after marginal association screening can be expressed by

$$\begin{gathered}\hat{\mu} = X\hat{\beta}, \\ \hat{\beta} = \begin{pmatrix} \hat{\beta}_{\mathcal{A}} \\ \hat{\beta}_{\mathcal{A}^c} \end{pmatrix} = \begin{pmatrix} (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T y \\ 0 \end{pmatrix}, \\ \mathcal{A} = \{j : T_j(y, X) > t\},\end{gathered} \tag{1}$$

where $X_{\mathcal{A}} = (X_j)_{j \in \mathcal{A}}$ and $\mathcal{A}^c$ indicates the complement set of $\mathcal{A}$. Note that the above procedure is similar to sure independence screening (Fan and Lv 2008), which uses predictor variables that are upper-ranked in marginal association analyses. The procedure (1) is feasible for $p \gg n$ data and is useful in building a predictive model. In view of the normal equations, it can be seen that $\hat{\beta}$ in (1) satisfies, for $j = 1, \ldots, p$,

$$(1-\hat{D}_j)\{\overset{T}{X}_j(X\hat{\beta}-y)\}+\hat{D}_j\hat{\beta}_j=0, \quad (2)$$

or, in vector form,

$$(I_p-\hat{D})\{X^T(X\hat{\beta}-y)\}+\hat{D}\hat{\beta}=0,$$

where $\hat{D}_j=1\{T_j(y,X)\le t\}$, where $1\{\cdot\}$ denotes the indicator function, $\hat{D}=\mathrm{diag}(\hat{D}_j:j)$, and I_p is the p-dimensional identity matrix. Obviously, for $j\in\mathcal{A}^c$, $\hat{D}_j=1$ and (2) reduces to $\hat{\beta}_j=0$, *i.e.*, a sparse solution; for $j\in\mathcal{A}$, $\hat{D}_j=0$ and the above normal equations reduce to $X_A^T(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}-y)=0$ because $\hat{\beta}_{\mathcal{A}^c}=0$. These are the normal equations for an ordinary least squares regression with design matrix $X_{\mathcal{A}}$. The resulting prediction process forms $\hat{\mu}(y)=X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}=X_{\mathcal{A}}(X_{\mathcal{A}}^TX_{\mathcal{A}})^{-1}X_{\mathcal{A}}^Ty$, which is discontinuous in y due to the thresholding induced by $\hat{D}_j$.

The main innovative idea in STMGP is to replace the discontinuous thresholding $\hat{D}_j$ in (2) with a smooth thresholding using the smooth-threshold estimating equations proposed by Ueki (2009). Following Ueki (2009), $\hat{D}_j=1\{T_j(y,X)\le t\}$ is replaced by an adaptive lasso smooth-thresholding function

$$\check{D}_j=\min[1,\{t/T_j(y,X)\}^{\frac{1+\gamma}{2}}], \quad (3)$$

where $\gamma>0$ is a tuning parameter. This smooth-thresholding function is chosen so as to be identical to the adaptive lasso estimator under the simplest least squares regression of $y=\beta+\epsilon$ (Ueki 2009). If $T_j(y,X)\le t$ (or $j\in\mathcal{A}^c$), $\check{D}_j=1$, producing a zero-valued regression coefficient; if $T_j(y,X)>t$ (or $j\in\mathcal{A}$), $\check{D}_j<1$ producing a nonzero regression coefficient. Therefore, the condition for a sparse solution with $\check{D}_j$ is the same as that with $\hat{D}_j$. Note that $\check{D}_j$ is monotonically decreasing in $T_j(y,X)$, so regression coefficients having large $T_j(y,X)$ are penalized to a lesser extent than those having small $T_j(y,X)$.

For a given screening cutoff value $t>0$, which gives a SNP set $\mathcal{A}=\{j:T_j(y,X)>t\}$, the estimates of the p regression coefficients are

$$\check{\beta}=\begin{pmatrix}\check{\beta}_{\mathcal{A}}\\ \check{\beta}_{\mathcal{A}^c}\end{pmatrix}=\begin{pmatrix}\{(I_{|\mathcal{A}|}-\check{D}_{\mathcal{A}})(\overset{T}{X}_{\mathcal{A}}X_{\mathcal{A}}+\lambda I_{|\mathcal{A}|})+\tau\check{D}_{\mathcal{A}}\}^{-1}(I_{|\mathcal{A}|}-\check{D}_{\mathcal{A}})\overset{T}{X}_{\mathcal{A}}y\\ 0\end{pmatrix}, \quad (4)$$

where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. The non-negative tuning parameters γ and τ are set to 1 and $n/\sqrt{\log n}$, respectively, following previous studies (Ueki and Tamiya 2016; Takahashi *et al.* 2020), and $\lambda>0$ is a small constant to avoid singularity of $X_{\mathcal{A}}^TX_{\mathcal{A}}$. The corresponding prediction of y_i is then $\check{\mu}_i(y)=X_i^T\check{\beta}$, where $\check{D}_j$ is an adaptive lasso smooth-thresholding function defined as $\check{D}_j=\min[1,\{t/T_j(y,X)\}^{\frac{1+\gamma}{2}}]$. Since $\check{D}_j=1$ if and only if $T_j(y,X)\le t$, the screened set $\mathcal{A}$ with $\check{D}_j$ is the same as that with $\hat{D}_j=1\{T_j(y,X)\le t\}$. It can be seen that $\check{D}_j$ replaces the discontinuous screening process $\hat{D}_j$ by a continuous function. As a result, $\check{\mu}_i(y)$ turns out to be continuous in y, enabling stable model selection (Breiman 1996).

According to Ueki (2009) and Ueki and Tamiya (2016), the regression coefficients for the screened set in (4) can equivalently be considered as the solution of the generalized ridge regression with loss $||y-X_{\mathcal{A}}\beta_{\mathcal{A}}||^2+\sum_{j\in\mathcal{A}}\beta_j^2W_j$, in which $W_j=\lambda+\tau\check{D}_j/(1-\check{D}_j)$. The ridge weight for each predictor variable, W_j, represents the uncertainty of the marginal association screening. If the marginal association is very weak, $\check{D}_j\approx1$ and W_j is large, and the corresponding regression coefficient is strongly shrunken toward zero. If the marginal association is strong, $\check{D}_j\approx0$ and $W_j\approx\lambda$, and the corresponding regression coefficient is less penalized. Continuity due to the smooth thresholding also allows computation of a C_p-type model selection criterion using SURE. The C_p-type criterion enables a computationally efficient choice of optimal P-value cutoff from the perspective of model selection. Details are provided in the Supplementary Material of Ueki and Tamiya (2016). We now outline the STMGP algorithm for $X=(G_1,\ldots,G_p)$.

Outline of the STMGP algorithm

Step 1. Perform single-SNP association analysis for p SNPs with a univariate model for each SNP.

Step 2. Retain SNPs whose single-SNP association P-value is less than $\alpha_{\max}$.

Step 3. Fix $\gamma=1$ and $\tau=n/\sqrt{\log n}$, and select an optimal α from candidate values in $[\alpha_{\min},\alpha_{\max}]$ by minimizing the C_p-type criterion:

$$C(\alpha)=\sum_{i=1}^{n}\{y_i-\check{\mu}_i(\alpha)\}^2+2\hat{\sigma}^2\mathrm{GDF}(\alpha).$$

Step 4. Compute $\check{\beta}$ in (4) by using the selected α in Step 3.

Here, $\check{\mu}_i(\alpha)$ denotes the predicted value for the ith subject at the P-value threshold α corresponding to the test statistic threshold t; $\alpha_{\max}$ is the maximum P-value in the search, which is set to make the expected number of screened SNPs to be on the order of n in practice; $\hat{\sigma}^2$ is an error variance estimate; and $\mathrm{GDF}(\alpha)$ denotes the generalized degrees of freedom (Ye 1998; Efron 2004). The univariate model for the jth variant G_j ($j=1,\ldots,p$) in Step 1 is

$$\mu_{01}=1_n\beta_{0j}+G_j\beta_{1j}. \quad (5)$$

Step 3 outputs estimates of regression coefficients, $\check{\beta}_0,\check{\beta}_1,\ldots,\check{\beta}_p$, for the intercept and each variant, which allows computation of the prediction model in an additive form. Some of the regression coefficients $\check{\beta}_1,\ldots,\check{\beta}_p$ can be exactly zero (*i.e.*, sparsity). The predicted value for a new individual who has variants $(G_j^\star)_{j=1,\ldots,p}$ can be calculated as $\check{\beta}_0+\sum_{j=1}^{p}G_j^\star\check{\beta}_j$. The above method assumes a linear regression model for a quantitative phenotype. For a binary phenotype, a logistic regression model is used.

