

# An approach to identify gene-environment interactions and reveal new biological insight in complex traits

Received: 22 September 2023

Accepted: 10 April 2024

Published online: 22 April 2024

 Check for updates

Xiaofeng Zhu<sup>1</sup>✉, Yihe Yang<sup>1</sup>, Noah Lorincz-Comi<sup>1</sup>, Gen Li<sup>1</sup>, Amy R. Bentley<sup>2</sup>, Paul S. de Vries<sup>3</sup>, Michael Brown<sup>3</sup>, Alanna C. Morrison<sup>3</sup>, Charles N. Rotimi<sup>2</sup>, W. James Gauderman<sup>4</sup>, Dabeeru C. Rao<sup>5</sup>, Hugues Aschard<sup>6,7</sup> & the CHARGE Gene-lifestyle Interactions Working Group\*

There is a long-standing debate about the magnitude of the contribution of gene-environment interactions to phenotypic variations of complex traits owing to the low statistical power and few reported interactions to date. To address this issue, the Gene-Lifestyle Interactions Working Group within the Cohorts for Heart and Aging Research in Genetic Epidemiology Consortium has been spearheading efforts to investigate  $G \times E$  in large and diverse samples through meta-analysis. Here, we present a powerful new approach to screen for interactions across the genome, an approach that shares substantial similarity to the Mendelian randomization framework. We identify and confirm 5 loci (6 independent signals) interacted with either cigarette smoking or alcohol consumption for serum lipids, and empirically demonstrate that interaction and mediation are the major contributors to genetic effect size heterogeneity across populations. The estimated lower bound of the interaction and environmentally mediated heritability is significant ( $P < 0.02$ ) for low-density lipoprotein cholesterol and triglycerides in Cross-Population data. Our study improves the understanding of the genetic architecture and environmental contributions to complex traits.

Current genome-wide association studies (GWAS) focus on detecting genetic variants that lead to different phenotype means across genotype groups<sup>1,2</sup>, and have identified a large number of genetic loci that, in some cases, explain large proportions of the trait's SNP-heritability<sup>3-5</sup>. While it is commonly agreed that complex traits are influenced by genetic and environmental factors and their interactions<sup>6-9</sup>, there is a long-standing disagreement about the

magnitude of the  $G \times E$  contribution to heritability because of different theoretical models and assumptions<sup>10,11</sup>. As pointed out in ref. 12, arbitrarily defined parameterizations of genetic effects with non-additive gene actions may explain the same degree of genetic variation as the currently prevailing additive model. Thus, while using additive genetic models such as polygenic risk scores to predict individual quantitative or qualitative phenotypes has become standard<sup>5</sup>, these

<sup>1</sup>Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, USA. <sup>2</sup>Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. <sup>3</sup>Human Genetics Center, Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. <sup>4</sup>Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA. <sup>5</sup>Center for Biostatistics and Data Science, Institute for Informatics, Data Science and Biostatistics, Washington University School of Medicine, St. Louis, MO, USA. <sup>6</sup>Institut Pasteur, Université Paris Cité, Department of Computational Biology, F-75015 Paris, France. <sup>7</sup>Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. \*A list of authors and their affiliations appears at the end of the paper. ✉ e-mail: [xxz10@case.edu](mailto:xxz10@case.edu)

models may not be fully informative in understanding genetic architecture.

Interactions are often studied secondarily in comparison to additive variance, whose advantage is to explain most of the correlations among relatives and fit natural selection model well<sup>10,13</sup>. Theoretical studies have demonstrated that a significant portion of variance can be explained by an additive model even when the genetic contribution to a phenotype is purely through  $G \times E$ <sup>14</sup>. This limitation is one of the key factors explaining the low power of approaches modeling interactions conditional on additive variance. As a result, studies focusing on detecting  $G \times E$  at the genome-wide level are seldom considered as primary analyses. By contrast, the joint evidence of main genetic and  $G \times E$  effects, in addition to the  $G \times E$  alone, is tested in the Gene-Lifestyle Interactions (GLI) Working Group within the Cohorts for Heart and Aging Research in Genetic Epidemiology Consortium (CHARGE)<sup>9,15</sup>, where only a modest number of genetic loci have been identified through testing for  $G \times E$  alone<sup>16–19</sup>. The joint test limits our ability to determine to what degree the currently identified loci reflect evidence for  $G \times E$  contribution, making it difficult to understand the precise interplay between genetic and environmental factors.

