GxEsum: genotype-by-environment interaction model based on summary statistics

Jisu Shin[1,2], **S Hong Lee**[1,2]*

[1]Australian Centre for Precision Health, University of South Australia Cancer Research Institute, University of South Australia, Adelaide, SA 5000, Australia

[2]UniSA: Allied Health and Human Performance, University of South Australia, Adelaide, SA 5000, Australia

*Correspondence

Abstract

Genetic variation in response to the environment is fundamental in biology and has been described as genotype-by-environment interaction (GxE), reaction norm or phenotypic plasticity. In the genomic era, there has been increasing interest in estimating GxE, using genome-wide SNPs, e.g. a whole-genome reaction norm model (RNM) that can estimate unbiased genome-wide GxE. However, the existing approach is computationally demanding and infeasible to handle large-scale biobank data. Here we introduce GxEsum, a model for estimating GxE based on GWAS summary statistics, which can be applied to a large sample size. In simulations, we show that GxEsum can control type I error rate and produce unbiased estimates in general. We apply GxEsum to UK Biobank to estimate genome-wide GxE for BMI and hypertension, and find that the computational efficiency of GxEsum is thousands of times higher than existing whole-genome GxE methods such as RNM. Because of its computational efficiency, GxEsum can achieve a higher precision (i.e. power) from a larger sample size. As the scale of available resources has been increased, GxEsum may be an efficient tool to estimate GxE that can be applied to large-scale data across multiple complex traits and diseases.

Introduction

The success of the human genome project has led to a paradigm-shift in the complex trait analysis that focuses on the genome-wide association studies (GWAS) [1]. GWAS have been incredibly successful to identify genome-wide significant single nucleotide polymorphisms (SNPs) that are associated with causal variants underlying complex traits [2,3]. Moreover, whole-genome approaches, using all common SNPs across the genome, have been useful to dissect the genetic architecture of complex traits, e.g. SNP-based heritability and genetic correlation [4]. However, the analytical modelling used in GWAS and whole-genome approaches usually assumes that there is no genotype-environment interaction (GxE), which can be often violated against the true genetic architecture of complex traits. Indeed, interaction is fundamental in biology and there has been increasing interest in estimating GxE, using genome-wide SNPs [5-7].

Current state-of-the-art whole genome methods for estimating GxE include genotype-covariate interaction genomic restricted maximum likelihood (GREML) and random regression GREML [8]. Recently, a multivariate reaction norm model (RNM) has been introduced, which can disentangle GxE from genotype-environmental correlation, providing more reliable GxE estimations. These methods typically employ the GREML approach that requires individual level genotypes and is computationally intensive. Especially when using biobank-scale data, the approach becomes computationally intractable.

To reduce the computational limitation of GREML, linkage disequilibrium score regression (LDSC) was introduced to estimate SNP-based heritability and genetic correlation [9]. LDSC is computationally efficient and requires no individual-level genotypes. Instead, it uses GWAS summary statistics, regressing the association test statistics of SNPs on their LD score. However, existing LDSC methods are limited to additive models only [10-13].

In this study, we propose a novel approach for GxE based on GWAS summary statistics (GxESum), which can estimate GxE using a large-scale biobank dataset, correctly accounting for genotype-environment correlation and scale effects. In simulated and real data analyses, we show that the computational efficiency of the proposed approach is substantially higher than RNM, an existing GREML-based method while the estimates are reasonably accurate and precise. Because of this computational advantage, GxEsum may be an efficient tool to estimate GxE that can be applied to large-scale data across multiple complex traits.

Results

Method Overview

We propose a method to estimate whole-genome GxE based on GWAS summary statistics, referred to as GxEsum. GxEsum can be a computationally efficient RNM using an extension of LDSC approach. For SNP effects modulated by an environment, the expected chi-square statistic (χ_j^2) is

$$\mathrm{E}\left[\chi_j^2 | \ell_j\right] = \frac{N\sigma_{g_1}^2}{M} * \ell_j + 1 + 2(\sigma_{g_1}^2 + \sigma_{\tau_1}^2)$$

where N is the number of individuals, M is the number of SNPs, $\sigma_{g_1}^2$ is the variance due to GxE, $\sigma_{\tau_1}^2$ is the variance due to residual heterogeneity or scale effects caused by residual-environment interaction (RxE), ℓ_j is the LD score at the variant j (please see Methods for a full derivation of this equation). If chi-square statistics from GWAS are regressed on LD scores, non-genetic interaction effects ($\sigma_{\tau_1}^2$) are captured by the intercept, from which GxE ($\sigma_{g_1}^2$) can be disentangled. Consequently, GxE effects estimated by GxEsum are equivalent to that adjusted for RxE when using RNM (Ni et al. 2019).

To validate the proposed model, i.e. GxEsum, we used various simulations that were based on real genotype data. In simulations with and without GxE, we assessed the type I error rates and the accuracy of estimated GxE. In the simulations, we deliberately generated confounding effects such as genotype-environment (G-E) correlation, RxE and residual-environment (R-E) correlation to see if the type I error rate and the accuracy of GxEsum were affected by these confounding factors.

In the real data analysis, we used the UK Biobank data with 288,837 unrelated individuals after stringent quality control. Subsets of the data with various sample sizes were analysed to compare the precision (i.e. power) and the computational efficiency of GxEsum and GREML-based GxE model (i.e. RNM).

Finally, we show how the genetic effects of trait (e.g. BMI or hypertension) are modulated by environment (e.g. neuroticism score, smoking, alcohol consumption or physical activity or age) by using the proposed method.