Incorporating GxE interactions with univariate regression approximation

In what follows, we describe our procedure to incorporate GxE interactions into the STMGP framework. Consider the standard GxE interaction model for the jth variant G_j and an environmental variable E,

$$\mu_{0123}=1_n\beta_{0j}+E\beta_{1j}+G_j\beta_{2j}+(G_j{}^\circ E)\beta_{3j}, \quad (6)$$

where $^\circ$ denotes the Hadamard product—*i.e.*, the ith element of $(G_j{}^\circ E)$ is given by $G_{ij}E_i$. As seen in Steps 1 and 2 of the STMGP algorithm, because the STMGP framework requires input of multiple

predictors that pass a marginal association P-value threshold from each univariate regression model, the above GxE interaction model does not directly fit the STMGP framework due to there being two regression coefficients—β_{2j} and β_{3j}—that associate with G_j. For example, if β_{2j} is highly significant but β_{3j} is not, it is uncertain whether we may include only G_j, because β_{2j} differs from the regression coefficient of G_j in the univariate regression model without interaction term $(G_j{}^{\circ}E)$. In contrast, if β_{3j} is highly significant but β_{2j} is not, then it is unclear whether we need $(G_j{}^{\circ}E)$ only, for the same reason. Furthermore, including both $(G_j{}^{\circ}E)$ and G_j might reduce predictive power by increasing the number of predictors included: in other words, the curse of dimensionality.

We propose a simple approximation to the above GxE interaction model by using a univariate regression model to eliminate these complications. To this end, we assume independence between E and each G_j. Such assumption is sometimes made in the literature on GxE interaction (Chatterjee and Carroll 2005; Mukherjee and Chatterjee 2007), and it is reasonable for many real GWAS data as the majority of variants have small marginal effects on environmental factors. Our proposed method (the main result) is simply to use the following univariate regression model instead of (6):

$$\mu_{013} = 1_n\beta_{0j} + \tilde{E}\beta_{1j} + (G_j{}^{\circ}\tilde{E})\beta_{3j}, \tag{7}$$

in which $\tilde{E}$ is the centered E as defined previously. In the Appendix we show that, under independence between G_j *and* E, the least squares estimate of the regression coefficient of $(G_j{}^{\circ}E)$ in (6) is approximated by that of $(G_j{}^{\circ}\tilde{E})$ in (7). This implies a one-to-one correspondence between the effects of the regression coefficient of $(G_j{}^{\circ}E)$ in (6) and that of the single predictor $(G_j{}^{\circ}\tilde{E})$. As a consequence, the STMGP framework can be directly applied by setting the following design matrix with $2p$ predictors:

$$X = (G_1, \ldots, G_p, G_1{}^{\circ}\tilde{E}, \ldots, G_p{}^{\circ}\tilde{E}).$$

If we have m environmental variables, $E_1, \ldots, E_m$, we may set

$$X = (G_1, \ldots, G_p, G_1{}^{\circ}\tilde{E}_1, \ldots, G_p{}^{\circ}\tilde{E}_1, \ldots, G_1{}^{\circ}\tilde{E}_m, \ldots, G_p{}^{\circ}\tilde{E}_m),$$

which has $(1+m)p$ predictors. To implement this proposal, we simply include an additional procedure into Steps 1 and 2 above. The following is the modification to include m environmental variables.

Steps 1 and 2 of STMGP algorithm modified to incorporate GxE interactions with m environmental variables $E_1, \ldots, E_m$

Step 1': Perform single-SNP association analysis for each of the p SNPs with a univariate model for each variant, and perform SNPx$\tilde{E}_k$ interaction analysis for each of the p SNPs and $\tilde{E}_k$ with the model (7) ($k = 1, \ldots, m$), where $\tilde{E}_k = E_k - \bar{E}_k 1_n$ with $\bar{E}_k$ the sample mean of E_k.

Step 2': Screen (retain) SNPs on the basis of single-SNP association P-values, and screen SNP–environmental variable pairs on the basis of SNPx$\tilde{E}_k$ interaction P-values ($k = 1, \ldots, m$) at $\alpha_{\max}$.

The above steps are easily performed with PLINK (Purcell *et al.* 2007; Chang *et al.* 2015), as follows. Prepare the centered environmental variable in a covariate file, say environment.cov. Then, the PLINK command option is –linear –covar environment.cov –interaction –parameters 1,2,3 –tests 1,3. It is also possible to include additional covariates. We have implemented the above algorithm in our STMGP package. We have also implemented a prediction model for binary traits with a logistic regression model based on the method developed in Ueki and Tamiya (2016).

Simulation study

To examine the performance of the proposed method, we conducted simulation studies based on real SNP-GWAS data analogous to those of Takahashi *et al.* (2020). We used an ADNI-GWAS dataset obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The goal of the ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org (Accessed: 2021 August 31). The ADNI is an ongoing, longitudinal study with the primary purpose being to explore the association of genetic and neuroimaging information with late-onset Alzheimer's disease. The study investigators recruited subjects older than 65 years of age comprising about 400 subjects with MCI, about 200 subjects with AD, and about 200 healthy controls. Each subject was followed for at least 3 years. During the study period, the subjects were assessed with MRI measures and psychiatric evaluation to determine the diagnostic status at each time point.

The ADNI-GWAS data were obtained from 818 DNA samples of ADNI1 participants by using the Illumina Human 610-Quad genotyping array (Shen *et al.* 2014). The data initially included 620,901 SNPs. We included the *apolipoprotein E* (*APOE*) SNPs rs429358 and rs7412 in our analysis. We used data from 684 non-Hispanic Caucasian samples after we excluded one pair showing cryptic relatedness (revealed by the PLINK pairwise $\hat{\pi}$ statistic being greater than 0.125) (Purcell *et al.* 2007), and we excluded subjects whose reported sex did not match the sex inferred from X-chromosome SNPs. We then applied further quality control measures by excluding SNPs with missing genotype rate > 0.1, Hardy–Weinberg equilibrium test P-value $< 10^{-6}$, and MAF $< 5\%$; the total number of remaining SNPs was 528,984, which is the value of P for this analysis.

For the 684 individuals, given that the above real genotype data remain fixed, we artificially generated a quantitative trait, which was used as a target variable to be predicted. We also simulated two environmental variables (sex, E_1, and years of education, E_2) as follows. E_1 was generated from a Bernoulli distribution with success probability 0.5. E_2 was generated from a standard normal distribution. Both variables were standardized to have mean zero and variance 1 in the generated sample. First, we denote by p_0 the number of causal variants for the main effects of genes, GxE$_1$ effects, and GxE$_2$ effects; note that the p_0 variants of each type are not the same. The corresponding $3p_0$ regression coefficients, β_j^* ($j = 1, \ldots, 3p_0$), were generated from pre-specified distributions. Specifically, the first p_0 regression coefficients were generated independently and identically from a normal, NEG2 (normal–exponential–gamma with shape parameter 2), or Laplace distribution with mean zero and variance h^2_G; the second p_0 regression coefficients were generated independently and identically from a normal, NEG2, or Laplace distribution with mean zero and variance $h^2_{G\times E_1}$; the remaining p_0 regression coefficients were generated independently and identically from a normal, NEG2, or Laplace distribution with mean zero and variance $h^2_{G\times E_2}$. Next, we randomly selected $3p_0$

causal variants, $G^*_1,\ldots,G^*_{3p_0}$, from among the p SNPs, $(G_1,\ldots,G_p)$. The first p_0 variants $(G^*_1,\ldots,G^*_{p_0})$ had a nonzero gene main effect, the second p_0 variants $(G^*_{1+p_0},\ldots,G^*_{2p_0})$ had a nonzero GxE interaction effect with E_1, and the remaining p_0 variants $(G^*_{1+2p_0},\ldots,G^*_{3p_0})$ had a nonzero GxE interaction effect with E_2.

Then, the conditional mean was set as

$$\mu_{\text{true}} = \frac{1}{\sqrt{p_0}}\sum_{j=1}^{p_0}\tilde{G}^*_j\beta^*_j + \frac{1}{\sqrt{p_0}}\sum_{j=1+p_0}^{2p_0}(G^*_j \tilde{\circ} E_1)\beta^*_j + \frac{1}{\sqrt{p_0}}\sum_{j=1+2p_0}^{3p_0}(G^*_j \tilde{\circ} E_2)\beta^*_j,$$

in which $\tilde{G}^*_j$, $(G^*_j \tilde{\circ} E_1)$, and $(G^*_j \tilde{\circ} E_2)$ denote the corresponding terms standardized to have mean zero and variance one. Finally, a quantitative trait was generated as $y = \mu_{\text{true}} + \epsilon$, where ϵ is an independently and identically distributed normal random variable with mean zero and variance $1 - h^2_G - h^2_{G\times E_1} - h^2_{G\times E_2}$. Note that $E(\frac{1}{\sqrt{p_0}}\sum_{j=1}^{p_0}\tilde{G}^*_j\beta^*_j) = \frac{1}{\sqrt{p_0}}\sum_{j=1}^{p_0}\tilde{G}^*_j E(\beta^*_j) = 0$ and $\text{Var}(\frac{1}{\sqrt{p_0}}\sum_{j=1}^{p_0}\tilde{G}^*_j\beta^*_j) = \frac{1}{p_0}\sum_{j=1}^{p_0}(\tilde{G}^*_j)^2\text{Var}(\beta^*_j) = h^2_G$, and, similarly, $\frac{1}{\sqrt{p_0}}\sum_{j=1+p_0}^{2p_0}(\tilde{G}^*_j \circ E_1)\beta^*_j$ and $\frac{1}{\sqrt{p_0}}\sum_{j=1+2p_0}^{3p_0}(\tilde{G}^*_j \circ E_2)\beta^*_j$ have mean zero and variance $h^2_{G\times E_1}$ and $h^2_{G\times E_2}$, respectively. Also note that the three terms in μ_{true} and ϵ are mutually independent. Thus, y has mean zero and variance 1, and the triplet $h^2 = (h^2_G, h^2_{G\times E_1}, h^2_{G\times E_2})$ is referred to as heritability throughout this paper. We considered a total of eight scenarios for h^2. First, we considered $(0.3,0,0)$, $(0.6,0,0)$, $(0,0.3,0)$, and $(0,0.6,0)$, where the first and second are scenarios with gene effect without GxE interactions, and the third and fourth are scenarios with GxE interactions only for E_1. Then we considered four additional scenarios: $(0,0.15,0.15)$, $(0,0.3,0.3)$, $(0,0,0.3)$, $(0,0,0.6)$, where the first and second are scenarios with GxE interactions both for E_1 and E_2, and the third and fourth are scenarios with GxE interactions only for E_2.