Concurrently, Mendelian randomization (MR) has been developed and widely applied to study causal relationships between risk exposures and outcomes in the post-GWAS era<sup>20,21</sup>. Although an MR approach has been used to explain  $G \times E$ <sup>22</sup>, the underlying connection between testing pleiotropic variants in the MR framework and the detection of  $G \times E$  is currently unclear. Here, we conceptually connect  $G \times E$  with MR framework, illuminate their similarities and demonstrate that the test of horizontal pleiotropy in MR<sup>23,24</sup> can be used for detecting  $G \times E$ . Based on this principle, one can identify  $G \times E$  using existing available GWAS and GWIS summary statistics. We applied this approach to the summary statistics from the Global Lipids Genetics Consortium study (GLGC,  $n = 1.65 \text{ M}$ )<sup>3</sup> and the summary statistics in the interaction analysis with cigarette smoking and alcohol drinking in the CHARGE GLI working group<sup>17</sup>, with replication using direct interaction tests performed in the UK Biobank (UKBB) data. Although the UKBB data accounted for about one third of sample in the GLGC consortium, theoretical work suggests that such replication is statistically independent (Supplementary Note)

## Results

### Testing $G \times E$ and mediation based on Mendelian randomization (MR)

Traditionally a genome-wide interaction study (GWIS) with an environmental exposure on a quantitative trait  $Y$  is modeled through a linear regression:

$$Y = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G \times E + \epsilon, \quad (1)$$

where  $\beta_1, \beta_2$  and  $\beta_3$  correspond to the ‘main’ effect of  $G$  (in the presence of  $E$ ), the main effect of  $E$  and the interaction effect of  $G \times E$ , respectively, and  $\epsilon$  is a random noise. Here  $G$ ,  $E$ , and  $G \times E$  represent a genotype value, environmental factor, and their interaction respectively. For simplicity, we do not include any covariates, but it will not affect the general conclusion. The interaction effect is evaluated by the *direct test* statistic  $T_{direct} = \hat{\beta}_3^2 / \text{var}(\hat{\beta}_3)$ , where  $\hat{\beta}_3$  refers the estimate from the regression model (1). Theoretical work indicates that the test statistics for the main effect  $\beta_1 = 0$  and the interaction effect  $\beta_3 = 0$  are correlated, with the correlation coefficient equal to  $-\mu_E / \sqrt{\mu_E^2 + \sigma_E^2}$ , where  $\mu_E$  and  $\sigma_E^2$  are the mean and variance of the environmental factor in the data<sup>14</sup>. However, the power of the direct test is usually low because of the collinearity between  $G$  and  $G \times E$  which induces a covariance between the estimates of  $\beta_1$  and  $\beta_3$ . This covariance produces uncertainty (i.e. larger standard error) which by itself reduces power for testing either  $\beta_1$  or  $\beta_3$ , even if the underlying true model includes  $G \times E$  alone (i.e.,  $\beta_1 = 0$  and  $\beta_3 \neq 0$ )<sup>10,14</sup>.

In practice, a GWAS is routinely conducted first when studying the genetic contribution to a trait, which is typically done through a linear regression model without including environmental factors, i.e.,

$$Y = \alpha_0 + \alpha G + \epsilon, \quad (2)$$

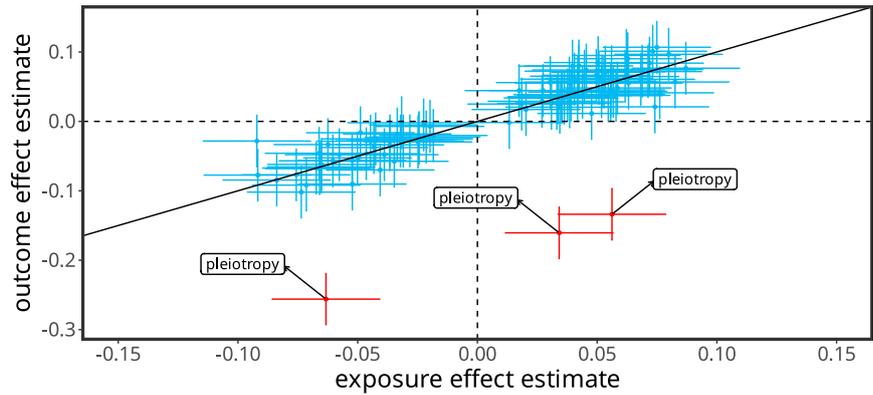
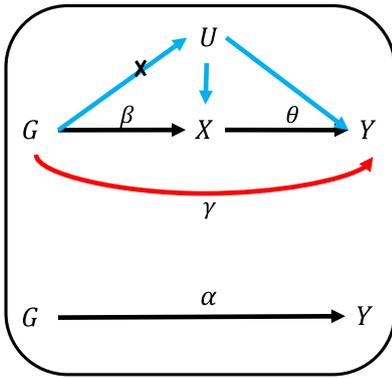
where we refer to  $\alpha$  as the ‘marginal’ effect from a GWAS (in the absence of  $E$ ) to differentiate from the main effect  $\beta_1$  in model (1). We show that  $\alpha - \beta_1 = \frac{\rho \sigma_{E1}}{\sigma_{G1}} \beta_2 + (\mu_{E1} + \frac{\rho \sigma_{E1}}{\sigma_{G1}}) \beta_3$ , where  $\rho$  is the mediation contribution of  $G$  through  $E$ ,  $\mu_{E1}$ ,  $\sigma_{E1}$ , and  $\sigma_{G1}$  represent the environmental mean, standard deviation, and genotype standard deviation in GWAS data, respectively, suggesting that testing the hypothesis  $H_0: \alpha - \beta_1 = 0$  for the difference between the marginal and main effects is equivalent to testing for the combined effect of  $G \times E$  and mediation, and further reduces to testing for the  $G \times E$  when  $G$  and  $E$  are independent (i.e.,  $\rho = 0$ , Supplementary Note). This hypothesis can be tested by the statistic  $T_{diff} = (\hat{\alpha} - \hat{\beta}_1)^2 / \text{var}(\hat{\alpha} - \hat{\beta}_1)$ , where  $\hat{\alpha}$ ,  $\hat{\beta}_1$ , and their corresponding standard errors are estimated from the GWAS and GWIS analyses, respectively. In fact,  $T_{diff}$  and  $T_{direct}$  are equivalent when GWAS and GWIS are performed in the same data. We verified this property using real data analysis in the GLI studies<sup>17</sup>, from which the summary statistics of the marginal, main, and interaction effects are available and the marginal effect was obtained after adjusting for  $E$ . We observed that the correlation between the statistics of the  $T_{diff}$  and the direct test is 0.98 for LDL and current smoking (Supplementary Fig. 5). However, GWAS is often performed in a much larger sample than the GWIS because of data availability. The environmental exposure may have different distributions in cohorts for conducting GWAS and GWIS (i.e., different mean and variance). Furthermore, models (1) and (2) are likely to be performed by two different groups of investigators, which will bring variation across studies in trait definitions, trait measurement procedures, quality control procedures, and covariates. Moreover, the summary statistics are obtained through meta-analyses in both GWAS and GWIS analyses, which can bring additional variation and confounding factors, including population stratification and cryptic relatedness, leading to a potentially invalid comparison between the marginal and main effects. In fact, it has been reported that the confounding of population stratification is not sufficiently corrected in large GWAS<sup>25,26</sup>. Therefore, directly using  $T_{diff}$  to screen the genome can be biased even for testing the combined contribution of interaction and mediation.

