Bioinformatics, 36(24), 2020, 5640–5648
doi: 10.1093/bioinformatics/btaa1083
Advance Access Publication Date: 16 January 2021
Original Paper

Genetics and population analysis

A two-step approach to testing overall effect of gene–environment interaction for multiple phenotypes

Arunabha Majumdar [1,2], Kathryn S. Burch[3], Tanushree Haldar[4], Sriram Sankararaman[5], Bogdan Pasaniuc[1,3], W. James Gauderman[6] and John S. Witte[2,*]

[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [2]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, USA, [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA, [4]Institute for Human Genetics, University of California, San Francisco, CA 94158, USA, [5]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095, USA and [6]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90007, USA

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz

Received on July 6, 2020; revised on December 9, 2020; editorial decision on December 11, 2020; accepted on December 17, 2020

Abstract

Motivation: While gene–environment (GxE) interactions contribute importantly to many different phenotypes, detecting such interactions requires well-powered studies and has proven difficult. To address this, we combine two approaches to improve GxE power: simultaneously evaluating multiple phenotypes and using a two-step analysis approach. Previous work shows that the power to identify a main genetic effect can be improved by simultaneously analyzing multiple related phenotypes. For a univariate phenotype, two-step methods produce higher power for detecting a GxE interaction compared to single step analysis. Therefore, we propose a two-step approach to test for an overall GxE effect for multiple phenotypes.

Results: Using simulations we demonstrate that, when more than one phenotype has GxE effect (i.e. GxE pleiotropy), our approach offers substantial gain in power (18–43%) to detect an aggregate-level GxE effect for a multivariate phenotype compared to an analogous two-step method to identify GxE effect for a univariate phenotype. We applied the proposed approach to simultaneously analyze three lipids, LDL, HDL and Triglyceride with the frequency of alcohol consumption as environmental factor in the UK Biobank. The method identified two loci with an overall GxE effect on the vector of lipids, one of which was missed by the competing approaches.

Availability and implementation: We provide an R package MPGE implementing the proposed approach which is available from CRAN: https://cran.r-project.org/web/packages/MPGE/index.html

Contact: jwitte@ucsf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene–environment (GxE) interactions contribute significantly to the genetic architecture underlying complex phenotypes (Dahl *et al.*, 2020). However, most GxE methods focus on testing a non-null effect of the interaction for one phenotype and one environmental factor at a time across genome-wide genetic variants (Gauderman *et al.*, 2017; Mukherjee *et al.*, 2012). Such methods include approaches to jointly testing marginal and interaction effects (Dai *et al.*, 2012a), empirical Bayes shrinkage methods (Mukherjee and Chatterjee, 2008), two-step approaches (Dai *et al.*, 2012b; Gauderman *et al.*, 2013; Hsu *et al.*, 2012; Zhang *et al.*, 2016), etc. While these approaches can increase power to detect GxE interactions, adequate power remains a concern. One possible approach to further increase the power of detecting GxE interactions is by modeling multiple related phenotypes together. Previous work indicates

that power to detect main genetic effects can be increased by modeling multiple correlated phenotypes; thus, one would expect similar gains to be available for assessing GxE interactions (Cornelis *et al.*, 2010; Zhang *et al.*, 2019).

There exists substantial shared genetic basis among different phenotypes (i.e. pleiotropy). Genome-wide association studies (GWAS) have shown overlap in the main genetic effects across various complex phenotypes. While extensive work has investigated approaches for assessing pleiotropy in main genetic effects (Bhattacharjee *et al.*, 2012; Galesloot *et al.*, 2014; Majumdar *et al.*, 2015, 2016, 2018; Ray *et al.*, 2016; Turley *et al.*, 2018), little has been done with regard to assessing pleiotropy in GxE effects. For example, an interaction between physical activity and a genetic variant can influence the levels of three lipids, LDL, HDL and Triglycerides, simultaneously (Kilpeläinen *et al.*, 2019). As another example, the pleiotropic genetic architecture of multiple smoking-related cancers (e.g. lung and head-neck) can be different among smokers and non-smokers (Jiang *et al.*, 2019; Schaal and Chellappan, 2014).

A recent study (Moore *et al.*, 2019) proposed a mixed-model approach to quantify the heritability of a complex phenotype explained due to GxE interaction across multiple environmental factors for a single phenotype. Another study (Yu *et al.*, 2018) proposed a subset-based multi-phenotype fixed-effects meta-analysis considering both marginal genetic effect and GxE effect across multiple phenotypes in the same model based on summary statistics of the corresponding effects. Another recent study (Zhang *et al.*, 2019) has proposed statistical methods for identifying aggregate-level GxE effect across multiple phenotypes for a gene instead of a single SNP. A simple strategy to test for an overall GxE effect across phenotypes is to perform a multivariate multiple linear regression including both the multivariate main genetic effect and the interaction effect terms in the model.

For a univariate phenotype, two-step methods can produce higher power for detecting GxE interactions compared to conventional approaches using a single analysis testing a GxE interaction (Dai *et al.*, 2012b; Gauderman *et al.*, 2013; Hsu *et al.*, 2012). Two-step approaches filter out less important genetic variants in the first step and test the more promising variants for GxE interaction in the second step to reduce the multiple testing burden. Among various strategies in the first step, a common approach is to test the SNPs for a marginal genetic association with the phenotype under the assumption that a SNP having a GxE interaction effect on the phenotype should also have a marginal genetic effect on the phenotype. Similarly, a two-step procedure for multivariate phenotypes should produce higher power for detecting an aggregate-level GxE effect compared to a simple one-step multivariate regression of testing an overall GxE effect across phenotypes.

In this article, we extend the two-step procedure to multivariate quantitative phenotypes, and investigate its relative performance compared to the one-step multivariate regression for testing an overall GxE effect. Our motivation is two-fold: in the 1st step, while filtering less important SNPs, simultaneously testing multiple related phenotypes should offer higher power for detecting SNPs having an overall marginal genetic effect (pleiotropy in main genetic effect); and in the 2nd step, testing such selected promising SNPs for an aggregate-level GxE effect on the multiple phenotypes should produce higher power due to pleiotropy in GxE effect across the phenotypes.

To adjust for multiple testing in the one-step and two-step approaches, we considered three different procedures: Bonferroni correction, subset testing and weighted hypothesis testing. We demonstrate by simulations that the multivariate two-step approach has a substantial power gain over the competing approaches. For real data application, we implement our approach to identify overall GxE effect of genome-wide SNPs and frequency of alcohol consumption on three lipids (LDL, HDL, Triglycerides) in the UK Biobank.

2 Materials and methods

We consider a cohort with individual-level data on a vector of multiple phenotypes, an environmental factor and genotypes of genome-wide SNPs. For each SNP, our approach consists of two steps applied on the same set of individuals. In the first step, we perform a test for marginal overall genetic association between the SNP and the multivariate phenotype. SNPs which show an evidence of overall genetic association are prioritized while testing for an overall GxE effect in the second step, i.e. we consider a more liberal threshold of significance level for these promising SNPs compared to the remaining SNPs while testing GxE. While combining the two steps to identify the genome-wide significant SNPs with an overall GxE effect, we adopt two different strategies for multiple testing adjustment: subset testing and weighted hypothesis testing.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k)'$ be multiple continuous phenotypes in a cohort, G denote genotypes at a SNP and E an environmental factor. E can be of arbitrary type, e.g. binary, categorical or continuous. We consider multivariate linear regression (MLR) to model the main genetic effect of the SNP on Y.

$$E(\boldsymbol{Y}) = \boldsymbol{\alpha} + \mathrm{G}\boldsymbol{\beta}_{\mathrm{G}}. \tag{1}$$