We used cross-validation to evaluate the prediction models. The data were randomly divided into two parts: 20% for training data and the remaining 80% for test data. The training dataset was used to build prediction models, and then the prediction accuracy of each model was evaluated on the basis of how well the simulated quantitative traits in the test dataset were predicted by the trained model. We used the prediction correlation coefficient (PCC) to measure the prediction accuracy. The above procedure was repeated 100 times. We note that the $3p_0$ causal SNPs and true regression coefficients differed for each replicate.

We also considered simulations for prediction of binary traits. A binary trait was generated by dichotomizing the quantitative trait on the basis of whether or not its value exceeded $\Phi^{-1}(0.7)$, in which Φ^{-1} is the standard normal quantile function. With a binary trait, the prediction accuracy of each model was evaluated by the area under the receiver operating characteristic curve (AUC).

Comparisons among prediction models

We compared the proposed extension of the STMGP method with other prediction models. We included the usual STMGP without GxE interaction as a competitor; specifically, the STMGP models compared were the STMGP without environmental variables, STMGP with environmental variable E_1, STMGP with environmental variable E_2, and STMGP with both environmental variables E_1 and E_2.

We also compared the proposed STMGP extension with other prediction models based on genomic BLUP. Specifically, we considered the following four genomic BLUP models,

$$\mu_b = 1_n\beta_0 + E_1\beta_{1,1} + E_2\beta_{2,1} + \sum_{j=1}^{p}\tilde{G}_j\beta_{j,2}, \quad (8)$$

$$\mu_{bge1} = 1_n\beta_0 + E_1\beta_{1,1} + E_2\beta_{2,1} + \sum_{j=1}^{p}\tilde{G}_j\beta_{j,2} + \sum_{j=1}^{p}(\tilde{G}_j \tilde{\circ} E_1)\beta_{j,3}, \quad (9)$$

$$\mu_{bge2} = 1_n\beta_0 + E_1\beta_{1,1} + E_2\beta_{2,1} + \sum_{j=1}^{p}\tilde{G}_j\beta_{j,2} + \sum_{j=1}^{p}(\tilde{G}_j \tilde{\circ} E_2)\beta_{j,3}, \quad (10)$$

$$\mu_{bge12} = 1_n\beta_0 + E_1\beta_{1,1} + E_2\beta_{2,1} + \sum_{j=1}^{p}\tilde{G}_j\beta_{j,2} + \sum_{j=1}^{p}(\tilde{G}_j \tilde{\circ} \bar{E}_{12})\beta_{j,3}, \quad (11)$$

where $\bar{E}_{12} = (E_1 + E_2)/2$, β_0 and β_1 are fixed effects, and $\beta_{j,2}$ and $\beta_{j,3}$ are random effects that are independently distributed as $N(0,\sigma^2_G)$ and $N(0,\sigma^2_{G\times E})$, respectively. Similar BLUP models have been considered in previous studies (e Sousa *et al.* 2017; Moore *et al.* 2018). We constructed the prediction model by BLUP implemented in the BGEE package for R (Granato *et al.* 2018) by using the BGEE function with options ite = 20000, burn = 1000, and thin = 3.

Application to prediction of real traits

We applied the proposed extension of the STMGP to the prediction of real traits. All variables were obtained from the ADNIMERGE package for R. We considered four cognitive scores as target traits for prediction: FAQ (Functional Assessment Questionnaire), CDRSB (Clinical Dementia Rating Sum of Boxes), MMSE (Mini-Mental State Examination), and ADAS11 [the 11-item ADAS-cog (Alzheimer's Disease Assessment Scale-Cognitive Subscale)]. We used SEX and EDU (years of education) as environmental variables. We also considered two additional covariates, AGE and *APOE4* genotype. The latter is a known risk allele for AD development. As with the above simulations, we evaluated prediction accuracy via cross-validation.

First, we randomly divided the 684 individuals into five groups of roughly equal size. Then, one of the five groups was selected as the test set and the remaining groups were used as the training set. Consequently, by repeating this with each group in turn acting as the test set, we had five different test/training sample combinations (*i.e.*, 5-fold cross-validation). For each of the five combinations, we built a prediction model based on the training set and predicted each trait value for the test set with the constructed prediction model.

For each training set, we used 528,984 SNPs as predictors as in the above simulation studies. The prediction models we compared were STMGP with SEX as the environmental variable, STMGP with EDU as the environmental variable, and STMGP with SEX and EDU both as environmental variables. BLUP-based prediction models are (8)–(11). Since the target traits are cognitive scores, we additionally studied regression models including *APOE4* genotype interaction without other variants; specifically, we considered the following models without GWAS data:

$$\mu_{l0} = 1_n\beta_0 + SEX\beta_{1,1} + EDU\beta_{2,1}, \quad (12)$$

$$\mu_l = 1_n\beta_0 + SEX\beta_{1,1} + EDU\beta_{2,1} + AGE\beta_{3,1} + APOE4\beta_{4,1}, \quad (13)$$

$$\mu_{lge1} = 1_n\beta_0 + SEX\beta_{1,1} + EDU\beta_{2,1} + AGE\beta_{3,1} + APOE4\beta_{4,1} + APOE4^{\circ}SEX\beta_{5,1}, \quad (14)$$

$$\mu_{lge2} = 1_n\beta_0 + SEX\beta_{1,1} + EDU\beta_{2,1} + AGE\beta_{3,1} + APOE4\beta_{4,1} + APOE4^{\circ}EDU\beta_{5,1}, \quad (15)$$

$$\mu_{lge12} = 1_n\beta_0 + SEX\beta_{1,1} + EDU\beta_{2,1} + AGE\beta_{3,1} + APOE4\beta_{4,1} + APOE4^{\circ}SEX\beta_{5,1} + APOE4^{\circ}EDU\beta_{6,1}. \quad (16)$$

Prediction accuracy was evaluated with PCC, which compares the predicted value with the actual trait in the test set.

Results

Simulation results

Results of the quantitative trait simulation are shown in Figures 1 and 2, and Supplementary Figure S1, where each cell exhibits mean PCC and the number of causal variants is $p_0 = 100$, 1000, and 500, respectively.

The first and second scenarios for h^2, $(0.3, 0, 0)$ and $(0.6, 0, 0)$, are those with gene effects but no GxE interactions. From Figures 1 and 2, and Supplementary Figure S1, all methods showed a higher predictive power in the latter scenario than in the former scenario due to the larger heritability. The four STMGP methods resulted in comparable predictive power, implying that the inclusion of GxE interactions had virtually no effect on predictive power, which is a reasonable result because no GxE interaction effects were assumed in the data-generating model. The BLUP models had lower predictive power than the STMGP methods, which is also reasonable because only a small proportion of variants was assumed to be causal and the BLUP models do not carry out variable selection. Indeed, by comparing Figures 1 and 2 and Supplementary Figure S1, it can be seen that an increase in the number of causal variants made the difference

Figure 1 Quantitative trait simulation with $p_0 = 100$. Average predictive correlation coefficient (PCC) for eight models. For each scenario (shown in rows), high values are highlighted in red and low values in white. s: STMGP with E_1 and E_2 as covariates; sge1: STMGP with E_1 and E_2 as covariates and E_1 as environmental variable for GxE interaction; sge2: STMGP with E_1 and E_2 as covariates and E_2 as environmental variable for GxE interaction; sge12: STMGP with E_1 and E_2 as covariates, and E_1 and E_2 as environmental variables for GxE interaction; bg: BLUP with E_1 and E_2 as covariates; bge1: BLUP with E_1 and E_2 as covariates and E_1 as environmental variable for GxE interaction; bge2: BLUP with E_1 and E_2 as covariates and E_2 as environmental variable for GxE interaction; bge12: BLUP with E_1 and E_2 as covariates, and average of E_1 and E_2 as environmental variable for GxE interaction. Scenarios are denoted as $(h^2_G, h^2_{G\times E_1}, h^2_{G\times E_2})$_ dist, where dist means effect size distribution: Normal, NEG2, or Laplace.