To overcome this bias, we note that the marginal effect estimate  $\hat{\alpha}$  and the main effect estimate  $\hat{\beta}_1$  have a linear relationship,

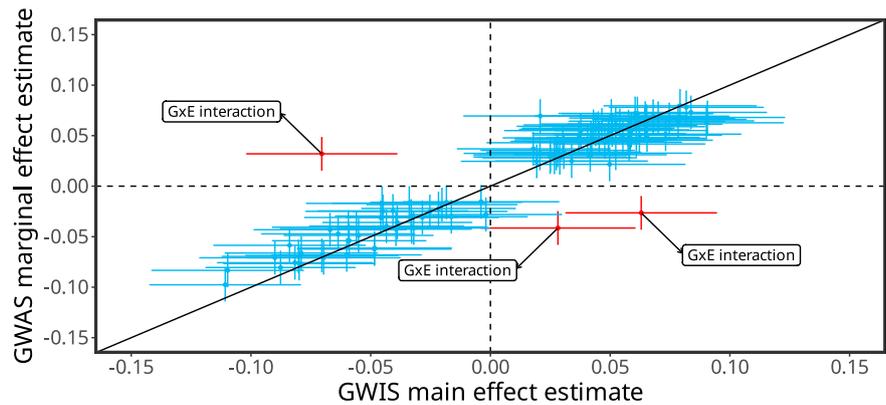
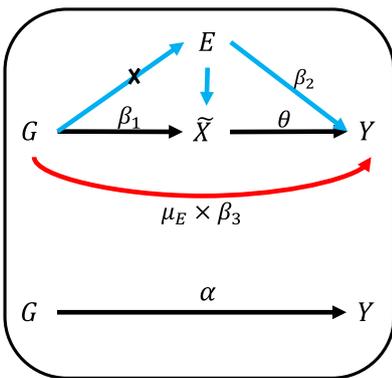
$$\hat{\alpha} = \theta \hat{\beta}_1 + \frac{\rho \sigma_{E1}}{\sigma_{G1}} \hat{\beta}_2 + \left( \mu_{E1} + \frac{\rho \sigma_{E1}}{\sigma_{G1}} \right) \hat{\beta}_3, \quad (3)$$

where  $\theta$  reflects the contribution of main effect to marginal effect, which converges to 1 when GWAS and GWIS are conducted using homogeneous measurements of phenotypes and environments (“Methods”). The genetic variants with no  $G \times E$  and no mediation will fall on the regression line but the variants with  $G \times E$  or mediation will depart from this line. We do not expect this pattern to be systematically impacted by the variation across studies. Therefore, we search the genetic variants that depart from this regression line to test the combined effect of  $G \times E$  and mediation, providing  $\theta$  can be correctly estimated. This idea is conceptually the same as the MR framework when we introduce a pseudo exposure  $\tilde{X}$ , representing a polygenic score of the trait (Fig. 1). We do not need to construct this pseudo exposure in our analysis because we work directly on the summary statistics under the MR framework. We then estimate the causal effect  $\theta$  of the pseudo exposure  $\tilde{X}$  on trait  $Y$  in the MR framework and the