Here, $\boldsymbol{\beta}_{\mathrm{G}} = (\beta_G^{(1)}, \ldots, \beta_G^{(k)})'$ and $\boldsymbol{\alpha} = (\alpha^{(1)}, \ldots, \alpha^{(k)})'$, and the error component is assumed to follow a multivariate normal distribution with zero mean vector and covariance matrix Σ_1. In the first step of the two-step procedure, we implement MLR to assess the overall main genetic effect of the SNP. In particular, we test $H_0 : \beta_G^{(1)} = \cdots = \beta_G^{(k)} = 0$ versus $H_1 : \beta_G^{(j)} \neq 0$, for at least one $j = 1, \ldots, k$. We note that the power of identifying a SNP having a marginal genetic effect should improve by modeling multiple related phenotypes instead of a single phenotype. In the second step, we consider multivariate multiple linear regression (MMLR) to incorporate the multivariate main effects of the SNP (G) and the environmental factor (E), and the multivariate interaction effect due to GxE.

$$E(\boldsymbol{Y}) = \boldsymbol{\alpha} + \mathrm{G}\boldsymbol{\beta}_{\mathrm{G}} + \mathrm{E}\boldsymbol{\beta}_{\mathrm{E}} + \mathrm{GE}\boldsymbol{\beta}_{\mathrm{GE}}. \tag{2}$$

Here, $\boldsymbol{\beta}_{\mathrm{E}} = (\beta_E^{(1)}, \ldots, \beta_E^{(k)})'$ and $\boldsymbol{\beta}_{\mathrm{GE}} = (\beta_{GE}^{(1)}, \ldots, \beta_{GE}^{(k)})'$, and the error component is assumed to follow multivariate normal with zero mean vector and a covariance matrix Σ_2. We implement the type II MANOVA to test $H_0 : \beta_{GE}^{(1)} = \cdots = \beta_{GE}^{(k)} = 0$ versus $H_0 : \beta_{GE}^{(j)} \neq 0$, for at least one j. In the type II MANOVA test, the following two models are compared: the unrestricted full model, $E(\boldsymbol{Y}) = \boldsymbol{\alpha} + \mathrm{G}\boldsymbol{\beta}_{\mathrm{G}} + \mathrm{E}\boldsymbol{\beta}_{\mathrm{E}} + \mathrm{GE}\boldsymbol{\beta}_{\mathrm{GE}}$, versus the restricted model, $E(\boldsymbol{Y}) = \boldsymbol{\alpha} + \mathrm{G}\boldsymbol{\beta}_{\mathrm{G}} + \mathrm{E}\boldsymbol{\beta}_{\mathrm{E}}$. Here, the unrestricted model reduces to the restricted model under H_0, when $\boldsymbol{\beta}_{\mathrm{GE}} = \boldsymbol{0}$. Thus, in the second step, we only test the null hypothesis that the vector of interaction effects $\boldsymbol{\beta}_{\mathrm{GE}} = \boldsymbol{0}$, leaving the vectors of main effects, $\boldsymbol{\beta}_{\mathrm{G}}$ and $\boldsymbol{\beta}_{\mathrm{E}}$, unrestricted. The power of detecting a GxE interaction effect should be increased if the interaction is shared across $Y_1, \ldots, Y_k$. We use the R package 'car' (Fox and Weisberg, 2018) to perform type II MANOVA.

In the two-step procedure, we combine the P-values obtained from 1st and 2nd steps to identify the SNPs that have a non-null overall GxE effect. We note that the linear model in Equation (1) is nested under the linear model in Equation (2). Hence, due to the general result in Dai *et al.* (2012b), the test statistic in the screening step to test $\boldsymbol{\beta}_{\mathrm{G}} = \boldsymbol{0}$ (Equation 1) and the test statistic testing $\boldsymbol{\beta}_{\mathrm{GE}} = \boldsymbol{0}$ in the second step (Equation 2) are independently distributed. This property is crucial to maintain the overall false positive rate of the combined two-step procedure at a desired level of significance. An important rationale behind expecting higher power from our two-step approach is that a SNP having a GxE effect should also have a marginal genetic effect. However, if this assumption does not hold for a SNP, one-step approaches would be more powerful. Below, we outline the subset testing (sst) approach and the weighted hypothesis testing (wht) approach to combine the two steps while adjusting for multiple testing to maintain the overall false positive rate.

2.1 Adjustment for multiple testing

Suppose we are considering m SNPs, and that we have two sets of P-values. One set is obtained from the first step testing for an overall main genetic effect across phenotypes (equation 1), $\boldsymbol{P}_G = (P_G^{(1)}, P_G^{(2)}, \ldots, P_G^{(m)})$; and the other set from the second step for the multivariate interaction effect (Equation 2), $\boldsymbol{P}_{GE} = (P_{GE}^{(1)}, P_{GE}^{(2)}, \ldots, P_{GE}^{(m)})$.

For the one-step multivariate GxE test, Bonferroni correction is applied to $\boldsymbol{P}_{GE}$. For the two-step approach, we consider the following two well-known procedures to combine the two steps (Gauderman *et al.*, 2013; Ionita-Laza *et al.*, 2007).

2.1.1 Subset testing

For $\boldsymbol{P}_G$, we consider a P-value threshold α_1 and filter out all SNPs for which $P_G^{(i)} > \alpha_1$, $i = 1, \ldots, m$. In the second step, we only consider the SNPs selected in the first step ($P_G < \alpha_1$), and apply a Bonferroni correction while testing for an overall GxE effect for these SNPs. Suppose, we consider a P-value threshold α_2 in the second step. If m_1 SNPs pass the first step, we compare P_{GE} to $\frac{\alpha_2}{m_1}$ for each of the selected m_1 SNPs to identify the SNPs having an overall GxE effect. A larger choice of α_1 will increase the possibility of sending the SNPs with a true GxE effect to the second step, but at the expense of a higher multiple testing burden in the second step (Gauderman *et al.*, 2013). Guided by Kooperberg and LeBlanc (2008), we considered the following choice of the P-value thresholds, $\alpha_1 = 0.005$ in step 1 and the standard choice of $\alpha_2 = 0.05$ in step 2.

2.1.2 Weighted hypothesis testing

Instead of completely dropping a set of less important SNPs in the second step, it has been argued that testing all SNPs in the second step while prioritizing them according to their relative ranking of importance obtained in the first step produces higher power to detect a GxE effect for a univariate phenotype (Hsu *et al.*, 2012; Ionita-Laza *et al.*, 2007; Wasserman and Roeder, 2006; Zhang *et al.*, 2016). Thus, we follow this approach and test all m SNPs using P_{GE} in step 2 based on a significance level weighted using the order of the P-values in step 1 (P_G). The weighting scheme uses an exponential weighting function, and allocates a larger fraction of the total significance level α to the most significant SNPs obtained in step 1 (Hsu *et al.*, 2012; Ionita-Laza *et al.*, 2007; Zhang *et al.*, 2016). In particular, while performing step 2, the k_1 most significant SNPs in the first bin in step 1 (lowest P_G) are tested at a significance level $\frac{1}{2k_1}\alpha$, the next k_2 ($= 2k_1$) most significant SNPs in the second bin in step 1 are tested at $\frac{1}{2^2k_2}\alpha$, the next k_3 ($= 2k_2$) at $\frac{1}{2^3k_3}\alpha$, and so on (Ionita-Laza *et al.*, 2007). For example, when $k_1 = 5$ and $\alpha = 0.05$, the top 5 SNPs from step 1 are tested at a significance level 0.005 in step 2, the next 10 at 0.00125, etc. This weighting scheme guarantees that the overall false positive rate for the entire procedure does not exceed α. Under this weighting scheme, the top SNPs from step 1 are tested at a more liberal significance threshold than the standard Bonferroni-corrected level required in a standard one-step exhaustive scan of all m SNPs. However, for the SNPs not in the top bins in step 1, weighted testing can have a more stringent threshold than Bonferroni correction. We used a standard choice of $k_1 = 5$ and $\alpha = 0.05$ (Hsu *et al.*, 2012; Ionita-Laza *et al.*, 2007; Zhang *et al.*, 2016).