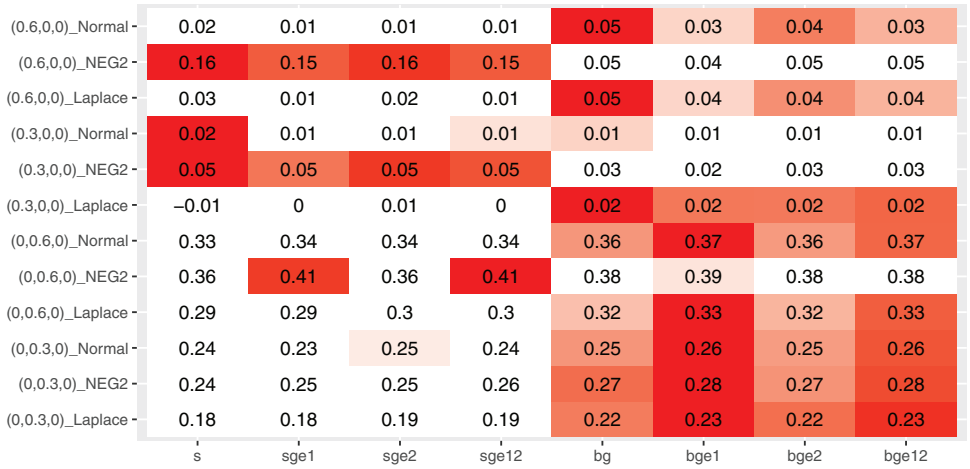

Figure 2 Quantitative trait simulation with $p_0 = 1000$. Average predictive correlation coefficient (PCC) for eight models. See Figure 1 for explanation of scenarios (shown in rows).

between the STMGP and BLUP methods smaller. The difference in effect size distribution had a non-negligible impact on predictive power. While the BLUP methods assume a normal distribution, the STMGP methods do not rely on the effect size distribution, and the STMGP methods had much higher predictive power than the BLUP methods, in particular, when the effect size distribution was non-normal. The difference between the STMGP and BLUP methods was pronounced under the NEG2 distribution, which has the heaviest tails among the three effect-size distributions compared. A similar result was observed in the simulation studies of Takahashi *et al.* (2020).

The third and fourth scenarios for h^2, $(0, 0.3, 0)$ and $(0, 0.6, 0)$, are those with GxE interactions only for E_1. As in the scenarios for $h^2 = (0.3, 0, 0)$ and $(0.6, 0, 0)$, all prediction models gave higher predictive power in the latter scenario than in the former scenario. Unlike the scenarios with no GxE interactions $h^2 = (0.3, 0, 0)$ and $(0.6, 0, 0)$, the STMGP methods incorporating GxE interaction effects had higher predictive power than the STMGP method without GxE interactions. For example, in scenario $h^2 = (0, 0.6, 0)$ under a normal effect-size distribution, the STMGP without GxE interaction produced mean PCC 0.36 (standard deviation 0.26), while the STMGP with GxE interaction on variable E_1 resulted in mean PCC 0.41 (standard deviation 0.22). On the other hand, the STMGP with GxE interaction on variable E_2 resulted in mean PCC 0.37 (standard deviation 0.26), which is comparable with STMGP without GxE interaction. This is reasonable since no GxE interaction effect on variable E_2 was assumed. The STMGP with GxE interaction on both E_1 and E_2 gave mean PCC 0.41 (standard deviation 0.23), a predictive power comparable to that of STMGP with GxE interaction on variable E_1. Total heritability and the difference in effect size distribution had a similar impact on predictive power in scenarios $(0.3, 0, 0)$ and $(0.6, 0, 0)$. For $p_0 = 100$ and the larger heritability scenario, $h^2 = (0, 0.6, 0)$, or under the NEG2 distribution, STMGP with GxE interaction on variable E_1 tended to produce higher predictive power than the BLUP methods, which is perhaps due to the fact that only a small proportion of variants was assumed to be causal. In the other cases among the third and fourth scenarios (any distribution with other than $(0, 0.6, 0)$ and $p_0 = 100$, or $p_0 = 100$ and NEG2 with any heritability [$(0, 0.3, 0)$ or $(0, 0.6, 0)$]), the STMGP methods did not always perform better than the BLUP methods.

Results of the additional four scenarios are shown in Supplementary Figures S3–S5. The first and second scenarios for h^2, $(0, 0.15, 0.15)$ and $(0, 0.3, 0.3)$, are the scenarios with GxE interactions both for E_1 and E_2. Unlike the scenarios $(0, 0.3, 0)$ and $(0, 0.6, 0)$, all three STMGP methods with GxE interaction had comparably higher predictive power than STGMP without GxE interaction. This is reasonable as GxE interaction was assumed for both variables, E_1 and E_2. The third and fourth scenarios for h^2, $(0, 0, 0.3)$ and $(0, 0, 0.6)$, are those with GxE interactions only for E_2. The results were similar to those for $(0, 0.3, 0)$ and $(0, 0.6, 0)$, in which the role of E_2 was replaced by E_1.

Results of the binary trait simulation are shown in Figures 3 and 4, and Supplementary Figure S2, in which each cell exhibits the mean AUC. The results were consistent overall with the results of the quantitative trait simulation, but differences in predictive power between methods were smaller than with the quantitative trait simulation.

Prediction of real quantitative trait

Results of predicting the four cognitive scores—FAQ, CDRSB, MMSE, and ADAS11—as target traits are shown in Table 1, which convey the five PCCs from 5-fold cross-validation. Generally, the prediction accuracy differed across the four traits. By comparing l0 with l, lge1, lge2, and lge12, which correspond to formulae (12)–(16), we see that inclusion of the *APOE4* genotype (without genome-wide variants) gave much higher predictive power. However, the observed comparable prediction ability among models l, lge1, lge2, and lge12 implies that the inclusion of an interaction between *APOE4* and either SEX or EDU did not impact predictive power. The BLUP methods, s, sge1, sge2, and sge12, resulted in performance that was comparable to those of l, lge1, lge2, and lge12, and did not show any extremely distinctive behavior. Similarly, the STMGP methods did not behave much differently from the other methods, but STMGP with a GxE interaction with EDU (sge2) tended to show slightly higher predictive power and improved upon the STMGP without GxE interaction. In particular, for prediction of FAQ, STMGP with a GxE

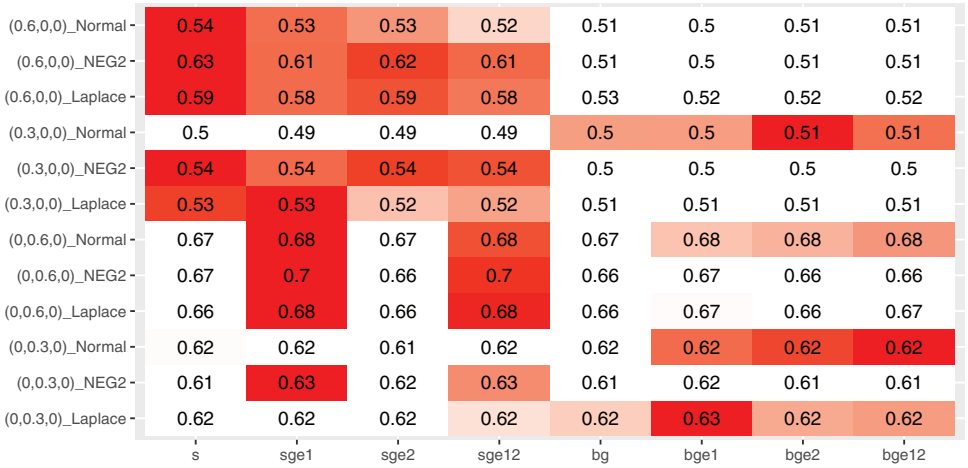

Figure 3 Binary trait simulation with $p_0 = 100$. Average area under the ROC curve (AUC) is shown for eight models. For each scenario (in rows), high values are highlighted in red and low values in white. s: STMGP with E_1 and E_2 as covariates; sge1: STMGP with E_1 and E_2 as covariates and E_1 as environmental variable for GxE interaction; sge2: STMGP with E_1 and E_2 as covariates and E_2 as environmental variable for GxE interaction; sge12: STMGP with E_1 and E_2 as covariates, and E_1 and E_2 as environmental variables for GxE interaction; bg: BLUP with E_1 and E_2 as covariates; bge1: BLUP with E_1 and E_2 as covariates and E_1 as environmental variable for GxE interaction; bge2: BLUP with E_1 and E_2 as covariates and E_2 as environmental variable for GxE interaction; bge12: BLUP with E_1 and E_2 as covariates, and average of E_1 and E_2 as environmental variable for GxE interaction. Scenarios are denoted as $(h^2_G, h^2_{G\times E_1}, h^2_{G\times E_2})$_ dist, where dist means effect size distribution: Normal, NEG2, or Laplace.