Simulations

Under the null (no GxE), whether or not there were confounding effects (RxE and G-E and R-E correlations), the type I error rate of GxEsum was not significantly inflated (Supplementary Table 1). Note that the use of 500 replicates for each simulation scenario can detect a type I error of greater than 0.07 or less than 0.03 as significantly different from 0.05, using the binomial distribution theory [14,15]. Even with larger confounding effects (Supplementary Table 2), there was no inflation for the type I error rate of GxEsum.

In simulation with non-zero interactions, estimated GxE (g1) was unbiased and not significantly different from the true values whether there were significant G-E and R-

E correlations or not (see supplementary Figure 1). It was noted that RxE component was correctly captured by the intercept and not confounded with GxE estimates (Supplementary Note 1 and Supplementary Table 3). In the absence of RxE, estimated GxE was also unbiased (Supplementary Figure 2). The estimated GxE seemed robust to different values of G-E and R-E correlations ranging from 0.05 to 0.2, respectively (Supplementary Figures 3 and 4).

On the other hand, estimated main genetic variance (g0) was slightly biased especially when using a large G-E or R-E correlation (Supplementary Figure 4). This is probably because of the fact that the main genetic effects are over-adjusted for the environment due to the large correlations (between the trait and environment) in the model.

We also validated that there was no inflation for the type I error rate when applying GxEsum to binary (disease) traits (Supplementary Tables 4 and 5), showing that GxEsum appears to be robust to false positives in the scenarios of various confounders. In addition, we estimated the variance component of GxE on the observed scale and transformed to that on the liability scale, using Robertson transformation [16]. As shown in Supplementary Figure 5, the transformed estimates coincided the true simulated values on the liability scale although the precision of estimates (represented as 95% CI) was shown to be decreased when the population prevalence approached an extreme (e.g. k=0.025). GxE estimates were biased when simulating a large effect size of GxE (e.g. 10% of phenotypic variance explained by GxE) in the case of k=0.025 (Supplementary Figure 6) although they were mostly unbiased in the case of k=0.1 (Supplementary Figure 7). The level of biasedness appeared to be increased when there were RxE effects (Supplementary Figure 6). Note that the large GxE (> 5% of phenotypic variance) may not be realistic as estimated GxE explain only a small proportion of phenotypic variance in general (e.g. < 1%; see real data analysis below). Finally, caution should be given in interpreting GxE estimates when there are large confounding effects such as substantial G-E and R-E correlations (Supplementary Figure 8). The inflated GxE estimates were probably due to the fact that the phenotypes were over-adjusted for the environment in the model because of the correlation between the main trait and environment (G-E and R-E correlations). This resulted in a reduced phenotypic variance (Supplementary Figure 9), hence an inflated GxE estimates that is the ratio of the estimated GxE variance to the phenotypic variance.

Nevertheless, those confounders including RxE interaction and G-E/R-E correlations would not produce false positives whether using continuous quantitative or binary responses as shown above (Supplementary Tables 1 – 5). We additionally tested if the type I error rate of GxEsum was controlled when there is collider bias, which is concerned especially when using a self-report study (e.g. UK Biobank data)[17]. In simulation with collider bias, although estimated SNP-heritability was substantially (and unrealistically) underestimated (Supplementary Figures 10 and 11), the type I error rate of GxEsum was well controlled whether using continuous or binary responses (Supplementary Tables 6 and 7).

The estimated variance of the main genetic effects was mostly unbiased when using binary disease traits without G-E/R-E correlations (Supplementary Figure 12). When there were significant G-E and /or R-E correlations, the estimated variance of the main genetic effects appeared to be underestimated especially when there was RxE interaction (Supplementary Figure 13), which confirmed the fact that the main genetic effects are over-adjusted for the environment due to correlations between the trait and environment (Supplementary Figure 9).

Precision and Computational Efficiency

The precision (i.e. power) was assessed by comparing the standard error (SE) of GxEsum and RNM estimates. The SE of GxEsum was obtained from LDSC software (using a jackknife method). The SE of RNM for GxE component can be obtained from the information matrix [18] or from well-established theory [19] (see Supplementary Table 8). Figure 1 shows that the SE of GxEsum was 1.65 times higher than that of RNM when using the same sample size of 50,000. However, when sample size increased for GxEsum up to 288,837, for which RNM estimation is infeasible, the ratio reduced to 0.2. GxEsum can use a larger sample size (e.g. > 1,000,000), for which the ratio is expected to be further decreased, although the largest sample size tested in this study was 288,837 (Supplementary Figure 14).

While the precision of GxEsum is competitive with that of RNM, the computational efficiency is dramatically different between two methods (Figure 2 and supplementary Table 9). For example, when using a sample size of 50,000, the computing time for RNM was taken more than a thousand times than GxEsum. Even for GxEsum with a sample size of 288,837, its computational efficiency was still substantially higher than RNM with a sample size of 50,000 (Supplementary Figure 10 and Supplementary Table 9). This justifies that GxEsum is a computationally efficient tool that can be applied to biobank scale data for multiple complex traits and diseases. It is noted that we assumed that preliminary analyses for each method were already done (e.g. GRM for RNM, and LD scores and GWAS for GxEsum) (Supplementary Table 9).