2.2 GxE tests for univariate phenotype

To test for a GxE interaction for a univariate phenotype, we consider the following existing methods (Dai *et al.*, 2012b; Gauderman *et al.*, 2013; Zhang *et al.*, 2016). Let Y denote a single continuous phenotype. In the one-step approach, we consider $E(Y) = \alpha + \beta_G \times G + \beta_E \times E + \beta_{GE} \times GE$, and test for $H_0 : \beta_{GE} = 0$ versus $H_1 : \beta_{GE} \neq 0$. In the two-step approach, we combine the step 1 model: $E(Y) = \alpha + \beta_G \times G$, with step 2 model: $E(Y) = \alpha + \beta_G \times G + \beta_E \times E + \beta_{GE} \times GE$. Here we consider the same multiple testing strategies as considered above for a multivariate phenotype.

3 Simulation study

3.1 Framework

We describe the simulation design for two phenotypes mainly for convenience in presenting the mathematical expressions. This can be extended for a larger number of phenotypes in a straightforward manner. Let Y_1 and Y_2 denote two phenotypes, G denote the genotypes at a SNP and E an environmental factor. We consider the following bivariate multiple linear regression to model the phenotypes.

$$(Y_1\, Y_2) = \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} + \begin{pmatrix}\beta_G^{(1)}\\ \beta_G^{(2)}\end{pmatrix} G + \begin{pmatrix}\beta_E^{(1)}\\ \beta_E^{(2)}\end{pmatrix} E + \begin{pmatrix}\beta_{GE}^{(1)}\\ \beta_{GE}^{(2)}\end{pmatrix} G \times E + \begin{pmatrix}\epsilon_1\\ \epsilon_2\end{pmatrix}. \quad (3)$$

We consider each of Y_1, Y_2, G and E to be mean-centered. We assume a bivariate normal distribution for $(\epsilon_1, \epsilon_2)'$. Under a fixed effects model, $V(Y_j) = \beta_G^{2(j)} V(G) + \beta_E^{2(j)} V(E) + \beta_{GE}^{2(j)} V(G \times E) + \sigma^2_{\epsilon_j}$, $j = 1, 2$. Under the assumption that G and E are independent in the population, we obtain that $V(G \times E) = V(G)V(E)$, since $E(G) = E(E) = 0$. Thus, $V(Y_j) = \beta_G^{2(j)} V(G) + \beta_E^{2(j)} V(E) + \beta_{GE}^{2(j)} V(G)V(E) + \sigma^2_{\epsilon_j}$, for $j = 1, 2$.

Let us denote $h_G^{2(j)} = \beta_G^{2(j)} V(G)$, $h_E^{2(j)} = \beta_E^{2(j)} V(E)$, $h_{GE}^{2(j)} = \beta_{GE}^{2(j)} V(G)V(E)$, and the total variance of j^{th} phenotype as $\sigma^2_{Y_j} = V(Y_j)$. Hence, $\sigma^2_{Y_j} = h_G^{2(j)} + h_E^{2(j)} + h_{GE}^{2(j)} + \sigma^2_{\epsilon_j}$. Without loss of generality, we assume that $\sigma^2_{Y_j} = 1$; so, $\sigma^2_{\epsilon_j} = 1 - (h_G^{2(j)} + h_E^{2(j)} + h_{GE}^{2(j)})$; $j = 1, 2$. Next, we derive the following: $cov(Y_1, Y_2) = cor(Y_1, Y_2) = \beta_G^{(1)} \beta_G^{(2)} V(G) + \beta_E^{(1)} \beta_E^{(2)} V(E) + \beta_{GE}^{(1)} \beta_{GE}^{(2)} V(G)V(E) + cov(\epsilon_1, \epsilon_2)$, where $cov(\epsilon_1, \epsilon_2)$ is the covariance between the noise terms in the two phenotypes. Thus, we can first fix the correlation between the phenotypes ($cor(Y_1, Y_2)$) and other simulation parameters which in turn determine the value of $cov(\epsilon_1, \epsilon_2)$ to be used in the simulations.

We simulate the genotype data at each SNP under the Hardy-Weinberg equilibrium using the probabilities of the three possible genotypes as: $P(AA) = p^2, P(Aa) = 2p(1-p), P(aa) = (1-p)^2$, where $P(A) = p$. We simulate the environmental factor as a binary random variable with $P(E = 1) = f$ and $P(E = 0) = 1 - f$.

3.2 Choice of parameters

In our simulation study, we consider three phenotypes for 20 000 individuals, and choose the pairwise phenotypic correlations randomly in the range 20–30%. We simulate the minor allele frequency at a SNP from Uniform $(0.05, 0.45)$, and the proportion of the reference category of the environmental factor, $f = P(E = 1)$, from Uniform $(0.2, 0.3)$. For each risk SNP with a marginal genetic effect on Y_j, we simulated $h_G^{2(j)}$ randomly between 0.1% and 0.2% so that, in aggregate, 100 such SNPs explain an average of 15% of the variance of $Y_j, j = 1, 2, 3$. When E has an effect on Y_j, we choose $h_E^{2(j)}$, the proportion of variance of Y_j explained due to E, randomly between 1% and 2%. Similarly, we randomly simulate $h_{GE}^{2(j)}$ in the range (0.01–0.05%) so that, in aggregate, 40 risk SNPs having a GxE effect on Y_j explain an average of 1.2% of the variance of Y_j.

We consider 100 000 null SNPs which have no marginal genetic association with any phenotype, and no GxE interaction on any phenotype. We consider a separate set of 100 non-null SNPs (denoted by m_G) each of which has a marginal genetic effect on at least one phenotype (Equation 1). Among these m_G non-null SNPs, m_{GE} SNPs have a GxE effect on at least one phenotype (Equation 2), and we vary $m_{GE} = 10, 20, 30, 40$. So, a subset of the risk SNPs having a marginal genetic effect are assumed to have a GxE effect (m_{GE} out of $m_G = 100$). We further assume that, if a SNP has a GxE effect on a phenotype, the same phenotype also has a marginal genetic effect due to the SNP. We consider three different scenarios. In the 1st scenario, each non-null SNP has a marginal genetic effect on the first phenotype but not the other two phenotypes; and if the SNP (one of m_{GE} SNPs) has a GxE effect, it has the interaction effect only on the first phenotype. Similarly in the 2nd scenario, the first two phenotypes (but not the last phenotype) have a marginal genetic effect due to each non-null SNP, and each of m_{GE} SNPs has a GxE effect on the first two phenotypes but not the last one. And in the 3rd scenario, all the three phenotypes have a marginal genetic effect from each non-null SNP, and each of m_{GE} SNPs has a GxE effect on every phenotype.

Under each scenario, we compare the performance of various tests of GxE interaction for each univariate phenotype and the multivariate phenotype. We apply three different procedures for multiple

testing adjustment to control the family-wise error rate (FWER) as outlined above: Bonferroni correction for the one-step methods (abbreviated as bonf), subset testing (sst) and weighted hypotheses testing (wht). For each method, we estimate the type I error rate and power based on 200 simulated datasets. Under a given simulation scenario, for each simulated dataset, we compute the proportion of the null SNPs at which the method of choice to test GxE (null hypothesis of no overall/univariate GxE effect) wrongly identified a genome-wide significant signal of interaction. We estimate the type I error rate as the mean of this proportion across 200 simulated datasets. We estimate the power using a similar procedure based on the risk SNPs only. It is important to explore whether we obtain higher power by testing GxE interaction for a multivariate phenotype compared to each univariate phenotype. For comparison, we can plot the power curve for the multivariate approach and the univariate approach for Y_1, Y_2, Y_3. However for the univariate approach, including three power curves for Y_1, Y_2, Y_3 may look repetitive without providing new insights. Hence, we instead used the following simple strategy. Under a given simulation scenario, let p_1, p_2, p_3 denote the power estimated by the univariate approach for Y_1, Y_2 and Y_3, respectively; and p denotes the power obtained by the multivariate approach. If $p > \max(p_1, p_2, p_3)$, the multivariate approach has higher power than the univariate approach for each of Y_1, Y_2, Y_3. Hence, for ease of presentation, when plotting the power obtained by the tests of GxE interaction for univariate phenotypes, we plot the maximum power obtained across three univariate phenotypes for each multiple testing adjustment procedure. Because, the main aim here is to compare the power of the multivariate approach with that of the univariate approach for each of the three phenotypes.