Figure 4 Binary trait simulation with $p_0 = 1000$. Average area under the ROC curve (AUC) for eight models. See Figure 3 for explanation of scenarios (shown in rows).

Table 1 Results of predicting four quantitative traits, FAQ, CDRSB, MMSE, and ADAS11

Trait[a]	Data[b]	l0[c]	l[d]	lge1[e]	lge2[f]	lge12[g]	s[h]	sge1[i]	sge2[j]	sge12[k]	bg[l]	bge1[m]	bge2[n]	bge12[o]
FAQ	CV 1	0.07	0.16	0.15	0.17	0.16	0.11	–0.01	0.15	0.05	0.14	0.13	0.13	0.12
	CV 2	0.17	0.35	0.33	0.36	0.34	0.26	0.24	0.32	0.31	0.32	0.35	0.33	0.33
	CV 3	0.19	0.15	0.15	0.16	0.16	0.19	0.13	0.21	0.15	0.17	0.15	0.18	0.17
	CV 4	0.01	0.26	0.26	0.27	0.27	0.31	0.18	0.24	0.19	0.23	0.28	0.25	0.23
	CV 5	0.08	0.16	0.16	0.10	0.09	0.15	0.14	0.17	0.15	0.17	0.14	0.17	0.15
	Mean	0.10	0.21	0.21	0.21	0.20	0.20	0.14	0.22	0.17	0.21	0.21	0.21	0.20
	SD	0.08	0.09	0.08	0.10	0.10	0.08	0.09	0.07	0.09	0.07	0.10	0.08	0.09
CDRSB	CV 1	0.07	0.13	0.13	0.12	0.12	0.21	0.18	0.22	0.17	0.12	0.13	0.10	0.11
	CV 2	0.16	0.38	0.37	0.36	0.35	0.33	0.28	0.33	0.30	0.34	0.36	0.34	0.33
	CV 3	0.22	0.26	0.26	0.26	0.26	0.28	0.26	0.26	0.25	0.25	0.25	0.26	0.27
	CV 4	0.10	0.37	0.37	0.37	0.37	0.44	0.36	0.41	0.31	0.36	0.39	0.37	0.36
	CV 5	0.19	0.27	0.26	0.25	0.22	0.27	0.25	0.28	0.27	0.27	0.25	0.27	0.27
	Mean	0.15	0.28	0.27	0.27	0.26	0.31	0.27	0.30	0.26	0.27	0.27	0.27	0.27
	SD	0.06	0.10	0.10	0.10	0.10	0.08	0.06	0.07	0.06	0.09	0.10	0.10	0.10
MMSE	CV 1	0.10	0.27	0.25	0.26	0.25	0.13	0.21	0.18	0.16	0.22	0.23	0.23	0.22
	CV 2	0.19	0.34	0.33	0.33	0.32	0.30	0.33	0.33	0.33	0.29	0.30	0.31	0.30
	CV 3	0.30	0.35	0.35	0.35	0.35	0.28	0.26	0.34	0.35	0.37	0.38	0.36	0.36
	CV 4	0.27	0.35	0.35	0.35	0.36	0.35	0.34	0.39	0.37	0.36	0.37	0.36	0.37
	CV 5	0.17	0.28	0.26	0.28	0.25	0.25	0.23	0.26	0.22	0.29	0.28	0.29	0.27
	Mean	0.21	0.32	0.31	0.31	0.31	0.26	0.27	0.30	0.29	0.31	0.31	0.31	0.30
	SD	0.08	0.04	0.05	0.04	0.05	0.08	0.06	0.08	0.09	0.06	0.06	0.05	0.06
ADAS11	CV 1	0.12	0.31	0.32	0.30	0.31	0.30	0.28	0.29	0.26	0.29	0.29	0.28	0.27
	CV 2	0.17	0.30	0.30	0.30	0.30	0.22	0.23	0.24	0.22	0.28	0.27	0.28	0.29
	CV 3	0.15	0.29	0.30	0.29	0.30	0.22	0.26	0.24	0.26	0.29	0.29	0.29	0.29
	CV 4	0.11	0.36	0.36	0.35	0.35	0.29	0.29	0.37	0.29	0.37	0.38	0.35	0.36
	CV 5	0.22	0.34	0.32	0.33	0.32	0.30	0.28	0.34	0.24	0.33	0.31	0.32	0.31
	Mean	0.15	0.32	0.32	0.31	0.32	0.27	0.27	0.30	0.25	0.31	0.31	0.30	0.30
	SD	0.04	0.03	0.03	0.03	0.02	0.04	0.02	0.06	0.03	0.04	0.04	0.03	0.03

[a] Prediction of each target trait is evaluated by the prediction correlation coefficient (PCC) from 5-fold cross-validation.
[b] Data used to calculate PCC (CV 1–CV 5 denote each cross-validated dataset from 5-fold cross-validation) for each model are shown in row together with mean and standard deviation (SD).
[c] Linear regression with SEX and EDU as predictors.
[d] Linear regression with SEX, EDU, AGE, and *APOE4* as predictors.
[e] Linear regression with SEX, EDU, AGE, *APOE4*, and *APOE4*xSEX as predictors.
[f] Linear regression with SEX, EDU, AGE, *APOE4*, and *APOE4*xEDU as predictors.
[g] Linear regression with SEX, EDU, AGE, *APOE4*, *APOE4*xSEX, and *APOE4*xEDU as predictors.
[h] STMGP with SEX, EDU, AGE, and *APOE4* as covariates.
[i] STMGP with SEX, EDU, AGE, and *APOE4* as covariates, and SEX as environmental variable for GxE interaction.
[j] STMGP with SEX, EDU, AGE, and *APOE4* as covariates, and EDU as environmental variable for GxE interaction.
[k] STMGP with SEX, EDU, AGE, and *APOE4* as covariates, and AGE and EDU as environmental variables for GxE interaction.
[l] BLUP with SEX, EDU, AGE, and *APOE4* as covariates.
[m] BLUP with SEX, EDU, AGE, and *APOE4* as covariates, and SEX as environmental variable for GxE interaction.
[n] BLUP with SEX, EDU, AGE, and *APOE4* as covariates, and EDU as environmental variable for GxE interaction.
[o] BLUP with SEX, EDU, AGE, and *APOE4* as covariates, and average of AGE and EDU as environmental variable for GxE interaction.

interaction with EDU (sge2) gave the highest mean PCC (0.22; standard deviation 0.07) among the methods. However, the differences among models were small: for example, the second best mean PCC was 0.21 for l, lge1, lge2, bg, bge1, bge2, and the mean PCC for the STMGP without GxE interaction was is 0.20 with standard deviation 0.08. On the other hand, the STMGPs with GxE interaction with SEX (sge1) or with both SEX and EDU (sge12) produced lower or more variable prediction results.

The above results indicate the possibility that incorporating GxE interactions leads to improved predictive performance. Of course, whether the predictive performance is improved or not depends on the choice of environmental variable, which was also observed in the simulation studies.

Finally, we checked the validity of the proposed univariate regression approximation in the real data application. Supplementary Figures S9–S16 show the accuracy of the proposed approximation, where each figure gives a scatter plot matrix of *P*-values associated with the GxE interaction term $G_j{}^{\circ}E$ from models (6) and (7) with environmental variables either centered or not. Since centering of environmental variable *E* does not change the model (6), we only compared three *P*-values: model (6), model (7) with centered *E*, and model (7) with non-centered *E*. Among the figures, Supplementary Figures S9, S11, S13, and S15 show the *P*-values associated with GxE interaction for SEX as the environmental variable, and Supplementary Figures S10, S12, S14, and S16 show the *P*-values associated with GxE interaction for EDU as the environmental variable. In all figures, the $-\log_{10}$ *P*-values for the GxE interaction term in the approximate univariate regression (*i.e.*, with no gene main effect) using a centered environmental variable were highly correlated (>0.99) with the $-\log_{10}$ *P*-values for the GxE interaction term in the interaction model having a gene main effect. On the other hand, with a non-centered environmental variable the same sets of $-\log_{10}$ *P*-values for the GxE interaction terms were either less correlated (correlation around 0.65 for SEX as *E*) or uncorrelated (< 0.02 for EDU). These results confirm the validity of the proposed univariate regression approximation.

Discussion

In this article, we presented a procedure to incorporate GxE interaction effects into our previously developed genetic modeling approach, the STMGP method. Since the STMGP method relies on univariate regression to screen for high-dimensional predictors, we developed a univariate regression approximation to the GxE interaction model so that the STMGP framework can be directly applied without modification. The approximation is simply to use "centered" environmental variables and remove gene main effect terms from the standard GxE interaction regression model. Simulation studies and real data analysis showed that incorporating GxE interactions may improve the performance of the STMGP, but, as expected, its effectiveness depends to a great extent on the underlying genetic structure.