A. Mendelian Randomization Pleiotropy Test.



B. Gene-Environment Interaction Test.



**Fig. 1 | Illumination of Mendelian randomization and  $G \times E$ .** **A** Left panel: the path diagram of the MR, where  $U$  refers to all confounders. Genetic variants ( $G$ ) contributing to outcome  $Y$  through mediation of exposure  $X$  are often selected as the valid genetic instrumental variables (black paths). Genetic variants contributing to  $Y$  through both black and red paths independently are horizontal pleiotropic variants. Genetic variants contributing to  $Y$  through confounders ( $U$ ) are invalid instrumental variables and need to be blocked ( $x$ ). Right panel: a scatter plot of effect sizes of genetic instrumental variants for an exposure and an outcome. Each  $+$  corresponds to the 95% confidence intervals of the exposure effect size (horizontal line segment) and the outcome effect size (vertical line segment). The horizontal pleiotropic variants (red  $+$ ) depart from the regression line and can be separated from the variants with no pleiotropic effect (blue  $+$ ). **B** Left panel: the  $G \times E$  framework, with the goal of testing  $G \times E$ . Instead of an explicit exposure, we create a pseudo exposure  $\tilde{X}$ , which can be viewed as a polygenic score for trait  $Y$  based on marginal effect sizes. However, our analysis does not require estimating this

pseudo exposure. The genetic variants associated with the pseudo exposure  $\tilde{X}$  but not through either the environment  $E$  or  $G \times E$  are valid instrumental variables. The genetic variants interacting with  $E$  can be viewed the same as horizontally pleiotropic variants in the MR framework. Genetic variants associated with  $Y$  via mediation through  $E$  can contribute to both the pseudo exposure  $\tilde{X}$  and  $Y$ , and thus have similar effects as  $G \times E$  and cannot be distinguished from  $G \times E$ . Thus, testing the combined effect of interaction and mediation is conceptually equivalent with testing the horizontally pleiotropic effect in the MR framework. Right panel: a scatter plot of genetic variants for GWIS main effects and GWAS marginal effects. Each  $+$  corresponds to the 95% confidence intervals of the GWIS main effect size (horizontal line segment) and the GWAS marginal effect size (vertical line segment). Like the horizontal pleiotropic variants in the MR framework,  $G \times E$  variants (red  $+$ ) depart from the regression line and can be separated from variants with no  $G \times E$  assuming no mediation.

$G \times E$  effect or mediation through  $E$  is tested in the same way as testing for horizontally pleiotropic variants<sup>23</sup>. In doing so, we first select a set of independent variants associated with trait  $Y$  and perform the inverse variance weighted analysis to estimate  $\theta$ , denoting as  $\hat{\theta}$ . Second, we test the  $G \times E$  or mediation of a genetic variant through  $E$  by the statistic  $T_{MR-GxE} = \frac{(\hat{\alpha} - \hat{\theta}\hat{\beta}_1)^2}{\text{var}(\hat{\alpha} - \hat{\theta}\hat{\beta}_1)} \sim \chi^2_1$ . This test can be performed by the iterative Mendelian randomization and pleiotropy (IMRP) approach<sup>23,27</sup>. The statistic  $T_{MR-GxE}$  is an asymptotically unbiased test for testing the combined effect of  $G \times E$  and mediation through  $E$  (Supplementary Note).

**Two-step procedure for testing  $G \times E$**

Note that  $T_{MR-GxE}$  likely tests for the combined effect of  $G \times E$  and mediation unless  $G$  and  $E$  are independent (i.e.,  $\rho = 0$ ). To test for  $G \times E$ , we propose a two-step procedure by using  $T_{MR-GxE}$  to screen the whole genome and then performing  $T_{direct}$  on the variants surviving the

$T_{MR-GxE}$  screen. We set the significance level at  $5 \times 10^{-8}$  for the first step ( $T_{MR-GxE}$  test), and the significance level at  $0.05/X$  for the step 2  $T_{direct}$  test, where  $X$  is the number of independent significant variants in the first step/test. This two-step procedure can increase power at the screening step when there is interaction and mediation and increases power at the direct testing step by substantially reducing the multiple comparison burden.  $T_{MR-GxE}$  and  $T_{direct}$  are not independent (Supplementary Note), and therefore, the variants detected by the two-step procedure could still reflect the contribution of mediation and  $G \times E$ , and it is necessary for further replication by performing  $T_{direct}$  in independent data. To mitigate this problem, we can exclude the variants identified through GWAS of  $E$ , which could represent large mediation effect.

**Type I error rate and power of  $T_{MR-GxE}$  and the two-step procedure**

We first performed a series of simulations to investigate the type-I error rate and power of  $T_{MR-GxE}$  in the absence of mediation. In

simulations we observed that  $E(\hat{\theta})$  is close to 1 and the estimate  $\hat{\theta}$  converges to 1 when sample size increases, which is expected by the theoretical prediction (Fig. 2A and Supplementary Fig. 6a). The direct estimate of the interaction effect  $\beta_3$  as well as of  $(\hat{\alpha} - \hat{\beta}_1\hat{\theta})/\mu_E$  is also unbiased (Fig. 2B, Supplementary Fig. 6b), although the standard error of  $(\hat{\alpha} - \hat{\beta}_1\hat{\theta})/\mu_E$  is affected by the environmental means in GWAS and GWIS. When no mediation is present, the type-I error rates for both  $T_{MR_{G \times E}}$  and the direct test are well controlled (Fig. 2C and Supplementary Fig. 6c). The power of  $T_{MR_{G \times E}}$  depends on multiple parameters, including  $\mu_E$  and allele frequency in GWAS and GWIS and is less powerful than  $T_{Direct}$  when the environmental mean in GWAS is lower (Fig. 2D and Supplementary Fig. 6d). Additional simulations for the estimates of  $\hat{\theta}$ , interaction effect  $(\hat{\alpha} - \hat{\beta}_1\hat{\theta})/\mu_E$ , type-I error rate and power are presented in Supplementary Figs. 7–9.