While evaluating the FWER in the above simulation scenarios, we assumed that the environmental factor has a marginal effect ($\boldsymbol{\beta}_{\mathrm{E}} \neq 0$) on the phenotypes (Equation 2). Under this assumption, we estimated FWER based on null SNPs ($\boldsymbol{\beta}_{\mathrm{GE}} = 0$) consisting of 100K SNPs which have neither a marginal genetic effect (referred as G effect in the following) nor a GxE effect, and (m_G-m_{GE}) null SNPs which have a G effect but no GxE effect. Here, $m_G = 100$ and $m_{GE} = 10, 20, 30, 40$. Next, we estimated FWER based on only the 100K SNPs that have neither G nor GxE effect. For a comprehensive evaluation of FWER, we also considered the following simulation scenarios. E has no marginal effect ($\boldsymbol{\beta}_{\mathrm{E}} = 0$), and FWER is estimated based on 100K null SNPs with no G and GxE effect. In the absence of a marginal effect of E, we also estimated FWER based on 100K SNPs having no G and GxE effect and (m_G-m_{GE}) SNPs with a G effect but no GxE effect.

In the above simulation scenarios when E has a marginal effect on the phenotypes, we assumed that the environmental factor is distributed independently of SNP genotypes. However, some null SNPs (no GxE effect) can be associated with the environmental factor. We assume that the number of SNPs associated with the environmental factor should be smaller than the number of SNPs (100) associated with the main phenotypes. Hence, under the above simulation scenarios, we also considered that 40 randomly selected null SNPs have a marginal genetic effect on E. To induce such an association, we simulated the binary E using a logistic regression model: $\mathrm{P}(E = 1|g_1, \ldots, g_{40}) = \frac{\exp(\sum_{j=1}^{40} \gamma_j g_j)}{1+\exp(\sum_{j=1}^{40} \gamma_j g_j)}$, where g_j is the genotype of the j^{th} selected SNP with an effect size γ_j. We simulated the odds ratios $\exp(\gamma_j), j = 1, \ldots, 40$, uniformly from $(1.1 - 1.3)$ or $(1/1.3 - 1/1.1)$ depending on the direction of association (positive/negative) chosen at random with probability half for each selected SNP. Under these simulation scenarios, we evaluate the overall type I error rate and power obtained by different univariate and multivariate approaches.

Table 1. Simulation results: estimated overall type I error rate obtained by different tests of overall GxE effect for a multivariate phenotype using various strategies of multiple testing adjustment

		1-step bonf	*2-step* subset	*2-step* weighted
#asso-pheno	m_{GE}	Correction	Testing	Hypothesis testing
1	10	0.04	0.04	0.1
1	20	0.05	0.04	0.06
1	30	0.09	0.06	0.04
1	40	0.06	0.05	0.03
2	10	0.04	0.04	0.06
2	20	0.06	0.06	0.06
2	30	0.08	0.04	0.02
2	40	0.06	0.07	0.04
3	10	0.03	0.05	0.04
3	20	0.03	0.05	0.05
3	30	0.06	0.02	0.02
3	40	0.07	0.07	0

Note: Here, #asso-pheno denotes the number of phenotypes that have a marginal genetic effect or an interaction effect due to risk SNPs; m_{GE} denotes the number of SNPs out of 100 (m_G) risk SNPs that have a GxE interaction effect. We present the overall type I error rate (FWER) at 0.05 level of significance with the desired level 5×10^{-7} per SNP, since we considered 100K null SNPs which have no marginal G or GxE effect. The type I error rate is estimated based on 200 simulated datasets with 20000 individuals in the sample.

3.3 Results

First, we present the estimated overall type I error rate (FWER) obtained from GxE tests for a multivariate phenotype. We present the FWER in Table 1 when the environmental factor E has a marginal effect on the phenotypes. Here, the FWER is estimated based on a set of null SNPs comprising 100K SNPs which have neither a G effect nor a GxE effect, and m_G-m_{GE} SNPs which have a G effect but no GxE effect. While the FWER appears to be controlled overall at the desired level of significance 0.05 (Table 1), we observe marginal inflation in some cases; this is mainly due to using 200 iterations of simulation (for computational feasibility) to estimate the FWER under a given simulation scenario. We present the estimated FWER of GxE tests for univariate phenotypes in Table 2 under the same simulation scenarios, and find that the FWER is controlled overall with marginal inflation in some cases. The FWER estimated based on only 100K null SNPs which have no G and GxE effect is controlled overall both for multivariate phenotype and univariate phenotypes (Supplementary Table S1). Interestingly, *2-step* procedure based on weighted hypothesis testing appears to control the FWER conservatively in this scenario (Supplementary Table S1).

When E has no effect on the phenotypes, the FWER estimated using 100K null SNPs with no G and GxE effect and another m_G-m_{GE} null SNPs with a G but no GxE effect is reasonably well controlled both for multivariate and univariate phenotypes (Supplementary Table S2). In this scenario, the FWER estimated based on only 100K null SNPs which have no G and GxE effect is overall controlled both for multivariate and univariate phenotypes (Supplementary Table S3). The *2-step* procedure based on weighted hypothesis testing controls FWER conservatively for some cases in this scenario.

In the simulation scenarios when 40 null SNPs (no GxE effect) are associated with the environmental factor, the FWER, estimated based on the null SNPs comprising 100K SNPs with no G and GxE effect and m_G-m_{GE} SNPs with a G effect but no GxE effect, was overall controlled with marginal inflation in some cases (Supplementary Table S4). We also find that the *2-step* approach based on weighted hypothesis testing controls the FWER overall better than the other approaches in this scenario (Supplementary Table S4).

We present the estimated power of GxE tests for multivariate and univariate phenotypes inFigures 1–3. First, we focus on the multivariate phenotype, and compare the power of *1-step* and *2-step* approaches to detect an overall effect of GxE interaction on multiple phenotypes. We find that both of the *2-step* procedures (subset testing and weighted hypothesis testing) produce higher power than the *1-step* approach (Bonferroni correction). We also observe that the weighted hypothesis testing (wht) performs better than the subset testing (sst) (Figs 1–3). We therefore focus on comparing the weighted hypothesis testing procedure with the Bonferroni correction to contrast the power of *2-step* and *1-step* approaches. In 3rd

Table 2. Simulation results: estimated overall type I error rate obtained by different tests of GxE effect for univariate phenotype using various strategies of multiple testing adjustment

		1-step bonf correction			*2-step* subset testing			*2-step* weighted hypothesis testing		
#asso-pheno	m_{GE}	pheno1	pheno2	pheno3	pheno1	pheno2	pheno3	pheno1	pheno2	pheno3
1	10	0.06	0.06	0.06	0.05	0.04	0.05	0.06	0.06	0.07
1	20	0.06	0.02	0.06	0.04	0.06	0.03	0.04	0.07	0.05
1	30	0.06	0.04	0.08	0.08	0.07	0.05	0.03	0.09	0.07
1	40	0.04	0.04	0.07	0.08	0.04	0.04	0.06	0.06	0.02
2	10	0.07	0.04	0.04	0.04	0.06	0.02	0.04	0.04	0.06
2	20	0.03	0.07	0.03	0.06	0.04	0.06	0.06	0.02	0.04
2	30	0.05	0.04	0.06	0.03	0.04	0.06	0.04	0.02	0.04
2	40	0.04	0.05	0.07	0.06	0.05	0.04	0.04	0.06	0.05
3	10	0.05	0.04	0.03	0.08	0.06	0.08	0.06	0.04	0.04
3	20	0.04	0.04	0.05	0.04	0.04	0.04	0.03	0.03	0.06
3	30	0.08	0.08	0.05	0.06	0.06	0.03	0.02	0.05	0.02
3	40	0.08	0.07	0.03	0.06	0.05	0.04	0.03	0.03	0.02

Note: Here, #asso-pheno denotes the number of phenotypes that have a marginal genetic effect or an interaction effect due to risk SNPs; m_{GE} denotes the number of SNPs out of 100 (m_G) risk SNPs that have a GxE interaction effect. We present the overall type I error rate (FWER) at 0.05 level of significance with the desired level 5×10^{-7} per SNP, since we considered 100K null SNPs which have no marginal G or GxE effect. Three univariate phenotypes are abbreviated as pheno1, pheno2 and pheno3, respectively. The type I error rate is estimated based on 200 simulated datasets with 20000 individuals in the sample.