An important point to note is that genome-wide GxE interaction analysis is more sensitive to model misspecification than marginal association analysis, as pointed out by Voorman *et al.* (2011), Almli *et al.* (2014), and Ueki *et al.* (2019). Since the model misspecification issue applies to all GxE interaction analyses, special care should be taken in modeling GxE interaction, such as selection of the environmental variable. We recommend using the check statistic proposed by Ueki *et al.* (2019) before performing a GxE interaction analysis; this enables prediction of problematic behavior in the GxE interaction analysis without having to perform the actual genome-wide scan.

Most of the existing genetic prediction models treat genetic data separately from non-genetic data. While the widely used additive models to combine genetic and non-genetic data are simple and easy to handle, there must be situations where non-additive models, such as models with GxE interactions, improve upon the additive models. However, studies have reported low power of GxE interaction analysis (Kraft *et al.* 2007). Nevertheless, analogous to the relationship between an association study and prediction modeling, the goal is not to discover GxE interactions but to have a better prediction model. Low statistical power is not necessarily a severe issue in this context: GxE interactions, even if not genome-wide significant, may be useful in helping to improve predictive power.

Data availability

All data necessary to reproduce the conclusions are fully presented in the paper. The authors do not have ownership of the data used; the data obtained were collected and are owned by the Alzheimer's Disease Neuroimaging Initiative (ADNI). Researchers may request and access the data through the ADNI website (http://adni.loni.usc.edu/ (Accessed: 2021 August 31)). The authors had no special access privileges to use these data. A computer program for the method proposed in this paper is available from the R package stmgp (version 1.0.4).

Supplementary material is available at *G3* online.

Acknowledgements

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and the DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, by the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; EliLilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org (Accessed: 2021 August 31)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Computations for the present work were performed by using the facilities of the Institute of Statistical

Mathematics. The authors are grateful for the helpful comments by two referees, associate editor, and Dr. John Cologne.

Funding

This work was supported by the JSPS KAKENHI Grant Numbers 20K11723 and 20H00576.

Conflicts of interest

None declared.

Literature cited

Abraham G, Kowalczyk A, Zobel J, Inouye M. 2013. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet Epidemiol. 37: 184–195.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: BN Petrov, F Caski, editors. Proceedings of the 2nd International Symposium on Information Theory. Budapest: Akadimiai Kiado. p. 267–281.

Almli LM, Duncan R, Feng H, Ghosh D, Binder EB, *et al.* 2014. Correcting systematic inflation in genetic association tests that consider interaction effects. JAMA Psychiatry. 71:1392–1399.

Arnau-Soler A, Macdonald-Dunlop E, Adams MJ, Clarke T-K, MacIntyre DJ, *et al.*; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. 2019. Genome-wide by environment interaction studies of depressive symptoms and psychosocial stress in UK biobank and generation scotland. Transl Psychiatry. 9:14.

Aschard H. 2016. A perspective on interaction effects in genetic association studies. Genet Epidemiol. 40:678–688.

Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, *et al.* 2012. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet. 131:1591–1613.

Ayers K, Cordell H. 2010. Snp selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol. 34:879–891.

Breiman L. 1996. Heuristics of instability and stabilization in model selection. Ann Stat. 24:2350–2383.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, *et al.* 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 7.4:

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 92:399–418.

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock S, *et al.* 2013. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 45: 400–405.

Cherlin S, Plant D, Taylor JC, Colombo M, Spiliopoulou A, *et al.* 2018. Prediction of treatment response in rheumatoid arthritis patients using genome-wide SNP data. Genet Epidemiol. 42: 754–771.

de Los Campos G, Vazquez A, Fernando R, Klimentidis Y, Sorensen D. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLOS Genet. 9:e1003608.

Dudbridge F. 2013. Power and predictive accuracy of polygenic risk scores. PLOS Genet. 9:e1003348.

e Sousa MB, Cuevas J, de Oliveira Couto EG, Pérez-Rodríguez P, Jarquín D, *et al.* 2017. Genomic-enabled prediction in maize using kernel models with genotype x environment interaction. G3 (Bethesda)). 7:1995–2014.

Efron B. 2004. The estimation of prediction error: covariance penalties and cross-validation. J Am Stat Assoc. 99:619–632.

Evans D, Visscher P, Wray N. 2009. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 18: 3525–3531.

Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion). J R Stat Soc Series B Stat Methodol. 70:903–911.

Gauderman WJ, Mukherjee B, Aschard H, Hsu L, Lewinger JP, *et al.* 2017. Update on the state of the science for analytical methods for gene-environment interactions. Am J Epidemiol. 186: 762–770.

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-López O, *et al.* 2018. BGGE: a new package for genomic-enabled prediction incorporating genotype x environment interaction models. G3 (Bethesda)). 8:3039–3047.

Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, *et al.* 2011. Genome-wide gene-environment study identifies glutamate receptor gene *GRIN2A* as a Parkinson's disease modifier gene via interaction with coffee. PLOS Genet. 7:e1002237.

Hoggart C, Whittaker J, Iorio M, Balding D. 2008. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. PLOS Genet. 4:e1000130.

Khoury MJ. 2017. Editorial: emergence of gene-environment interaction analysis in epidemiologic research. Am J Epidemiol. 186: 751–752.

Kooperberg C, LeBlanc M. 2008. Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol. 32:255–263.

Kraft P, Aschard H. 2015. Finding the missing gene–environment interactions. Eur J Epidemiol. 30:353–355.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 63:111–119.

Lello L, Avery SG, Tellier L, Vazquez AI, de los Campos G, *et al.* 2018. Accurate genomic prediction of human height. Genetics. 210: 477–497.

Maher B. 2008. Personal genomes: the case of the missing heritability. Nature. 456:18–21.

Maier R, Moser G, Chen G-B, Ripke S, Coryell W, *et al.*; Cross-Disorder Working Group of the Psychiatric Genomics Consortium. 2015. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 96:283–294.

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, *et al.* 2011. Beyond missing heritability: prediction of complex traits. PLOS Genet. 7:e1002051.

Manolio T. 2013. Bringing genome-wide association findings into clinical use. Nat Rev Genet. 14:549–558.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, *et al.* 2009. Finding the missing heritability of complex diseases. Nature. 461: 747–753.

McAllister K, Mechanic LE, Amos C, Aschard H, Blair IA, *et al.* 2017. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. Am J Epidemiol. 186: 753–761.

Meijsen JJ, Campbell A, Hayward C, Porteous DJ, Deary IJ, *et al.* 2018. Phenotypic and genetic analysis of cognitive performance in major depressive disorder in the generation Scotland: Scottish family health study. Transl Psychiatry. 8:63.

Moore R, Casale FP, Jan Bonder M, Horta D, Franke L, *et al.*; BIOS Consortium. 2018. A linear mixed-model approach to study multivariate gene–environment interactions. Nat Genet. 51:180–186.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, *et al.* 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLOS Genet. 11: e1004969.

Mukherjee B, Chatterjee N. 2007. Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics. 64:685–694.

Ober C, Vercelli D. 2011. Gene–environment interactions in human disease: nuisance or opportunity? Trends Genet. 27:107–115.

Osazuwa-Peters OL, Waken RJ, Schwander KL, Sung YJ, de Vries PS, *et al.* 2020. Identifying blood pressure loci whose effects are modulated by multiple lifestyle exposures. Genet Epidemiol. 44: 629–641.

Privé F, Aschard H, Blum MGB. 2019. Efficient implementation of penalized regression for genetic risk prediction. Genetics. 212: 65–74.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, *et al.* 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559–575.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, *et al.*; International Schizophrenia Consortium. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 460:748–752.

Ritchie MD, Davis JR, Aschard H, Battle A, Conti D, *et al.* 2017. Incorporation of biological knowledge into the study of gene-environment interactions. Am J Epidemiol. 186:771–777.

Shen L, Thompson P, Potkin S, Bertram L, Farrer L, *et al.*; Alzheimer's Disease Neuroimaging Initiative. 2014. Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers. Brain Imaging Behav. 8:183–207.

Stein C. 1981. Estimation of the mean of a multivariate normal distribution. Ann Stat. 9:1135–1151.

Sung YJ, de las Fuentes L, Schwander KL, Simino J, Rao DC. 2014. Gene–smoking interactions identify several novel blood pressure loci in the framingham heart study. Am J Hypertens. 28: 343–354.

Sung YJ, Winkler TW, Manning AK, Aschard H, Gudnason V, *et al.* 2016. An empirical comparison of joint and stratified frameworks for studying g x e interactions: Systolic blood pressure and smoking in the CHARGE gene-lifestyle interactions working group. Genet Epidemiol. 40:404–415.

Takahashi Y, Ueki M, Tamiya G, Ogishima S, Kinoshita K, *et al.* 2020. Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes. Transl Psychiatry. 10:294.

Takane Y, Yanai H. 1999. On oblique projectors. Linear Algebra Appl. 289:297–310.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Methodol. 58:267–288.

Ueki M. 2009. A note on automatic variable selection using smooth-threshold estimating equation. Biometrika. 96: 1005–1011.