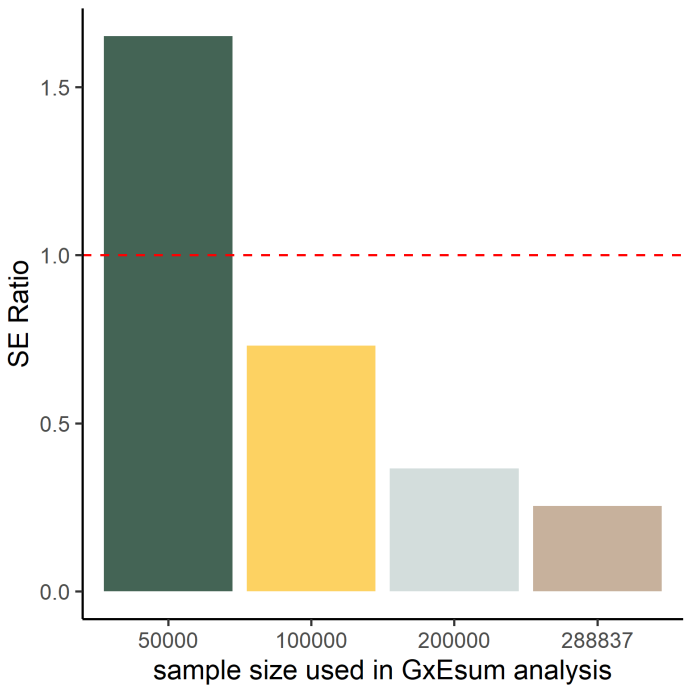

Figure 1. The ratio of standard error (SE) from GxEsum to that from RNM using UK Biobank data. The SEs of GxE variance estimated from GxEsum with various sample sizes ranging from 50,000 to 288,837 were obtained, and they were compared to the SE of GxE variance estimated from RNM with a sample size of 50,000. The dashed horizontal line represents the ratio as 1.

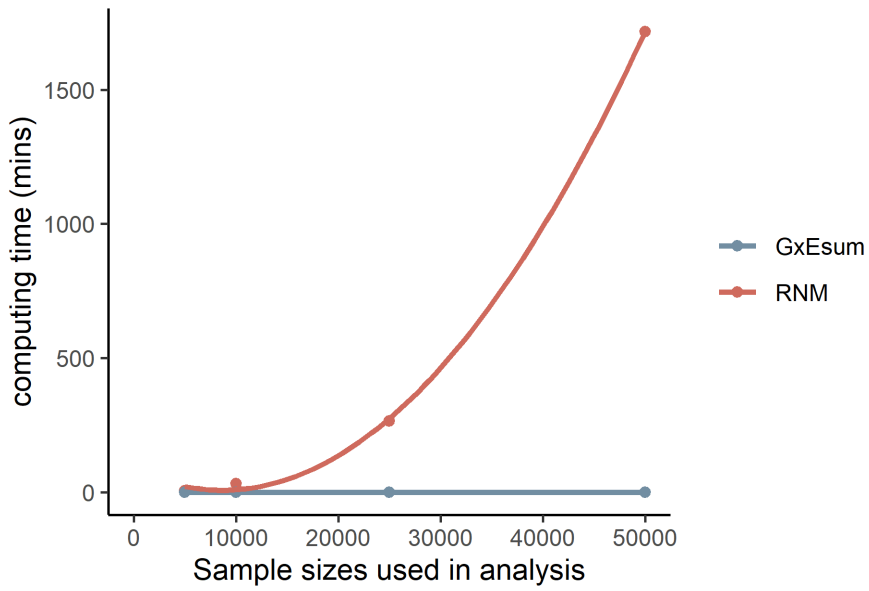

Figure 2. Computing time with various sample sizes used in GxEsum and RNM analyses. As the sample size increases, the computing time of RNM (red) increases exponentially, while that of GxEsum (blue) is almost invariant (less than a minute).

Real data analyses

We applied GxEsum to estimate genetic effects of body mass index (BMI) that were modulated by an environment such as age, alcohol consumption, smoking, neuroticism scores or physical activity. The significant GxE were observed from the analyses using neuroticism scores and smoking. On the other hand, we did not find significant GxE when using age, alcohol consumption and physical activity after Bonferroni correction (Bonferroni p-value = 0.05/8 = 0.006 since there were 8 significance tests in this study) (Table 1). The GxEsum approach applied to a binary disease was conducted using hypertension as the main trait, and BMI, waist-hip ratio (WHR) or body fat percentage (BFP) as an environmental variable. Table 1 shows that the genetic effects of hypertension are significantly modulated by BMI, but not by WHR or BFP.

Because not all variables were without missing, we imputed missing phenotypes using the mean value for each variable in the analyses, in order to maximise the sample size. In this real data analysis, there was no remarkable difference in the results whether using phenotypic imputation or not although some variables improved their significance, e.g. NEU and SMK (Supplementary Tables 10 and 11).

Table 1. Estimates obtained from GxEsum analysis using real data.

Main trait	Environmental variable	Main genetic effect	The proportion of phenotypic variance explained by GxE effect	p-value for GxE
BMI	Age	0.216 (0.007)	0.004 (0.0017)	1.86E-02
	NEU[a]	0.216 (0.007)	0.0069 (0.0016)	1.61E-05
	SMK[b]	0.212 (0.007)	0.0059 (0.002)	3.18E-03
	PA[c]	0.218 (0.007)	0.0029 (0.0013)	2.57E-02
	ALC[d]	0.216 (0.007)	0.0032 (0.0017)	5.98E-02
Hypertension	BMI	0.152 (0.008)	0.006 (0.002)	2.09E-03
	WHR[e]	0.154 (0.008)	0.005 (0.002)	3.21E-02
	BFP[f]	0.151 (0.008)	0.008 (0.003)	2.66E-02

We used a quantitative trait (BMI) and a binary disease trait (hypertension) because BMI is known to be modulated by age/lifestyle such as NEU, SMK, ALC, PA [8,20,21], and hypertension is known to be caused by obese traits such as BMI and WHR[22,23]. The p-value is from a Wald test for the estimated GxE variance not being different from zero. The estimates on the observed scale for the binary trait, hypertension, were transformed to those on the liability scale using Robertson transformation[16,24].