We next investigated the performance of  $T_{Direct}$ ,  $T_{MR_{G \times E}}$  and the two-step procedure when mediation is present and multiple variants are tested. We generated 20 independent variants with one variant having mediation, interaction, or both. All three tests have well controlled type I error rates when mediation is absent (Fig. 2E and Supplementary Fig. 10A). When mediation is present, the type-I error rate was still well controlled, although inflation can be observed for the two-step test and  $T_{MR_{G \times E}}$  when  $E$  contributes to 5% of the outcome variation and the samples between GWAS and GWIS are completely overlapped (Supplementary Fig. 10B, C). This inflation was caused by mediation and quickly disappeared when the overlapping rate between GWAS and GWIS subjects was reduced. The statistical power of  $T_{MR_{G \times E}}$  and the two-step procedure for testing  $G \times E$  was much more improved than  $T_{Direct}$  when mediation was present (Fig. 2F and Supplementary Fig. 10D–F).

### Identifying gene-smoking and gene-alcohol drinking interactions to serum lipids

We applied the two-step procedure to search for genetic variants interacting with cigarette smoking and alcohol drinking for serum lipids, using the summary statistics of high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and triglycerides (TG) from the GLGC ( $n=1.65$  M) and the CHARGE GLI ( $n=134$  K). To mitigate the effects of mediation through cigarette smoking or alcohol drinking, we excluded all loci with  $P$ -value  $< 5 \times 10^{-7}$  reported in the early GWAS of cigarette smoking status or alcohol drinking<sup>28</sup>, which represent relatively large effect sizes of variants on cigarette smoking and alcohol drinking. We observed that  $\hat{\theta}$  ranged from 0.92–1.33, 0.95–1.62, 0.83–1.25, 0.87–1.37, and 0.95–1.28 for European, African, Asian, Hispanic, and Cross-population data, respectively (Supplementary Data S1). The departure of  $\hat{\theta}$  from 1 suggests that the phenotype treatments, analysis protocols, and corrections for population structure were not identical between the GLGC and CHARGE GLI consortiums. For example, CHARGE GLI performed a natural logarithmic transformation to the lipid measurements, whereas GLGC further performed an inverse normal transformation. The number of principal components (PCs) for correcting populations was also different between GLGC and CHARGE GLI. Despite these discrepancies, we did not observe an inflation for  $T_{MR_{G \times E}}$ , with the genomic control  $\lambda$  values being close to 1 (range 0.93–1.05, Supplementary Data S2).

Using  $T_{MR_{G \times E}}$  to screen the genome, we observed 15 genome-wide significant loci consisting of 17 independent signals ( $P < 5 \times 10^{-8}$ ,  $r^2 < 0.1$ ), including 4 and 5 loci for LDL-C, 7 and 5 loci for HDL-C, and 5 and 6 loci for TG, interacting with cigarette smoking and alcohol drinking or mediating through them, respectively (Fig. 3A–C, G–I, Supplementary Data S3a). All but 3 loci have been reported to be associated with either cigarette smoking or alcohol drinking in the recent largest GWAS study with over 3 million samples<sup>29</sup>, suggesting the contribution of both  $G \times E$  and mediation. Since we already excluded the cigarette smoking and alcohol drinking variants

identified from a relatively smaller study<sup>28</sup>, these detected variants should represent modest mediation effects. Locus-specific plots of all significantly associated loci are presented in Supplementary Fig. 11, which suggest that multiple protein-coding genes are present in these loci. Strikingly, all the loci have previously been mapped to lipids traits except *RPL5P26* on chromosome 10. The  $G \times E$  or mediation loci are clearly departing from most of the lipids-associated variants (Fig. 3D–F, J–L). The population-specific  $T_{MR_{G \times E}}$  results are presented in Supplementary Fig. 12, which are also consistent with the Cross-population results, although the main contribution comes from the European population.

By applying the two-step procedure, we observed that 8 of the 17 independent signals are significant when using the direct test  $T_{Direct}$  after correcting for the 17 tests and 4 environmental factors (Table 1,  $P < 7.35 \times 10^{-4}$ ). In comparison to the direct test in GWIS, the two-step procedure identified more  $G \times E$  signals for each of the three lipid traits and four environmental factors (Supplementary Data S3b). This provides additional support for the enhanced statistical power of the two-step procedure. The tissue enrichment analysis using the GWAS-based pathway analysis tools, MAGMA<sup>30</sup> and FUMA<sup>31</sup>, suggest that these loci are enriched in liver, hippocampus, small intestine, and stomach tissues (Supplementary Fig. 13). Multiple loci were colocalized with expression quantitative trait loci (eQTLs) in the corresponding liver, lung, and blood tissues in the genotype-tissue expression database (GTEx)<sup>32</sup> (Supplementary Fig. 14).