Ueki M, Fujii M, Tamiya G; for Alzheimer's Disease Neuroimaging Initiative and the Alzheimer's Disease Metabolomics Consortium. 2019. Quick assessment for systematic test statistic inflation/deflation due to null model misspecifications in genome-wide environment interaction studies. PLOS One. 14: e0219825.

Ueki M, Kawasaki Y. 2013. Multiple choice from competing regression models under multicollinearity based on standardized update. Comput Stat Data Anal. 63:31–41.

Ueki M, Tamiya G; Alzheimer's Disease Neuroimaging Initiative 2016. Smooth-threshold multivariate genetic prediction with unbiased model selection. Genet Epidemiol. 40:233–243.

Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet. 97:576–592.

Voorman A, Lumley T, McKnight B, Rice K. 2011. Behavior of qq-plots and genomic control in studies of gene-environment interaction. PLOS One. 6:e19416.

Warren H, Casas J, Hingorani A, Dudbridge F, Whittaker J. 2013. Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores. Genet Epidemiol. 38: 72–83.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 88:76–82.

Ye J. 1998. On measuring and correcting the effects of data mining and model selection. J Am Stat Assoc. 93:120–131.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J R Stat Soc B. 67:301–320.

Communicating editor: J. G. D. Prendergast

Appendix

Derivation of the univariate regression approximation

We consider the GxE interaction model $y_i = \beta_{0j} + E_i\beta_{1j} + G_{ij}\beta_{2j} + E_iG_{ij}\beta_{3j} + \epsilon_i$, where the ϵ_i are independently and identically distributed with mean zero and variance σ_0^2. Here, we assume that $G_{ij}$and E_i are independent, and that each is independently and identically distributed for $i = 1, \ldots, n$. We also assume that σ_0^2, $v_{G_j} = \mathrm{Var}(G_{ij})$, and $v_E = \mathrm{Var}(E_i)$ are finite. Let $P_X = X(X^TX)^{-1}X^T$ be the projection matrix onto the column space of X, and let $Q_X = I_n - P_X$. Then, for n-dimensional one-vector 1_n, the operator $Q_{1_n} = I_n - P_{1_n}$ gives centering to have mean zero. Let $\tilde{E} = Q_{1_n}E$ and $\tilde{G}_j = Q_{1_n}G_j$. Then, for large n,

$$n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{E} = \mathrm{E}(n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{E}) + O_p\{\mathrm{Var}(n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{E})^{1/2}\} = O_p(n^{-1/2}). \tag{A1}$$

Finally, let $W_j = E^{\circ}G_j$ and $\tilde{W}_j = \tilde{E}^{\circ}G_j$.

Note that $P_X = P_{n^{-1/2}X}$ for any given matrix X. Thus, $Q_X = Q_{n^{-1/2}X}$ also holds. By Ueki and Kawasaki (2013) and Ueki *et al.* (2019), the least squares estimate of regression coefficient β_{3j} in the model $\mu = \mu(E, G_j) = 1_n\beta_{0j} + E\beta_{1j} + G_j\beta_{2j} + (G_j{}^{\circ}E)\beta_{3j}$, model (6), is

$$\overset{0123}{\underset{3j}{\hat{\beta}}}(E, G_j) = \frac{y^TQ_{(1_n,E,G_j)}W_j}{||Q_{(1_n,E,G_j)}W_j||^2} = \frac{n^{-1}y^TQ_{(1_n,E,G_j)}W_j}{n^{-1}||Q_{(1_n,E,G_j)}W_j||^2}. \tag{A2}$$

Similarly, the least squares estimate of regression coefficient β_{3j} in the model $\mu = \mu(E, G_j) = 1_n\beta_{0j} + E\beta_{1j} + G_j\beta_{2j} + (G_j{}^{\circ}E)\beta_{3j}$ is

$$\overset{013}{\underset{3j}{\hat{\beta}}}(E, G_j) = \frac{y^TQ_{(1_n,E)}W_j}{||Q_{(1_n,E)}W_j||^2} = \frac{n^{-1}y^TQ_{(1_n,E)}W_j}{n^{-1}||Q_{(1_n,E)}W_j||^2}. \tag{A3}$$

We utilize the decomposition of a projection matrix or blockwise formula (Takane and Yanai 1999, Lemma 3 (iii)), $P_{(A,B)} = P_A + P_{Q_AB}$ for two matrixes A and B. Note that $P_AP_{Q_AB} = P_{Q_AB}P_A = O$ since $Q_AA = O$. Then, $P_{(1_n,E)} = P_{1_n} + P_{\tilde{E}}$. Using this, and by the blockwise formula again, we have $P_{(1_n,E,G_j)} = P_{1_n} + P_{(\tilde{E},\tilde{G}_j)} = P_{1_n} + P_{\tilde{E}} + P_{Q_{\tilde{E}}\tilde{G}_j} = P_{(1_n,E)} + P_{Q_{\tilde{E}}\tilde{G}_j}$. Thus,

$$Q_{(1_n,E,G_j)} = Q_{(1_n,E)} - P_{Q_{\tilde{E}}\tilde{G}_j},$$

and applying this identity to (A2),

$$\begin{aligned}\overset{0123}{\underset{3j}{\hat{\beta}}}(E, G_j) &= \frac{n^{-1}y^TQ_{(1_n,E,G_j)}W_j}{n^{-1}\overset{T}{\underset{j}{W}}Q_{(1_n,E,G_j)}W_j} \\ &= \frac{n^{-1}y^T\{Q_{(1_n,E)} - P_{Q_{\tilde{E}}\tilde{G}_j}\}W_j}{n^{-1}\overset{T}{\underset{j}{W}}\{Q_{(1_n,E)} - P_{Q_{\tilde{E}}\tilde{G}_j}\}W_j} \\ &= \frac{n^{-1}y^TQ_{(1_n,E)}W_j - n^{-1}y^TP_{Q_{\tilde{E}}\tilde{G}_j}W_j}{n^{-1}||Q_{(1_n,E)}W_j||^2 - n^{-1}\overset{T}{\underset{j}{W}}P_{Q_{\tilde{E}}\tilde{G}_j}W_j},\end{aligned} \tag{A4}$$

which differs from (A3) unless $n^{-1}W_j^TP_{Q_{\tilde{E}}\tilde{G}_j}W_j$ and $n^{-1}W_j^TP_{Q_{\tilde{E}}\tilde{G}_j}W_j$ are both negligible. Let $\overline{G}_j = Q_{\tilde{E}}\tilde{G}_j$. The second term of the numerator of (A4) can be written as

$$\begin{aligned}n^{-1}y^TP_{\overline{G}_j}W_j &= n^{-1}y^T\overline{G}_j(\overset{T}{\underset{j}{\overline{G}}}\overline{G}_j)^{-1}\overset{T}{\underset{j}{\overline{G}}}W_j \\ &= (n^{-1}y^T\overline{G}_j)(n^{-1}\overset{T}{\underset{j}{\overline{G}}}\overline{G}_j)^{-1}(n^{-1}\overset{T}{\underset{j}{W}}\overline{G}_j).\end{aligned}$$

To begin with, by (A1) the left, middle, and right terms reduce to

$$\begin{aligned}n^{-1}y^T\overline{G}_j = n^{-1}y^TQ_{n^{-1/2}\tilde{E}}\tilde{G}_j &= n^{-1}y^T\tilde{G}_j - \frac{(n^{-1}y^T\tilde{E})(n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{E})}{||n^{-1/2}\tilde{E}||^2} \\ &= n^{-1}y^T\tilde{G}_j + o_p(1),\end{aligned} \tag{A5}$$

$$\begin{aligned}n^{-1}\overset{T}{\underset{j}{\overline{G}}}\overline{G}_j = n^{-1}\overset{T}{\underset{j}{\tilde{G}}}Q_{n^{-1/2}\tilde{E}}\tilde{G}_j &= n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{G}_j - \frac{(n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{E})^2}{||n^{-1/2}\tilde{E}||^2} \\ &= n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{G}_j + o_p(1),\end{aligned} \tag{A6}$$

$$\begin{aligned}n^{-1}\overset{T}{\underset{j}{W}}\overline{G}_j = n^{-1}\overset{T}{\underset{j}{W}}Q_{n^{-1/2}\tilde{E}}\tilde{G}_j &= n^{-1}\overset{T}{\underset{j}{W}}\tilde{G}_j - \frac{(n^{-1}\overset{T}{\underset{j}{W}}\tilde{E})(n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{E})}{||n^{-1/2}\tilde{E}||^2} \\ &= n^{-1}\overset{T}{\underset{j}{W}}\tilde{G}_j + o_p(1),\end{aligned} \tag{A7}$$

respectively. Combining (A5)–(A7), the numerator of (A4) reduces to

$$\begin{aligned}&n^{-1}y^TQ_{(1_n,E)}W_j - n^{-1}y^TP_{Q_{\tilde{E}}\tilde{G}_j}W_j \\ &= n^{-1}y^TQ_{(1_n,E)}W_j - \frac{(n^{-1}y^T\tilde{G}_j)(n^{-1}\overset{T}{\underset{j}{W}}\tilde{G}_j)}{n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{G}_j} + o_p(1).\end{aligned} \tag{A8}$$