[a]NEU: Neuroticism Score
[b]SMK: Smoking
[c]PA: Physical Activity
[d]ALC: Alcohol consumption
[e]WHR: Waist-Hip Ratio
[f]BFP: Body Fat Percentage

Discussion

In this study, we propose GxEsum, a novel whole-genome GxE method, of which the computational efficiency is a thousand times higher than existing methods. The estimation of GxE using GWAS summary statistics has great flexibility in the application of the method to multiple complex traits and diseases. The proposed method and theory have been explicitly verified using comprehensive simulations that were carried out for both quantitative trait and binary disease. Moreover, we showed that the type I error rate of the proposed method was not inflated by moderate to severe collider bias [17] that caused a substantial underestimation of heritability shown in our simulation (Supplementary Figures 10 and 11).

In the real data analysis, we show that the genetic effects of BMI were significantly modulated by NEU and SMK, which agrees with previous studies [8,25]. It is noted that the significance of GxE was improved because we used a larger sample size, compared with the previous studies. Our result agrees with Robinson et al. (2017) who found no significant GxE evidence for age when analysing BMI using the UK Biobank in which the participants aged 40-69 at the recruitment. However, a dataset with a wider range of ages is desirable, which would increase the power to detect GxE on age. For example, a significant GxE was found in a BMI-age analysis using a dataset including samples aged 18-80 at the recruitment[8]. For hypertension, its causal relationship with BMI has been reported by a number of studies using Mendelian randomization [22,26]. However, it was not clear if the causal relationship is due to GxE or something else, e.g. unknown non-genetic effects of the disease modulated by BMI status. We show that the causal relationship between hypertension and BMI is significantly explained by genome-wide GxE. Interestingly, there is no significant evidence of genome-wide GxE for hypertension-WHR or hypertension-BFP causal relationship.

The estimated intercept from GxEsum should be interpreted with caution. We show that estimated intercepts were unbiased from the theoretically predicted values when using the simulation of quantitative traits, as a proof of concept, i.e. the phenotypic variance explained by RxE effects ($h^2_{\tau_1}$) can be obtained as $h^2_{\tau_1}$= (*intercept* – 1 – $2h^2_{g_1}$) / 2 (from eq. (4)). However, in real data analyses, there may be additional confounding effects such as scale effects, residual heteroscedasticity or/and sample heterogeneity that are often attributed to unknown factors. Moreover, when using binary traits, substantial scale effects can be generated (statistical RxE effects) because only affected and unaffected status are observed and individual differences within affected or unaffected group are ignored. These additional confounding effects and statistical scale effects are captured and estimated as an intercept in GxEsum[9], resulting in unreliable RxE estimates. It is noted that RxE estimation is not the main interest of GxEsum and can be more reliably estimated in RNM that is designed to model both GxE and RxE.

The existing GxE methods require individual-level genotype data which often has a restriction to share, and their computational burden is typically high. Moreover, it is not clear how they perform when the representativeness of the samples is limited, e.g. selection bias due to a collider in the UK Biobank samples. On the contrary, the

proposed approach, GxEsum, is computationally efficient and can detect GxE interaction correctly for both quantitative and binary disease traits even when there is moderate to server collider bias. If GWAS summary statistics of estimated main additive and interaction effects can be made publicly available, a meta-analysis across multiple cohorts can be possible for an ever-large GxE study (like the context of LDSC SNP-heritability meta-analysis). There are some issues that the measure of environmental variable may not be standardised across study cohorts, and the environmental variable may be even unavailable in some cohorts. However, these issues can be remedied when the information of exposome that is the standardised measure of all exposures for individuals, complemented to the genome, is available.

There is a GxE method that can use GWAS summary statistics, i.e. VarExp, which is recently published. While VarExp benefits computationally from using GWAS summary statistics, it needs to invert the correlation matrix between SNPs, which prevents from using a large number of SNPs[27]. Furthermore, the theoretical frameworks of GxEsum and VarExp are fundamentally different in that the latter does not account for confounding effects such as scale effects, residual heterogeneity or RxE that can be captured by the estimated intercept of GxEsum. Finally, the performance of VarExp has been verified with a limited magnitude of interaction effects up to 1.5% and 0.25% of the phenotypic variance for quantitative and binary traits, respectively[27].

Like RNM, GxEsum can fit environmental exposures such that the genetic effects of a trait can be modelled as a nonlinear function of a continuous environmental gradient. The potential modifier of the genetic effects is not limited to environmental exposures but can be extended to novel variables from multi-omics data such as gene-expression, protein expression and methylation data. Polygenic risk scores can also be considered as an environmental variable in the model. This novel approach may allow dissecting a latent biological architecture of a complex trait in a future application of GxEsum.