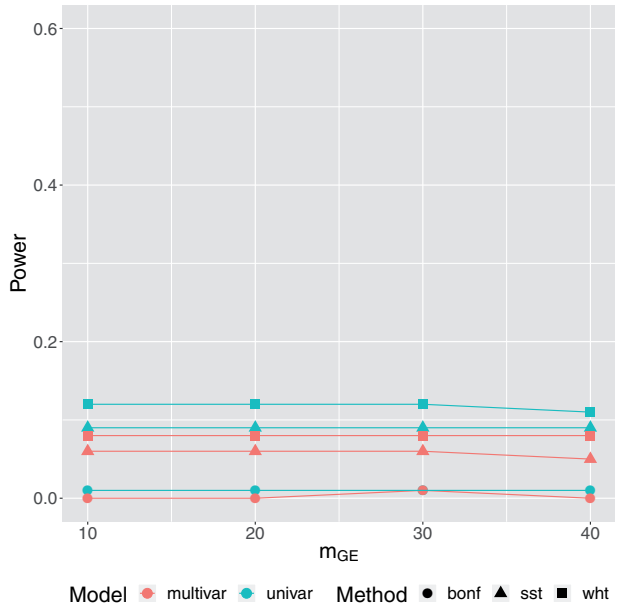

Fig. 1. Simulation results: estimated power obtained by different tests of overall GxE effect for multivariate phenotype (multivar), and tests of GxE effect for univariate phenotype (univar) using various strategies of multiple testing adjustment: *1-step* Bonferroni correction (bonf), *2-step* subset testing (sst) and *2-step* weighted hypothesis testing (wht). Here, 1st phenotype (but not 2nd and 3rd) has a marginal genetic effect or an interaction effect due to risk SNPs. We denote the number of SNPs out of 100 risk SNPs which have a GxE effect as m_{GE}. The power is estimated based on 200 simulated datasets

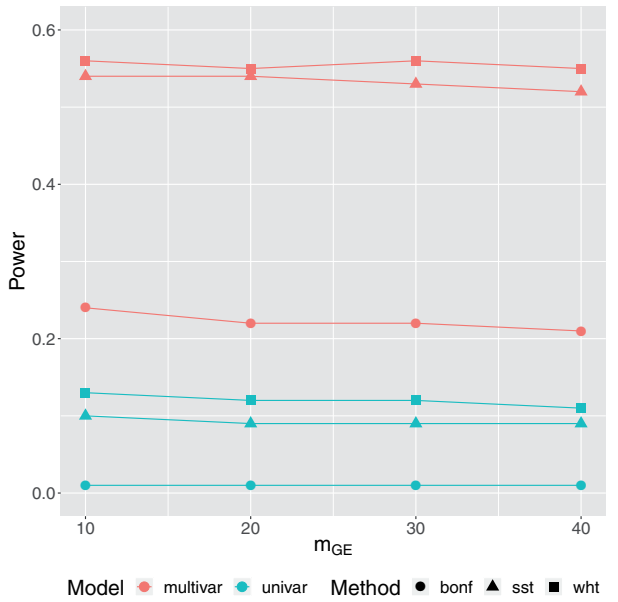

Fig. 3. Simulation results: estimated power obtained by different tests of overall GxE effect for multivariate phenotype (multivar), and tests of GxE effect for univariate phenotype (univar) using various strategies of multiple testing adjustment: *1-step* Bonferroni correction (bonf), *2-step* subset testing (sst) and *2-step* weighted hypothesis testing (wht). Here, all three phenotypes have a marginal genetic effect or an interaction effect due to risk SNPs. We denote the number of SNPs out of 100 risk SNPs which have a GxE effect as m_{GE}. The power is estimated based on 200 simulated datasets

simulation scenario, when all three phenotypes have a GxE effect from each of m_{GE} SNPs, the *2-step* approach implemented by weighted hypothesis testing (abbreviated as *2-step*-wht) produces 32–34% higher power than the Bonferroni correction implementing the *1-step* approach (*1-step*-bonf) (Fig. 3). In 2nd scenario, when two phenotypes have a GxE effect, *2-step*-wht produces 23–24% power increase than *1-step*-bonf (Fig. 2). In the absence of pleiotropy in GxE effect, i.e. when one phenotype has a GxE effect, *2-step*-wht yields 7–8% power gain than *1-step*-bonf (Fig. 1). Hence, overall our *2-step* approach offers substantial power gain compared to the *1-step* approach when testing for an overall GxE effect on a multivariate phenotype. We also find that *2-step*-wht performs marginally better than *2-step*-sst (subset testing) and produces a power gain of 1–3% in the third simulation scenario (Fig. 3), 3–4% in the second scenario (Fig. 2), and 2–3% in the first scenario (Fig. 1).

For GxE tests with univariate phenotypes, the *2-step* approaches (wht and sst) perform better than the *1-step* approach (Bonferroni correction) which is consistent with findings from previous studies (Gauderman *et al.*, 2013; Zhang *et al.*, 2016). Between the two strategies of *2-step* approaches for univariate phenotype, wht produces marginally higher power than sst.

Next, we contrast the performance of *2-step*-wht for multivariate phenotypes (multivar_wht) with that of *2-step*-wht for

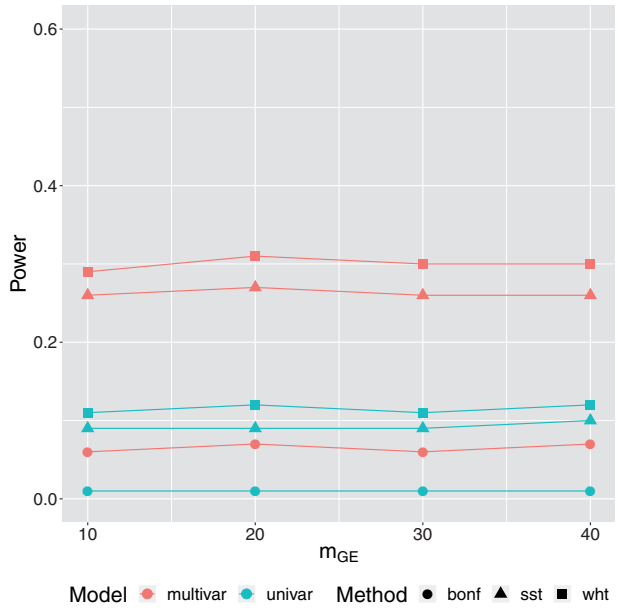

Fig. 2. Simulation results: estimated power obtained by different tests of overall GxE effect for multivariate phenotype (multivar), and tests of GxE effect for univariate phenotype (univar) using various strategies of multiple testing adjustment: *1-step* Bonferroni correction (bonf), *2-step* subset testing (sst) and *2-step* weighted hypothesis testing (wht). Here, first two phenotypes (but not 3rd) have a marginal genetic effect or an interaction effect due to risk SNPs. We denote the number of SNPs out of 100 risk SNPs which have a GxE effect as m_{GE}. The power is estimated based on 200 simulated datasets

univariate phenotypes (univar_wht). We contrast the power obtained by multivar_wht with the maximum power obtained across the three univariate phenotypes each obtained by univar_wht (as discussed above). In the presence of pleiotropy in the GxE effect, multivar_wht produces 44% higher power than univar_wht under the third simulation scenario (Fig. 3), and 18–19% power gain under the second scenario (Fig. 2). However, in the absence of pleiotropy in GxE effect under the first simulation scenario, univar_wht offers marginal power gain (3–4%) over multivar_wht (Fig. 1). Taken together, the multivariate approach produces substantially higher power than the univariate approach in the presence of pleiotropy in GxE effect, and loses marginal power in its absence.