By analogous calculations, the denominator of (A4) reduces to

$$n^{-1}||Q_{(1_n,E)}W_j||^2 - n^{-1}\overset{T}{\underset{j}{W}}P_{Q_{\tilde{E}}\tilde{G}_j}W_j \tag{A9}$$

$$= n^{-1}||Q_{(1_n,E)}W_j||^2 - \frac{(n^{-1}\overset{T}{\underset{j}{W}}\tilde{G}_j)^2}{n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{G}_j} + o_p(1). \tag{A10}$$

Substituting (A8) and (A10) into (A4),

$$\overset{0123}{\underset{3j}{\hat{\beta}}}(E, G_j) = \frac{n^{-1}y^TQ_{(1_n,E)}W_j - \frac{(n^{-1}y^T\tilde{G}_j)(n^{-1}\overset{T}{\underset{j}{W}}\tilde{G}_j)}{n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{G}_j}}{n^{-1}||Q_{(1_n,E)}W_j||^2 - \frac{(n^{-1}\overset{T}{\underset{j}{W}}\tilde{G}_j)^2}{n^{-1}\overset{T}{\underset{j}{\tilde{G}}}\tilde{G}_j}} + o_p(1). \tag{A11}$$

This approximates (A3) if $(n^{-1}y^T\tilde{G}_j)(n^{-1}W_j^T\tilde{G}_j)$ and $(n^{-1}W_j^T\tilde{G}_j)^2$ are both negligible, which, however, might not be true in general.

Instead, we consider the case where E is replaced by $\tilde{E} = Q_{1_n}E = E - \overline{E}1_n$ in (A2). In this case, the estimate of regression coefficient (A3) is

$$\overset{013}{\hat{\beta}}_{3j}(\tilde{E}, G_j) = \frac{y^T Q_{(1_n,\tilde{E})}\tilde{W}_j}{||Q_{(1_n,\tilde{E})}\tilde{W}_j||^2} = \frac{n^{-1}y^T Q_{(1_n,\tilde{E})}\tilde{W}_j}{n^{-1}||Q_{(1_n,\tilde{E})}\tilde{W}_j||^2}, \tag{A12}$$

and the corresponding model is $\mu = \mu(\tilde{E}, G_j) = 1_n\beta_{0j} + \tilde{E}\beta_{1j} + G_j\beta_{2j} + (G_j{}^{\circ}\tilde{E})\beta_{3j}$ (*i.e.*, model (7)). By an argument analogous to that which leads to (A11),

$$\begin{aligned}\overset{0123}{\hat{\beta}}_{3j}(\tilde{E}, G_j) &= \frac{y^T Q_{(1_n,\tilde{E},G_j)}\tilde{W}_j}{||Q_{(1_n,\tilde{E},G_j)}\tilde{W}_j||^2} = \frac{n^{-1}y^T Q_{(1_n,\tilde{E},G_j)}\tilde{W}_j}{n^{-1}||Q_{(1_n,\tilde{E},G_j)}\tilde{W}_j||^2} \\ &= \frac{n^{-1}y^T Q_{(1_n,\tilde{E})}\tilde{W}_j - \dfrac{(n^{-1}y^T\tilde{G}_j)(n^{-1}\overset{T}{\tilde{W}}_j\tilde{G}_j)}{n^{-1}\overset{T}{\tilde{G}}_j\tilde{G}_j}}{n^{-1}||Q_{(1_n,\tilde{E})}\tilde{W}_j||^2 - \dfrac{(n^{-1}\overset{T}{\tilde{W}}_j\tilde{G}_j)^2}{n^{-1}\overset{T}{\tilde{G}}_j\tilde{G}_j}} + o_p(1).\end{aligned} \tag{A13}$$

Here we focus on the quantity

$$\begin{aligned}n^{-1}\overset{T}{\tilde{W}}_j\tilde{G}_j &= n^{-1}\sum_{i=1}^{n}\tilde{E}_i G_{ij}\tilde{G}_{ij} \\ &= \mathrm{E}(n^{-1}\sum_{i=1}^{n}\tilde{E}_i G_{ij}\tilde{G}_{ij}) + O_p\{\mathrm{Var}(n^{-1}\sum_{i=1}^{n}\tilde{E}_i G_{ij}\tilde{G}_{ij})^{1/2}\}.\end{aligned}$$

By the independence between E and G_j,

$$\mathrm{E}(n^{-1}\sum_{i=1}^{n}\tilde{E}_i G_{ij}\tilde{G}_{ij}) = n^{-1}\sum_{i=1}^{n}\mathrm{E}(\tilde{E}_i)\mathrm{E}(G_{ij}\tilde{G}_{ij}) = 0,$$

where the last identity is due to the fact that $\mathrm{E}(\tilde{E}_i) = \mathrm{E}(E_i - \bar{E}) = 0$ for any i. As a consequence,

$$n^{-1}\overset{T}{\tilde{W}}_j\tilde{G}_j = O_p(n^{-1/2}),$$

and by substituting the above into (A13),

$$\overset{0123}{\hat{\beta}}_{3j}(\tilde{E}, G_j) = \frac{n^{-1}y^T Q_{(1_n,\tilde{E})}\tilde{W}_j + (n^{-1}y^T\tilde{G}_j)O_p(n^{-1/2})}{n^{-1}||Q_{(1_n,\tilde{E})}\tilde{W}_j||^2} + o_p(1). \tag{A14}$$

This representation reveals that, if $n^{-1}y^T Q_{(1_n,\tilde{E})}\tilde{W}_j$ dominates $(n^{-1}y^T\tilde{G}_j)n^{-1/2}$, $\hat{\beta}_{3j}^{0123}(\tilde{E}, G_j)$ (equation (A14)) is approximated by $\hat{\beta}_{3j}^{013}(\tilde{E}, G_j)$ (equation (A12)). In other words, the approximation breaks down only if $n^{-1/2}(n^{-1}y^T\tilde{G}_j)$ cannot be ignored in comparison to $n^{-1}y^T Q_{(1_n,\tilde{E})}\tilde{W}_j$ for large n, which is the case when the jth variant has a large marginal effect on y while the GxE interaction effect is weak or absent. Such variants should in principle be captured by the marginal association scan. The proposed algorithm thus implements the marginal association scan in addition to the GxE interaction scan, which avoids missing variants that have strong marginal effects. Supplementary Figures S9–S16 confirm that the approximation works well in practice with real data, in which we can see the importance of centering E (see "Prediction of real quantitative trait" section).

Invariance of regression coefficient estimate for GxE interaction

Here we show that the least squares estimate of regression coefficient β_{3j} in the model $\mu = \mu(E, G_j) = 1_n\beta_{0j} + E\beta_{1j} + G_j\beta_{2j} + (G_j{}^{\circ}E)\beta_{3j}$, model (6), is invariant if E is replaced by $E^a = E - a1_n$ and/or G_j is replaced by $G_j^b = G_j - b1_n$ for any scalar values a and b. Recall (A2),

$$\overset{0123}{\hat{\beta}}_{3j}(E, G_j) = \frac{y^T Q_{(1_n,E,G_j)}W_j}{||Q_{(1_n,E,G_j)}W_j||^2},$$

where $W_j = (E{}^{\circ}G_j)$. Therefore,

$$\overset{0123}{\hat{\beta}}_{3j}(E^a, \overset{b}{G}_j) = \frac{y^T Q_{(1_n,E^a,\overset{b}{G}_j)}\overset{a,b}{W}_j}{||Q_{(1_n,E^a,\overset{b}{G}_j)}\overset{a,b}{W}_j||^2}, \tag{B1}$$

where $W_j^{a,b} = (E^a{}^{\circ}G_j^b) = (E - a1_n)^{\circ}(G_j - b1_n) = W_j - bE - aG_j + ab1_n$. Note that $Q_{(1_n,E^a,G_j^b)} = I_n - P_{(1_n,E^a,G_j^b)} = I_n - P_{(1_n,E-a1_n,G_j-b1_n)} = I_n - P_{(1_n,E,G_j)} = Q_{(1_n,E,G_j)}$. Hence,

$$Q_{(1_n,E^a,\overset{b}{G}_j)}\overset{a,b}{W}_j = Q_{(1_n,E,G_j)}(W_j - bE - aG_j + ab1_n) = Q_{(1_n,E,G_j)}W_j,$$

in which the second identity is due to the fact that bE, aG_j, and $ab1_n$ are included in the linear span by $(1_n, E, G_j)$. Therefore, by (B1), for any scalar values a and b, the following identity holds:

$$\overset{0123}{\hat{\beta}}_{3j}(E^a, \overset{b}{G}_j) = \frac{y^T Q_{(1_n,E,G_j)}W_j}{||Q_{(1_n,E,G_j)}W_j||^2} = \overset{0123}{\hat{\beta}}_{3j}(E, G_j). \tag{B2}$$

It is noteworthy that the invariance is essentially due to the involvement of both E and G_j, so it is not guaranteed to hold in the absence of either of the two terms.