In the analysis of binary disease traits, estimates on the liability scale, transformed from those on the observed scale using Robertson transformation, should be interpreted with caution. Biased estimates on the liability scale are likely due to the violation of the normality assumption that is essentially required for the Robertson transformation, i.e. large interaction effects can cause a non-normal phenotypic distribution. It is also known that if the transformation involves substantial non-additive effects, it can give biased estimates on the liability scale [16,24,28]. However, when non-additive effects are small, the transformation can give reasonably accurate estimates on the liability scale, which is also evidenced by our simulations with small interaction effects. As shown in the real data analysis, the magnitude of genome-wide GxE is not large (< 1% of the phenotypic variance), showing that the bias of transformation due to the assumption violation may not be substantial in general. Nevertheless, it is required to develop a better transformation method for large interaction effects in a further study, e.g. when using multiple environmental variables simultaneously, the interaction effects are aggregated and can be substantially large.

There are a number of limitations in our study. First, like RNM, GxEsum does not determine the causal direction between variables, which can be provided from previous studies or other epidemiologic methods, e.g. Mendelian randomisation, as prior information. Second, we only modelled the first order of random regression coefficients with a single environmental variable, and there may be significant additional effects when modelling a higher-order interaction or multiple environmental variables. Third, the estimation for the main genetic effects can be biased when there are large G-E and/or R-E correlations. Because of such correlations, the main genetic effects are over-adjusted when the phenotypes of the main trait are adjusted for the environmental variable in the model. Therefore, a careful interpretation of the estimated main genetic effects is required when using GxEsum. Lastly, when using the same sample size, the precision of GxEsum is not better than GREML-based GxE methods, implying that the former is only useful when using a large sample size that the latter cannot handle.

In conclusion, despite these caveats, GxEsum can be a useful tool to estimate whole-genome GxE as it can achieve a higher precision (i.e. power) from a larger sample size, compared to existing GxE methods. Especially when the scale of available resources increases, GxEsum may be a unique method that can be applied to large-scale data across multiple complex traits and diseases in the context of GxE.

Methods

GxEsum

Following Ni et al. (2019), RNM can be written as

$$\mathbf{y} = \mathbf{b} + \mathbf{g} + \boldsymbol{\tau} = \mathbf{b} + \mathbf{g_0} + \mathbf{g_1} \times \mathbf{E} + \boldsymbol{\tau}_\mathbf{0} + \boldsymbol{\tau}_\mathbf{1} \times \mathbf{E}$$

where $\mathbf{y}$ is *N* vector of phenotypic observations, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{g}$ is *N* individual genetic effects, which can be decomposed into the first and second order of genetic random regression coefficients, $\mathbf{g_0}$ and $\mathbf{g_1}$, $\boldsymbol{\tau}$ is residual effects, decomposed into the first and second order of residual random regression coefficients, $\boldsymbol{\tau}_\mathbf{0}$ and $\boldsymbol{\tau}_\mathbf{1}$, and $\mathbf{E}$ is an *N* vector of environmental variable. Note that $\mathbf{E}$ can be also any covariate variable (e.g. smoking, alcohol consumption).

Assuming that the phenotypes ($\mathbf{y}$) are pre-adjusted for the main genetic effects ($\mathbf{g_0}$), environmental or covariate variable ($\mathbf{E}$) and other fixed effects ($\mathbf{b}$), the model can be rewritten as

$$\mathbf{y} = \mathbf{g_1} \times \mathbf{E} + \boldsymbol{\tau}_\mathbf{0} + \boldsymbol{\tau}_\mathbf{1} \times \mathbf{E} = \mathbf{X}\boldsymbol{\beta}_\mathbf{1} \times \mathbf{E} + \boldsymbol{\tau}_\mathbf{0} + \boldsymbol{\tau}_\mathbf{1} \times \mathbf{E}$$

where $\mathbf{X}$ is an *N* x *M* standardised genotype matrix for *M* SNPs, $\boldsymbol{\beta_1}$ is an *M* vector of SNP interaction effects modulated by the environment (i.e. GxE SNP effects). It is noted that $\boldsymbol{\tau_0}$ is residual effects that are consistent across environment whereas $\boldsymbol{\tau_1}$ captures heterogeneous residual effects across environment (i.e. RxE).

Following Bulik-Sullivan et al. (2015), assuming $\mathbb{E}[\mathbf{g_1}] = \mathbb{E}[\boldsymbol{\beta_1}] = 0$, the expected chi-square statistics of variant *j* for the GxE is

$$\mathbb{E}\left[\chi_j^2\right] = N \cdot \mathrm{Var}(\widehat{\beta_{1j}}) \qquad \text{eq. (1)}$$

Using the law of total variance, $\mathrm{Var}[\widehat{\beta_{1j}}]$ can be obtained as

$$\begin{aligned}\mathrm{Var}(\widehat{\beta_{1j}}) &= \mathbb{E}\left[\mathrm{Var}\left(\widehat{\beta_{1j}}\middle|\mathbf{EX}\right)\right] + \mathrm{Var}\left[\mathbb{E}(\widehat{\beta_{1j}}|\mathbf{EX})\right] \\ &= \mathbb{E}\left[\mathrm{Var}\left(\widehat{\beta_{1j}}\middle|\mathbf{EX}\right)\right]\end{aligned}$$

where $\mathbf{EX}$ is an *N* x *M* matrix with each column having the Hadamard product between $\mathbf{E}$ and $\mathbf{X}_j$(standardised genotypes at the *j* th SNP) and the conditional expectation of $\widehat{\beta_{1j}}$ is $\mathbb{E}\left(\widehat{\beta_{1j}}|\mathbf{EX}\right) = 0$.