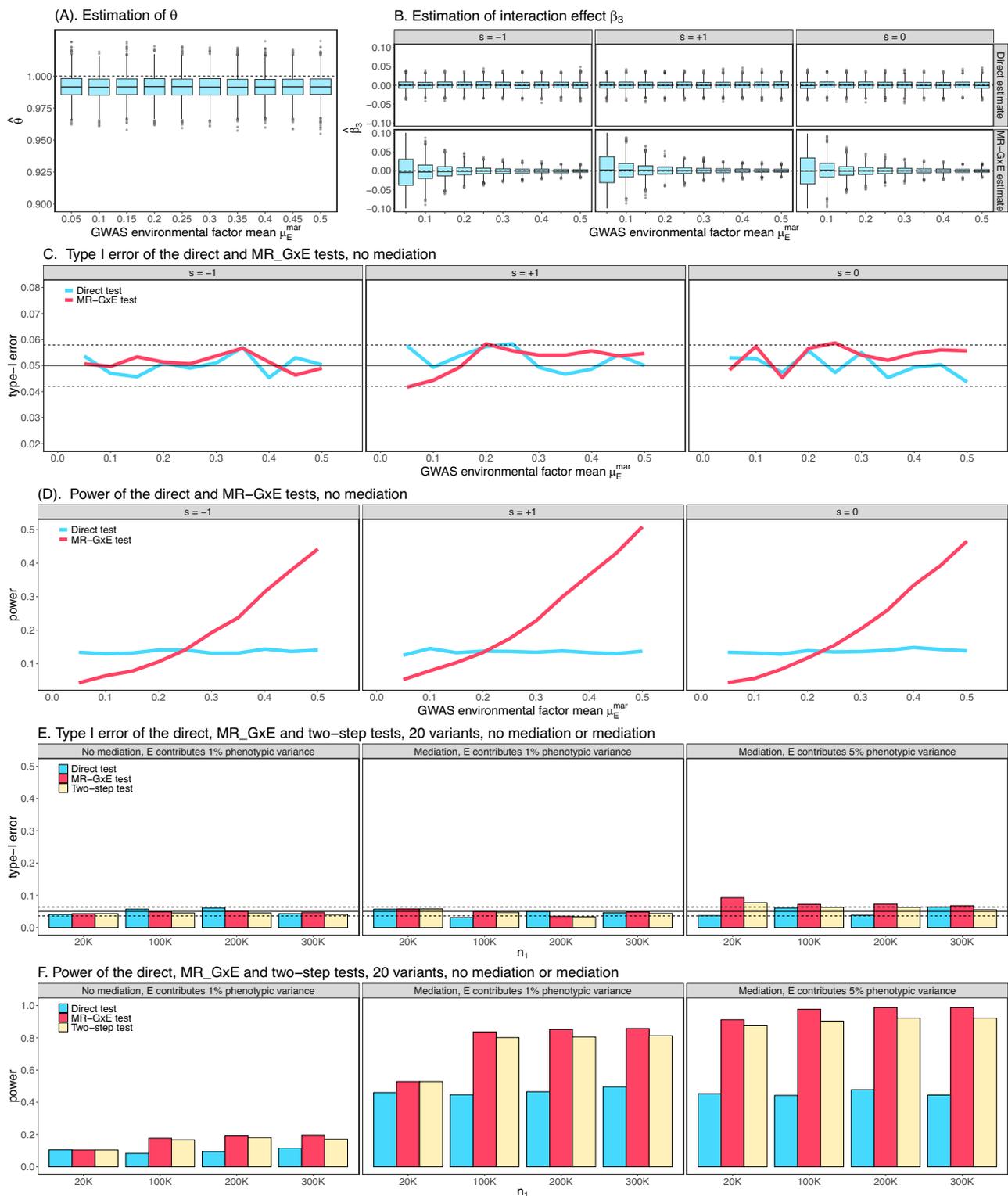
### Independent replication

We next attempted to replicate the evidence for these 8 independent signals in the UKBB. Although the UKBB data accounted for about one third of samples in the GLGC consortium, the direct test statistic  $T_{Direct}$  calculated in the UKBB is independent of  $T_{MR_{G \times E}}$ , so are the  $T_{Direct}$  test statistics calculated in UKBB and CHARGE GLI, thus qualifying this as an independent replication (Supplementary Note). Six of the 8 signals are replicated in the UKBB after adjusting for 32 tests ( $P < 1.56 \times 10^{-3}$ ), and 5 of them were genome-wide significant by the  $T_{Direct}$  test in combined CHARGE GLI and UK Biobank data (Table 1). All 8 independent signals have the same interaction direction in CHARGE GLI and UKBB except *LPL*, which is not significant in UKBB (Supplementary Data S3a). The *CETP* and *SMARCA4* loci are the only two loci with no reported mediation evidence through either cigarette smoking or alcohol drinking.

We now aim to understand if the interaction evidence observed in this study has an alternative explanation<sup>33</sup> because of linkage disequilibrium (LD) with a variant which has causal effect on cigarette smoking or alcohol drinking. To examine this, we searched if there exists a variant(s) at each of the loci in Table 1 explaining the observed interaction evidence in the UKBB. However, we did not observe such variants (Supplementary Fig. 15), suggesting that the interaction evidence presented in Table 1 is genuine. In total, we identified 5 loci consisting of 6 independent signals that have evidence of interaction with either cigarette smoking or alcohol drinking.