When some null SNPs have a marginal effect on the environmental factor, the comparative performance of the multivariate and univariate approaches with respect to power of detecting GxE effect remain similar (Supplementary Figs S1–S3).

We note that in Figures 1–3, the power curve for a given method remains flat irrespective of the total number of GxE risk SNPs included in the phenotype simulation model. In each iteration under a simulation scenario, we simulated the proportion of variance explained in the phenotype due to the GxE effect per SNP uniformly from a fixed range. Hence, on average across the iterations, each GxE SNP has a similar GxE heritability under a given simulation scenario. If the total number of GxE SNPs increases, the total GxE heritability of the phenotype increases, but per-SNP GxE heritability remains similar. As a result, the estimated power curve remains flat irrespective of the number of GxE SNPs considered. We also repeated the power comparison corresponding to Figures 1–3 fixing the number of GxE risk SNPs as 30 while simulating the pairwise phenotypic correlation uniformly from the following increasing ranges: (0.1-0.2), (0.2-0.3), (0.3-0.4), (0.4-0.5). As expected, the estimated power of a multivariate approach overall increases as the pairwise phenotypic correlation increases (Supplementary Figs S4–S6).

In our simulation design, we always assumed that each risk SNP with a GxE effect also has a marginal genetic effect on the phenotypes. However, if a GxE SNP does not have a marginal genetic effect, a one-step approach will be more powerful to detect such GxE signals compared to a two-step approach. More specifically, a two-step method based on subset testing will miss such SNPs, because it transfers a SNP to the second step for GxE testing only if the SNP shows a marginal genetic association in the first step. Also, a two-step method based on weighted hypothesis testing will allocate such a SNP near the tail of the relative ranking of importance in the first step; and hence is likely to miss the GxE signal in the second step due to a stringent threshold of significance level used for the SNP.

4 Real data application

We considered three lipids LDL, HDL and Triglycerides in the UK Biobank as the multivariate phenotype, and the frequency of alcohol consumption as the environmental factor. We note that LDL was measured directly instead of using Friedewald equation. We removed individuals with missing values of phenotypes or relevant covariates from the sample, leaving 253 653 White-British unrelated individuals. While a similar proportion of individuals belong to the two different sex categories, the maximum proportion of individuals (44%) belong to the age group of 55–65 (Supplementary Table S5). First, we applied the log-transformation on the observed values of each lipid. Next, we adjust each log-transformed lipid for age, sex and 20 principal components (PCs) of genetic ancestry by linear regression. Finally, we apply the inverse rank normal transformation on the adjusted residuals obtained from the linear regression for each lipid, and considered them as the final phenotype vectors, $Y_j, j = 1, 2, 3$. The frequency of alcohol consumption was coded as an integer-valued variable (UK Biobank Data-Coding 100402) with the following categories: 1 (daily or almost daily), 2 (three or four times a week), 3 (once or twice a week), 4 (one to three times a month), 5 (special occasions only), 6 (never).

Genotype data in UK Biobank were assayed using two similar genotyping arrays. A subset of 49 950 individuals in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study were genotyped using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix. Subsequently, 438 427 participants were genotyped using the closely related Applied Biosystems UK Biobank Axiom Array (Bycroft *et al.*, 2018). We removed SNPs with minor allele frequency < 0.01, and genotype missingness > 0.01. Next we removed SNPs deviating from Hardy Weinberg Equilibrium based on a *P*-value threshold 10^{-7}. We finally tested 459 792 genotyped SNPs in the UKB one at a time. We applied the different procedures presented above for testing the GxE interaction. We provide the results from multivariate and univariate tests of GxE interaction in Table 3. QQ plots for the genome-wide GxE testing (Supplementary Fig. S10) indicate that adjusting for the selected relevant covariates was adequate for both testing an overall GxE effect on the lipids (using multivariate multiple linear regression) and testing GxE effect on each lipid separately (using a multiple linear regression). For every GxE testing approach, we used a threshold of the overall type I error rate (FWER) as 0.05. For subset testing in a 2-step procedure, we considered a *P*-value threshold of 0.005 to assess marginal genetic association in the first step. We also note that the adjusted GxE test *P*-values obtained from the 2-step method based on weighted hypothesis testing (Table 3) should be compared to the FWER level 0.05, not the Bonferroni corrected significance level using the total number of SNPs tested.

For eight SNPs on chromosome 1 and 8, the *2-step* weighted hypothesis testing (*2-step*-wht) identified a genome-wide significant overall effect of GxE interaction on the lipids, whereas the *2-step* subset testing (*2-step*-sst) approach identified an overall GxE effect for only the three SNPs on chromosome 8, also identified by *2-step*-wht (Table 3). The *1-step* Bonferroni correction (*1-step*-bonf) did not detect any genome-wide significant signal of an overall GxE effect. All the SNPs on chromosome 1 and 8 detected by *2-step*-wht are in linkage disequilibrium (LD). We identified the lead SNP (highlighted with bold font) on each chromosome based on r^2 threshold of 0.2 (Table 3). Thus, *2-step*-wht detected two loci on chromosome 1 and 8 to have an overall GxE effect; *2-step*-sst detected the GxE signal only on chromosome 8.

Table 3. Real data results: genome-wide significant signals of aggregate-level GxE interaction effect on the vector of three lipids (LDL, HDL, Triglycerides) and univariate GxE effect on HDL obtained by the *2-step* multivariate and univariate approaches, respectively

Multivariate 2-step weighted hypothesis testing for lipids					
SNP	CHR	BP	P.g (MLR)	P.ge (MMLR)	P.ge.wht
rs7528419	1	109817192	< E-300	0.001	0.01
rs12740374	1	109817590	< E-300	0.002	0.02
rs660240	1	109817838	< E-300	0.001	0.01
rs629301	1	109818306	< E-300	0.001	0.01
rs646776	1	109818530	< E-300	0.001	0.01
rs6984305	8	9178268	9.09E-112	6.64E-07	0.007
rs6601299	8	9184691	6.06E-106	1.94E-06	0.02
rs2126259	8	9185146	3.40E-108	1.91E-06	0.02
Multivariate 2-step subset testing for lipids					
			P.g (MLR)	P.ge (MMLR)	
rs6984305	8	9178268	9.09E-112	6.64E-07	
rs6601299	8	9184691	6.06E-106	1.94E-06	
rs2126259	8	9185146	3.40E-108	1.91E-06	
Univariate 2-step weighted hypothesis testing for HDL					
SNP	CHR	BP	P.g (ULR)	P.ge (UMLR)	P.ge.wht
rs6984305	8	9178268	1.23E-78	1.40E-07	0.0004
rs9987289	8	9183358	6.20E-99	9.43E-06	0.006
rs6601299	8	9184691	1.29E-79	9.68E-07	0.002
rs2126259	8	9185146	1.00E-81	7.93E-07	0.002
rs11779870	8	9211723	1.50E-51	1.33E-05	0.03
Univariate 2-step subset testing for HDL					
SNP	CHR	BP	P.g (ULR)	P.ge (UMLR)	
rs6984305	8	9178268	1.23E-78	1.40E-07	
rs6601299	8	9184691	1.29E-79	9.68E-07	
rs2126259	8	9185146	1.00E-81	7.93E-07	