Noting that the least-square estimate of $\widehat{\beta_{1j}}$ can be obtained as $\widehat{\beta_{1j}} = (\mathbf{EX}_j)'\mathbf{y}/N$, $\mathrm{Var}\left(\widehat{\beta_{1j}}|\mathbf{EX}\right)$ can be rearranged as

$$\begin{aligned}
&\mathrm{Var}\left(\widehat{\beta_{1j}}|\mathbf{EX}\right) = \mathrm{Var}[(\mathbf{EX}_j)'\mathbf{y}/N|\mathbf{EX}] \\
&= \frac{1}{N^2}\mathrm{Var}\left[\left(\mathbf{EX}_j\right)'\mathbf{y}\middle|\mathbf{EX}\right] \\
&= \frac{1}{N^2}\left(\mathbf{EX}_j\right)'\mathrm{Var}(\mathbf{y}|\mathbf{EX})\left(\mathbf{EX}_j\right) \\
&= \frac{1}{N^2}(\mathbf{EX}_j)'\mathrm{Var}[(\mathbf{EX})\boldsymbol{\beta_1} + \boldsymbol{\tau_0} + \mathbf{E}\boldsymbol{\tau_1}|\mathbf{EX}](\mathbf{EX}_j) \\
&= \frac{1}{N^2}\left(\mathbf{EX}_j\right)'\mathrm{Var}[(\mathbf{EX})\boldsymbol{\beta_1}|\mathbf{EX}]\left(\mathbf{EX}_j\right) \\
&+ \frac{1}{N^2}\left(\mathbf{EX}_j\right)'\left(\mathbf{EX}_j\right)\mathrm{Var}(\boldsymbol{\tau_0}) + \frac{1}{N^2}(\mathbf{EX}_j)'(\mathbf{E})(\mathbf{E})'(\mathbf{EX}_j)\mathrm{Var}(\boldsymbol{\tau_1}) \\
&= \frac{1}{N^2}(\mathbf{EX}_j)'(\mathbf{EX})(\mathbf{EX})'(\mathbf{EX}_j)\mathrm{Var}(\boldsymbol{\beta_1}|\mathbf{EX}) \\
&+ \frac{1}{N^2}\left[N\mathrm{Var}(\boldsymbol{\tau_0}) + (\mathbf{EX}_j)'(\mathbf{E})(\mathbf{E})'(\mathbf{EX}_j)\mathrm{Var}(\boldsymbol{\tau_1})\right] \\
&= \frac{1}{N^2}\frac{h_{g_1}^2}{M}\left(\mathbf{EX}_j\right)'(\mathbf{EX})(\mathbf{EX})'\left(\mathbf{EX}_j\right) \\
&+ \frac{1}{N^2}\left[N\left(1 - h_{g_1}^2 - h_{\tau_1}^2\right) + h_{\tau_1}^2(\mathbf{EX}_j)'(\mathbf{E})(\mathbf{E})'(\mathbf{EX}_j)\right]
\end{aligned}$$

where $h^2_{g_1}$ and $h^2_{\tau_1}$ is the proportion of phenotypic variance explained by GxE and RxE, respectively.

Therefore, $\mathbb{E}[\chi_j^2]$ in eq. (1) can be written as

$$\mathbb{E}[\chi_j^2] = N \cdot \mathrm{Var}(\widehat{\beta_{1j}}|\mathbf{EX})$$

$$= \frac{1}{N}\left[\frac{h^2_{g_1}}{M}(\mathbf{EX}_j)'(\mathbf{EX})(\mathbf{EX})'(\mathbf{EX}_j) + N(1 - h^2_{g_1} - h^2_{\tau_1}) + h^2_{\tau_1}(\mathbf{EX}_j)'(\mathbf{E})(\mathbf{E})'(\mathbf{EX}_j)\right] \quad \text{eq. (2)}$$

According to Bulik-Sullivan et al. (2015), the products of the standardised genotypes at variant *j* and other variants can be expressed as a function of LD scores, i.e.

$$\frac{1}{N^2}(\mathbf{X_j})'(\mathbf{X})(\mathbf{X})'(\mathbf{X_j}) = \frac{1}{N^2}\left(\mathrm{N}^2 + \left[\mathrm{N} * (1 - \tilde{r}_j^2) + \tilde{r}_j^2 * \mathrm{N}^2\right] * (\mathrm{M} - 1)\right)$$

$$= \mathbf{1} + \left[\frac{1 - \tilde{r}_j^2}{\mathrm{N}} + \tilde{r}_j^2\right] * (\mathrm{M} - 1)$$

$$= \ell_j + \left[\frac{(\mathrm{M} - 1)(1 - \tilde{r}_j^2)}{\mathrm{N}}\right]$$

$$\approx \ell_j + \frac{M - \ell_j}{N}$$

where $\tilde{r}_j^2$ defined as the expected sample correlation between genotypes at the *j*th variant and the other (M-1) variants, and $\ell_j = 1 + \tilde{r}_j^2(M - 1)$ is the LD scores of the *j*th SNP.