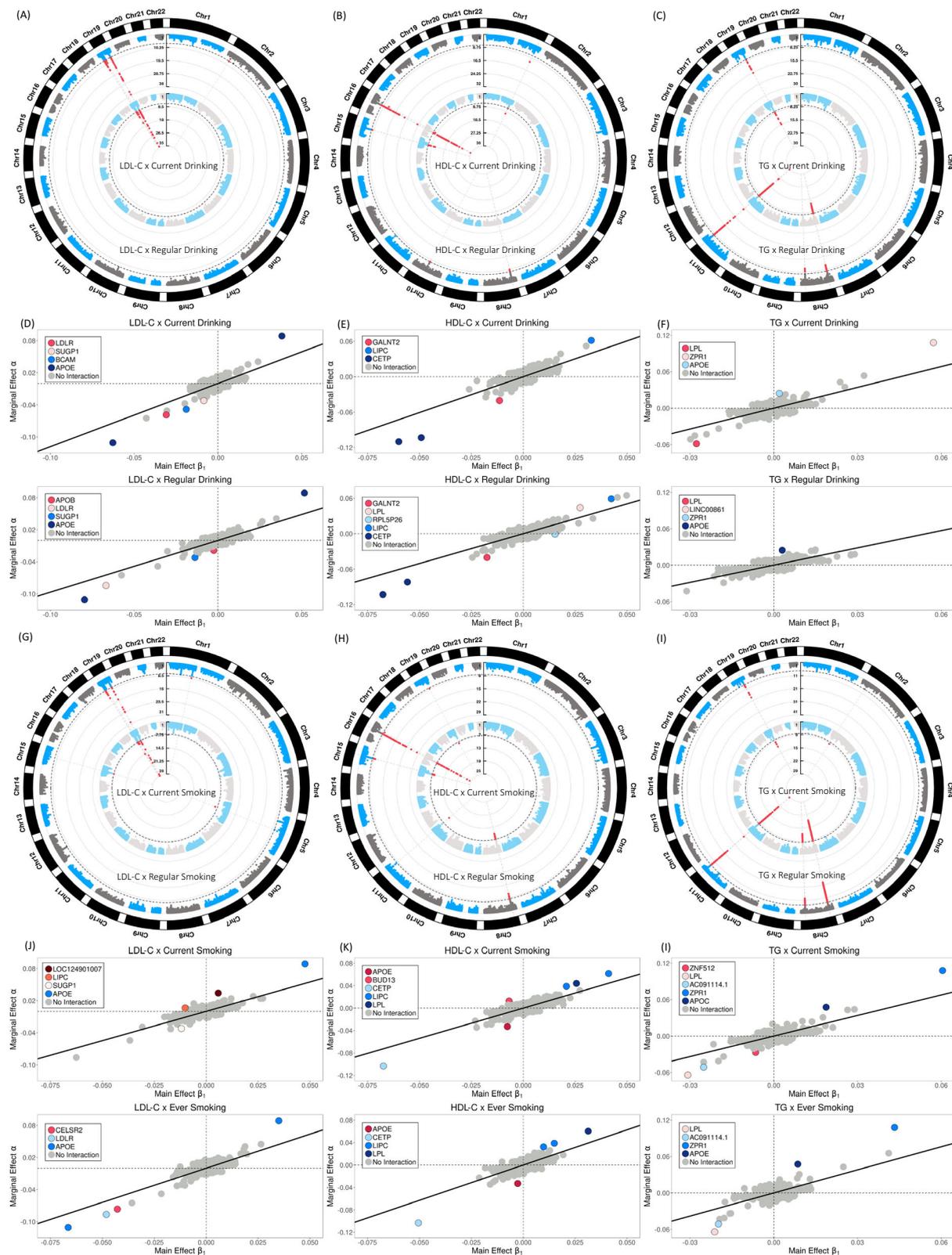
### $G \times E$ interaction and mediation to SNP heritability

Since  $\hat{\alpha} - \hat{\beta}_1\hat{\theta}$  refers to the combined interaction and mediation contribution to the marginal effect, we use  $\hat{\alpha} - \hat{\beta}_1\hat{\theta}$  to estimate the heritability contributed by the interaction and mediation through the LD score (LDSC) regression<sup>34</sup>. Note that this heritability is a lower bound of the phenotype variance contributed by the  $G \times E$  and mediation through  $E$  and is a part of the heritability estimated through the marginal effect, which is often referred to as the SNP-heritability in GWAS. In both Cross-Population (Fig. 4A) and European population (Fig. 4B), we observed significant interaction and mediation heritability ( $P < 0.03$ ) with ever cigarette smoking for LDL-C, and alcohol consumption or cigarette smoking for TG, suggesting that the heritability estimates based on marginal effects also include significant