Note: The frequency of alcohol consumption is considered as the environmental factor. CHR denotes chromosome, and BP denotes base pair position. In multivariate analysis, P.g (MLR) denotes the *P*-value of testing the marginal multivariate genetic association between the SNP and the lipids using multivariate linear regression (MLR); P.ge (MMLR) denotes the *P*-value of testing overall GxE effect on the lipids using multivariate multiple linear regression (MMLR) prior to adjustment for multiple testing; P.ge.wht denotes the adjusted *P*-value of testing the overall GxE effect obtained by the *2-step* approach based on weighted hypothesis testing. In univariate analysis, P.g (ULR) denotes the *P*-value of testing univariate marginal genetic association between the SNP and HDL using univariate linear regression (ULR); P.ge (UMLR) denotes the *P*-value of testing univariate GxE effect on HDL using univariate multiple linear regression (UMLR) prior to adjustment for multiple testing; P.ge.wht denotes the adjusted *P*-value of testing univariate GxE effect using *2-step* weighted hypothesis testing. The lead SNPs, i.e. the independent SNPs having the strongest genome-wide significant signal of GxE effect, are highlighted with bold font. The lead SNPs were obtained based on r^2 threshold 0.2. P.ge.wht is the adjusted *P*-value obtained by the *2-step* weighted hypothesis testing and was compared to the FWER level 0.05.

In the univariate GxE analysis, *1-step* Bonferroni correction did not identify any genome-wide significant GxE effect. *2-step* univariate approaches detected GxE effect for HDL, but none for LDL and triglycerides (lower part of Table 3). *2-step*-wht identified five SNPs on chromosome 8, and *2-step*-sst identified a subset of these SNPs (Table 3). Even though univariate *2-step*-wht found two additional SNPs on chromosome 8 compared to the *2-step* multivariate approaches, all the five SNPs are in strong LD resulting in the same lead SNP rs6984305, which was also identified by the *2-step* multivariate tests (Table 3). For the common SNPs detected by both univariate and multivariate *2-step*-wht, the univariate *2-step*-wht produced more significant *P*-values compared to multivariate *2-step*-wht (Table 3). Multivariate *2-step*-wht identified the GxE SNPs on chromosome 1 which were missed by the univariate approaches. Thus, the univariate approach identified fewer loci compared to the multivariate approach.

At rs7528419 on chromosome 1, the univariate GxE test (multiple linear regression) *P*-value for three lipids were 0.1, 0.008 and 0.0001, none of which is genome-wide significant. Even though this seems to be a moderate evidence of pleiotropy in GxE effect, the *1-step* multivariate test (MMLR) *P*-value for an overall GxE effect across lipids was 0.001 which is also not genome-wide significant. However, since this SNP has a strong evidence of pleiotropy in marginal main genetic effect with univariate *P*-values across lipids as 1.6×10^{-283}, 4.4×10^{-20}, 0.0001, and also a *P*-value $< 10^{-300}$ for the multivariate main genetic association, the multivariate *2-step*-wht approach prioritized this SNP in the second step while testing GxE, and identified a genome-wide significant overall GxE effect for this SNP.

At rs6984305 on chromosome 8 which is mapped to a nearby gene AC022784.1, a previous study (Noordam *et al.*, 2019) identified an effect of GxE interaction on HDL with sleep duration as the environmental factor. At rs7528419 on chromosome 1 (mapped to gene CELSR2), previous studies (De Vries *et al.*, 2019) found GxE interaction effect on LDL as well as HDL with alcohol consumption as the environmental factor in multiple different populations. Therefore, our method replicated these signals of GxE interaction for lipids; however, the environmental factor was different for the signal on chromosome 8. In NHGRI-EBI GWAS catalog, rs6984305 on chromosome 8 is also reported to be marginally associated with liver enzyme levels and serum alkaline phosphatase levels; and rs6984305 on chromosome 1 is reported to be associated with LDL (Zhu *et al.*, 2017), response to statin (Theusch *et al.*, 2014), etc.

5 Discussion

We have proposed a two-step approach to test for an aggregate-level gene–environment interaction across multiple related phenotypes. Using simulations, we demonstrate that our method produces substantially higher power than the Bonferroni-corrected one-step test

of overall effect of GxE interaction on a vector of multiple phenotypes. While our proposed two-step multivariate approach also provides substantially higher power than competing univariate approaches in the presence of pleiotropic GxE effect, in the absence of pleiotropy, the method only loses marginal power compared to the analogous two-step univariate approach. We demonstrate our *2-step* approach by applying it to a vector of three lipid phenotypes in the UK Biobank with the frequency of alcohol consumption as the environmental factor. Our method identified a pair of independent genome-wide significant signals of overall effect of GxE interaction on the three lipids. Previous studies reported these SNPs to have GxE effect on LDL & HDL with alcohol consumption as the environmental factor, and on HDL with sleep duration as the environmental factor. A limitation of the UKB lipid data which we analyzed is that the information on any medication to control or lower lipid levels of the participants were not available.

In a GxE study for a univariate phenotype, a common approach is to perform a joint 2 degree of freedom test to detect either a marginal genetic effect or a GxE effect (Kraft *et al.*, 2007; Manning *et al.*, 2011). It aims to improve the discovery of SNPs associated with the phenotype in the presence of heterogeneity in the main genetic effect due to the environmental factor. A joint test of either a marginal genetic effect or an interaction effect cannot distinguish a significant signal of GxE effect. In this paper, we solely focused on testing a non-null effect of the GxE interaction. Hence, for a meaningful comparison of our approach with the 2 degree of freedom test, we improvised an analogous testing procedure. We devised a Wald test based on the least square estimates of univariate β_{GE} (Manning *et al.*, 2011) across the three phenotypes and their corresponding covariance structure (outlined in Appendix). Using simulations we compared this approach with our 2-step multivariate approach, and found that the 2-step approach consistently produces higher power. Moreover, the improvised approach produces very similar power as the multivariate multiple linear regression (MMLR), because the former can be viewed as a summary statistics based version of the later.

Zhang *et al.* (2019) proposed a statistical approach to test for an overall GxE effect on multiple phenotypes at a gene-level. In contrast, our approach focuses on one SNP at a time. Hence, for a SNP-level comparison, we designed an analogous testing procedure following the main idea of Zhang *et al.* (2019). We consider the principal components (PCs) of the multivariate phenotype as the newly derived phenotypes which are uncorrelated between each other. For each PC, we perform the univariate GxE test using a multiple linear regression in presence of a possible main effect due to the SNP and environmental factor. Finally, we combine the GxE test *P*-values across the PCs using the Fisher's meta analysis to obtain the *P*-value of testing for an overall GxE effect. Using simulations, we find that our proposed multivariate 2-step approach consistently produces a higher power than this approach (Supplementary Figs S7–S9). Moreover, the approach based on PCs loses marginal power (around 1%) compared to the multivariate multiple linear regression (MMLR).