According to the central moment theory of standard normal distribution of three independent random variables ($\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{E}$), each with an N vector, useful equations are

$$\mathbb{E}[(\mathbf{X_1})'(\mathbf{X_1})] = N$$

$$\mathbb{E}[(\mathbf{X_1})'(\mathbf{X_2})(\mathbf{X_2})'(\mathbf{X_1})] = N$$

$$\mathbb{E}[(\mathbf{X_1})'(\mathbf{X_1})(\mathbf{X_1})'(\mathbf{X_1})] = N^2$$

$$\mathbb{E}[(\mathbf{EX_1})'(\mathbf{E})(\mathbf{E})'(\mathbf{EX_1})] = \mathbb{E}[(\mathbf{EX_1})'(\mathbf{EX_2})(\mathbf{EX_2})'(\mathbf{EX_1})] = 3N$$

and

$$\mathbb{E}[(\mathbf{EX_1})'(\mathbf{EX_1})(\mathbf{EX_1})'(\mathbf{EX_1})] = N^2.$$

Therefore, assuming that $\mathbf{E}$ and $\mathbf{X_j}$ have negligible correlation (as expected), the term $(\mathbf{EX}_j)'(\mathbf{EX})(\mathbf{EX})'(\mathbf{EX}_j)$ can be expressed as a function of LD scores as

$$\frac{1}{N^2}(\mathbf{EX}_j)'(\mathbf{EX})(\mathbf{EX})'(\mathbf{EX}_j) = \frac{1}{N^2}\left(N^2 + \left[3N(1 - \tilde{r}_j^2) + \tilde{r}_j^2 * N^2\right] * (M - 1)\right)$$

$$= 1 + \left[\frac{3\left(1-\tilde{r}_j^2\right)}{N} + \tilde{r}_j^2\right] * (M-1)$$

$$= \ell_j + \frac{3(M-\ell_j)}{N}$$

Thus, a part in eq. (2) can be rearranged as

$$\frac{1}{N}\left[\frac{h_{g_1}^2}{M}\left(\mathbf{EX}_j\right)'(\mathbf{EX})(\mathbf{EX})'\left(\mathbf{EX}_j\right) + N\left(1-h_{g_1}^2\right)\right]$$

$$= \frac{1}{N}\left[\frac{N^2 h_{g_1}^2}{M}\left(\ell_j + \frac{3\left(M-\ell_j\right)}{N}\right) + N\left(1-h_{g_1}^2\right)\right]$$

$$= \frac{N h_{g_1}^2}{M}\left(\ell_j + \frac{3\left(M-\ell_j\right)}{N}\right) + 1 - h_{g_1}^2$$

$$= \frac{(N-3)h_{g_1}^2}{M}\ell_j + 1 + 2h_{g_1}^2$$

$$= \frac{N(1-3/N)h_{g_1}^2}{M}\ell_j + 1 + 2h_{g_1}^2$$

$$= \frac{N h_{g_1}^2}{M}\ell_j + 1 + 2h_{g_1}^2 \qquad \text{eq. (3)}$$

The term, $1-3/N$, in eq. (3) can be approximated as 1 in the analysis using biobank scale data, which contains over 10^5 samples.

The remaining part in eq. (2) can be rearranged as

$$\frac{1}{N}\left[N\left(-h_{\tau_1}^2\right) + h_{\tau_1}^2(\mathbf{EX}_j)'(\mathbf{E})(\mathbf{E})'(\mathbf{EX}_j)\right] = 2h_{\tau_1}^2$$

where $\mathbb{E}[(\mathbf{EX_1})'(\mathbf{E})(\mathbf{E})'(\mathbf{EX_1})] = 3N$ according to the central moment theory of standard normal distribution (see above), assuming that $\mathbf{E}$ and each column of $\mathbf{X}$ have a negligible correlation, which satisfies if $\mathbf{E}$ is an environmental variable or a polygenic trait.

Therefore,

$$\mathbb{E}\left[\chi_j^2\right] = N \cdot \mathrm{Var}\left(\widehat{\beta_{1j}}|\mathbf{EX}\right) = \frac{N h_{g_1}^2}{M}\ell_j + 1 + 2h_{g_1}^2 + 2h_{\tau_1}^2 \qquad \text{eq. (4)}$$

where $1 + 2h_{g_1}^2 + 2h_{\tau_1}^2$ can be obtained as the intercept of the outcome by fitting to the proposed model (GxEsum), which is verified using simulations (see Supplementary Note 1 and Supplementary Table 11).

To validate the proposed model in general, we used comprehensive phenotypic simulations that were based on real genotype data (see Supplementary Note 2).

Real data

UK Biobank data were used, which contains 0.5 million individuals aged between 40-69 years. The data consists of health-related information for each participant who was recruited in 2006-2010, and their imputed genomic data (~92 million SNPs) has been distributed through European Genome-phenome Archive. A stringent quality control process for individuals was set as followings: 1) who were reported as non-white British, 2) who were having mismatched gender between the reported and the inferred by the genotypic data, 3) who were having missing rate over 0.05, 4) who were having putative sex chromosome aneuploidy. In addition, only HapMap3 SNPs were used which were passed from the stringent quality controls for SNPs. The filter for SNPs is set as followings: 1) which were having INFO score less than 0.6, 2) which were having a MAF less than 1%, 3) which were having Hardy-Weinberg Equilibrium (HWE) P-value less than 1E-4, 4) one of which from the duplicated SNPs. From those passing the tough procedures, we additionally excluded one of pair of samples who were having the genomic relationship higher than 0.05. After quality control, 288,837 individuals and 1,133,273 SNPs were remained.

Among trait phenotypes available in the UK biobank, we arbitrarily selected BMI (a quantitative trait) and hypertension (a binary disease trait) and tested if the genetic effects of BMI or hypertension was significantly modulated by an environmental variables, i.e. NEU, smoking, ALC, PA or age for testing the quantitative trait, and BMI, WHR or BFP for testing the binary disease trait. The phenotypes of the main trait were adjusted for potential confounders such as age, gender, year of birth, assessment centre, genetic batch, household income, educational qualification, the first 10 principal components, and the environmental variable. For any phenotypic missing value for each variable, we used the mean of the phenotypes of the variable, i.e. phenotypic imputation with the mean. Further details of the variables used on this study are in Supplementary Note 2.

In GWAS, we used a linear model for quantitative traits as well as for binary responses. The use of a linear model applied to binary responses is because it has been reported that a logistic regression may generate biased estimates in some instances[29] and our simulations (Supplementary Note 3) were based on a probit model (i.e. a linear transformation of the inverse standard normal distribution) that can be well approximated by a linear model [30].

URLs

The UK Biobank via https://www.ukbiobank.ac.uk/

The PLINK version 1.9 via https://www.cog-genomics.org/plink/1.9/

The MTG 2 version 2.15 via https://sites.google.com/site/honglee0707/mtg2

The LDSC via https://github.com/bulik/ldsc

The ARIC study via https://sites.cscc.unc.edu/aric/

GxEsum script via https://sites.google.com/site/honglee0707/mtg2

Code Availability

MTG2 version 2.15 for using RNM https://sites.google.com/site/honglee0707/mtg2

GxEsum model is implemented in the script that are publicly available in https://sites.google.com/site/honglee0707/mtg2

Acknowledgements

This research is supported by Australian Research Council (DP 190100766, FT 160100229), and the HPC resources were provided by the Australian Government through Gadi under the National Computational Merit Allocation Scheme (NCMAS) and by the university of South Australia through Tango 2.0. We would like to thank staff and participants of the ARIC study and the UK Biobank for their valuable contributions. The access code for the ARIC study is phs000090, and the reference number approved by the UK Biobank is 14575. UK Biobank data was approved by the North West Research Ethics Committee (REC, reference number 11/NW/0382).

References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24 (2012).
2. Sud, A., Kinnersley, B. & Houlston, R.S. Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer* **17**, 692-704 (2017).
3. Pasaniuc, B. & Price, A.L. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18**, 117-127 (2017).
4. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
5. Arnau-Soler, A. *et al.* Genome-wide by environment interaction studies of depressive symptoms and psychosocial stress in UK Biobank and Generation Scotland. *Translational Psychiatry* **9**, 14 (2019).
6. Gong, J. *et al.* Genome-Wide Interaction Analyses between Genetic Variants and Alcohol Consumption and Smoking for Risk of Colorectal Cancer. *PLoS Genet* **12**, e1006296 (2016).
7. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics* **44**, 659-669 (2012).
8. Robinson, M.R. *et al.* Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat Genet* **49**, 1174-1181 (2017).
9. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
10. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
11. Gazal, S., Marquez-Luna, C., Finucane, H.K. & Price, A.L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat Genet* **51**, 1202-1204 (2019).

12. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat Genet* **51**, 1244-1251 (2019).
13. Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics, C., Wray, N.R. & Lee, S.H. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *Am J Hum Genet* **102**, 1185-1194 (2018).
14. Rosner, B. *Fundamentals of biostatistics*, (Nelson Education, 2015).
15. Austin, P.C. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *Int J Biostat* **5**, Article 13 (2009).
16. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics* **88**, 294-305 (2011).
17. Munafò, M.R., Tilling, K., Taylor, A.E., Evans, D.M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology* **47**, 226-235 (2017).
18. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*, (Sinauer Sunderland, MA, 1998).
19. Visscher, P.M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* **10**, e1004269 (2014).
20. Sutin, A.R., Ferrucci, L., Zonderman, A.B. & Terracciano, A. Personality and obesity across the adult life span. *Journal of personality and social psychology* **101**, 579-592 (2011).
21. Rask-Andersen, M., Karlsson, T., Ek, W.E. & Johansson, Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS genetics* **13**, e1006977-e1006977 (2017).
22. Hyppönen, E., Mulugeta, A., Zhou, A. & Santhanakrishnan, V.K. A data-driven approach for studying the role of body mass in multiple diseases: a phenome-wide registry-based case-control study in the UK Biobank. *The Lancet Digital Health* **1**, e116-e126 (2019).
23. Lee, M.-R., Lim, Y.-H. & Hong, Y.-C. Causal association of body mass index with hypertension using a Mendelian randomization design. *Medicine* **97**, e11252-e11252 (2018).
24. Dempster, E.R. & Lerner, I.M. Heritability of Threshold Characters. *Genetics* **35**, 212-236 (1950).
25. Ni, G. *et al.* Genotype-covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nat Commun* **10**, 2239 (2019).
26. Larsson, S.C., Bäck, M., Rees, J.M.B., Mason, A.M. & Burgess, S. Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study. *European heart journal* **41**, 221-226 (2020).
27. Laville, V. *et al.* VarExp: estimating variance explained by genome-wide GxE summary statistics. *Bioinformatics* **34**, 3412-3414 (2018).
28. Van Vleck, L.D. Estimation of Heritability of Threshold Characters. *Journal of Dairy Science* **55**, 218-225 (1972).
29. Sun, R., Carroll, R.J., Christiani, D.C. & Lin, X. Testing for gene-environment interaction under exposure misspecification. *Biometrics* **74**, 653-662 (2018).
30. Lee, S.H., Goddard, M.E., Wray, N.R. & Visscher, P.M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology* **36**, 214-224 (2012).