There are some limitations to our approach and potential for future improvement. First, a crucial assumption underlying the two-step approaches is that if a genetic variant has a GxE effect, it should also have a marginal genetic effect on the phenotype. This assumption is expected to hold in most cases (Paré *et al.*, 2010). For example, for a binary environmental factor (e.g. smoking status), a SNP with GxE effect has no marginal genetic association in the following scenario. Its marginal genetic effect on the phenotype of individuals belonging to each subgroup of the population classified by the status of the environmental factor (e.g. smokers versus non-smokers) not only has an opposite direction, but also perfectly cancels out each other across the subgroups to become zero in the population. In general, this is a strong condition to be satisfied for many SNPs. Thus, it is more plausible that a genetic variant with a GxE effect also has a marginal genetic association (at least with a weak effect) (Kooperberg and LeBlanc, 2008; Murcray *et al.*, 2008, 2011; Paré *et al.*, 2010). In support of this assumption, our real data analysis for lipids indeed identified SNPs which show an evidence of both GxE effect and marginal genetic effect. However, it is possible that a few SNPs can have a GxE effect in spite of having no marginal genetic effect. If a GxE SNP does not have any marginal genetic effect, one-step GxE testing procedures are expected to be more powerful than two-step approaches. Second, as commonly practiced in standard GWAS, the multiple testing procedures implemented in our approach to identify the genome-wide significant signals of GxE effect do not explicitly account for LD among the SNPs. Hence, the procedures are expected to be conservative in nature, limiting the power of the tests. An interesting future direction of research will be to adjust the multiple testing strategies for LD, in particular, the weighted hypothesis testing and subset testing procedures. Third, we have initially developed our approach and explored its performance for quantitative phenotypes. However, the phenotypes can also be of a mixed type, e.g. blood pressure (continuous) and stroke (binary). One strategy is to convert the binary trait into a liability scale using its estimate of prevalence in the population, and then apply our method on the pair of continuous phenotypes. To preserve the original mixed type of the phenotypes in the analysis, we plan to develop new methods using a generalized estimating equations (GEE) approach under the framework of seemingly unrelated regressions (SUR) (Liu *et al.*, 2009). In future work, we also plan to extend the approach for multiple related case-control phenotypes. Fourth, another future direction is to develop a GxE test for multiple phenotypes using the minimum of univariate GxE test *P*-values across the phenotypes as the test statistic. It is important to derive the null distribution of the test statistic accounting for any possible correlation between the *P*-values. Fifth, we have considered one environmental factor in the model. However, considering multiple relevant environmental factors at the same time as multiple phenotypes should further improve the power of detecting an overall GxE effect.

In summary, studying GxE interactions is crucial to better understand how the interplay among the genetic and environmental factors underlies the phenotypic variation. However, detecting GxE interactions can be challenging due to limited statistical power which can be substantially increased with new methodological approaches. Here we present a novel approach that improves the statistical power to identify GxE interactions in the presence of pleiotropy in the GxE effect, which may be valuable for future studies. The approach is theoretically sound and computationally efficient. We provide an R package MPGE for general use of the method by other investigators.

Acknowledgements

This research was conducted using the UK Biobank Resource under applications 24129 and 33297. The authors thank the participants of UK Biobank for making this work possible. They thank Yi Ding for providing useful information regarding UK Biobank data.
Financial Support: none declared.

Conflict of Interest: none declared.

References

Bhattacharjee,S. *et al.* (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.*, **90**, 821–835.

Bycroft,C. *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.

Cornelis,M.C. *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.

Dahl,A. *et al.* (2020) A robust method uncovers significant context-specific heritability in diverse complex traits. *Am. J. Hum. Genet.*, **106**, 71–91.

Dai,J.Y. *et al.* (2012a) Simultaneously testing for marginal genetic association and gene–environment interaction. *Am. J. Epidemiol.*, **176**, 164–173.

Dai,J.Y. *et al.* (2012b) Two-stage testing procedures with independent filtering for genome-wide gene–environment interaction. *Biometrika*, **99**, 929–944.

De Vries,P.S. *et al.*; InterAct Consortium. (2019) Multiancestry genome-wide association study of lipid levels incorporating gene-alcohol interactions. *Am. J. Epidemiol.*, **188**, 1033–1054.

Fox,J. and Weisberg,S. (2018) *An R Companion to Applied Regression*. Sage Publications, Thousand Oaks.

Galesloot,T.E. *et al.* (2014) A comparison of multivariate genome-wide association methods. *PLoS One*, **9**, e95923.

Gauderman,W.J. *et al.* (2013) Finding novel genes by testing G × E interactions in a genome-wide association study. *Genet. Epidemiol.*, **37**, 603–613.

Gauderman,W.J. *et al.* (2017) Update on the state of the science for analytical methods for gene–environment interactions. *Am. J. Epidemiol.*, **186**, 762–770.

Hsu,L. *et al.* (2012) Powerful cocktail methods for detecting genome-wide gene–environment interaction. *Genet. Epidemiol.*, **36**, 183–194.

Ionita-Laza,I. *et al.* (2007) Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *Am. J. Hum. Genet.*, **81**, 607–614.

Jiang,X. *et al.* (2019) Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.*, **10**, 431.

Kilpeläinen,T.O. *et al.*; Lifelines Cohort Study. (2019) Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. *Nat. Commun.*, **10**, 1–11.

Kooperberg,C. and LeBlanc,M. (2008) Increasing the power of identifying gene × gene interactions in genome-wide association studies. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.*, **32**, 255–263.

Kraft,P. *et al.* (2007) Exploiting gene–environment interaction to detect genetic associations. *Hum. Hered.*, **63**, 111–119.

Liu,J. *et al.* (2009) Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet. Epidemiol.*, **33**, 217–227.

Majumdar,A. *et al.* (2015) Semiparametric allelic tests for mapping multiple phenotypes: binomial regression and mahalanobis distance. *Genet. Epidemiol.*, **39**, 635–650.

Majumdar,A. *et al.* (2016) Determining which phenotypes underlie a pleiotropic signal. *Genet. Epidemiol.*, **40**, 366–381.

Majumdar,A. *et al.* (2018) An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS Genet.*, **14**, e1007139.

Manning,A.K. *et al.* (2011) Meta-analysis of gene–environment interaction: joint estimation of SNP and SNP × environment regression coefficients. *Genet. Epidemiol.*, **35**, 11–18.

Moore,R. *et al.*; BIOS Consortium. (2019) A linear mixed-model approach to study multivariate gene–environment interactions. *Nat. Genet.*, **51**, 180–186.

Mukherjee,B. and Chatterjee,N. (2008) 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.

Mukherjee,B. *et al.* (2012) Testing gene–environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.*, **175**, 177–190.

Murcray,C.E. *et al.* (2008) Gene–environment interaction in genome-wide association studies. *Am. J. Epidemiol.*, **169**, 219–226.

Murcray,C.E. *et al.* (2011) Sample size requirements to detect gene–environment interactions in genome-wide association studies. *Genet. Epidemiol.*, **35**, 201–210.

Noordam,R. *et al.* (2019) Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nat. Commun.*, **10**, 1–13.

Paré,G. *et al.* (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. *PLoS Genet.*, **6**, e1000981.

Ray,D. *et al.* (2016) USAT: a unified score-based association test for multiple phenotype–genotype analysis. *Genet. Epidemiol.*, **40**, 20–34.

Schaal,C. and Chellappan,S.P. (2014) Nicotine-mediated cell proliferation and tumor progression in smoking-related cancers. *Mol. Cancer Res.*, **12**, 14–23.

Theusch,E. *et al.* (2014) Ancestry and other genetic associations with plasma pcsk9 response to simvastatin. *Pharmacogenet. Genomics*, **24**, 492–500.

Turley,P. *et al.*; 23andMe Research Team. (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.*, **50**, 229–237.

Wasserman,L. and Roeder,K. (2006) Weighted hypothesis testing. *arXiv preprint math/0604172*.

Yu,Y. *et al.* (2018) Subset-based analysis using gene–environment interactions for discovery of genetic associations across multiple studies or phenotypes. *Hum. Hered.*, **83**, 283–314.

Zhang,J. *et al.* (2019) Test gene–environment interactions for multiple traits in sequencing association studies. *Hum. Hered.*, **84**, 170–196.

Zhang,P. *et al.* (2016) Detecting gene–environment interactions for a quantitative trait in a genome-wide association study. *Genet. Epidemiol.*, **40**, 394–403.

Zhu,Y. *et al.* (2017) Susceptibility loci for metabolic syndrome and metabolic components identified in Han Chinese: a multi-stage genome-wide association study. *J. Cell. Mol. Med.*, **21**, 1106–1116.