**Fig. 2 | Simulation performance of  $T_{MR\_GxE}$  and the two-step procedure. A–D** No medication was present. The simulation details were described in “Methods”. **A** Box plots of  $\hat{\theta}$  in simulations under different environments in GWAs data. The top and bottom edges of the box plots represent the 25th and 75th percentiles of  $\hat{\theta}$ , and the horizontal middle line represents the 50th percentile. The vertical bars extend from the 25th (or 75th) percentile of  $\hat{\theta}$  to the minimum (or maximum) value of simulated data.  $E(\hat{\theta})$  is close to 1 as expected. **B** Box plots of the direct estimate of  $\beta_3$  in GWIS (top panel) and by  $(\hat{\alpha} - \hat{\beta}_1 \hat{\theta})/\mu_e$  through MR-GxE analysis (bottom panel). The box plots are interpreted the same as in (A) accordingly. Both the estimates of  $\beta_3$  and that by  $(\hat{\alpha} - \hat{\beta}_1 \hat{\theta})/\mu_e$  are unbiased. Here  $s = -1$  refers to the scenario when the

main effect and interaction effect have opposite effect directions;  $s = 0$  refers to no main effect; and  $s = 1$  refers to the scenario when the main effect and interaction effect have the same effect direction. **C** Type I error rate comparison between  $T_{MR\_GxE}$  and the direct test for different main and interaction effect directions. Both  $T_{MR\_GxE}$  and the direct test maintain the type I error rate well. **D** Power comparison between  $T_{MR\_GxE}$  and the direct test for different main and interaction effect directions. **E, F** 20 variants were tested when mediation was present or not. The simulation details were described in “Methods”. **E** Type I error comparison for  $T_{Direct}$ ,  $T_{MR\_GxE}$  and two-step procedure. The dash lines represent the 95% confidence interval. **F** Power comparison for  $T_{Direct}$ ,  $T_{MR\_GxE}$  and two-step procedure.



**Fig. 3 | Manhattan plots, marginal and main effect size comparisons.** The circle Manhattan plots of gene  $\times$  alcohol drinking interactions for **A** LDL-C; **B** HDL-C; and **C** TG, respectively. The genome-wide significant loci are presented in red dots. The marginal and main effect sizes corresponding to alcohol drinking for **D** LDL-C; **E** HDL-C, and **F** TG, respectively. The colored circles

represent the genome-wide significant loci and gray circles represent insignificant loci by  $T_{MR\_LGE}$  test. The circle Manhattan plots of gene  $\times$  cigarette smoking interactions for **G** LDL-C; **H** HDL-C; and **I** TG, respectively. The marginal and main effect sizes corresponding to cigarette smoking for **J** LDL-C; **K** HDL-C, and **L** TG, respectively.

**Table 1 | Interaction loci screened by  $T_{MR\_GxE}$  and followed by the direct test  $T_{Direct}$  in GLI (two-step test) and replicated by the direct test  $T_{Direct}$  in UK Biobank**

Mapping Gene	CHR: BP	Lead SNP	Environmental factor	Lipid traits	MR_GxE test P-value	GLI direct test P-value	UKBB direct test P-value	Combined GLI and UKBB direct test P-value
Signals identified by $T_{MR\_GxE}$ ( $P < 5E-08$ ), by $T_{Direct}$ ( $P < 7.35E-04$ ) and replicated by $T_{Direct}$ in UKBB ( $P < 1.56E-3$ ) or combined GLI and UKBB $T_{Direct}$ $P < 5E-8$								
<i>BUD13</i> *	11:116637146	rs12294259	Regular Drinking	TG	<b>2.47E-18</b>	<b>3.61E-06<sup>a</sup></b>	<b>1.97E-04<sup>a</sup></b>	<b>2.14E-08</b>
	11:116657561	rs3741298	Current Smoking	TG	<b>2.80E-13</b>	<b>1.16E-10<sup>a</sup></b>	4.24E-01 <sup>a</sup>	<b>6.99E-10</b>
<i>CETP</i>	16:57000696	rs8045855	Current Drinking	HDL-C	<b>6.12E-24</b>	<b>1.85E-07<sup>a</sup></b>	<b>4.97E-07<sup>a</sup></b>	<b>4.05E-12</b>
	16:57006829	rs289713	Regular Drinking	HDL-C	<b>5.01E-19</b>	<b>3.63E-07<sup>a</sup></b>	<b>3.16E-06<sup>a</sup></b>	<b>4.62E-11</b>
<i>BCAM</i> *	19:45392254	rs6857	Regular Drinking	LDL-C	<b>4.02E-12</b>	<b>1.28E-06<sup>a</sup></b>	<b>2.95E-04<sup>a</sup></b>	<b>8.57E-09</b>
<i>NECTIN2</i> *	19:45422946	rs4420638	Regular Drinking	LDL-C	<b>6.55E-36</b>	<b>4.41E-05<sup>a</sup></b>	<b>1.95E-06<sup>a</sup></b>	<b>2.08E-09</b>
<i>TOMM40</i>								
<i>APOE</i>								
<i>APCO1</i>								
<i>LPL</i> *	8:19830170	rs1569209	Current Smoking	TG	<b>4.77E-10</b>	<b>1.01E-13<sup>b</sup></b>	3.49E-02 <sup>b</sup>	<b>1.04E-13</b>
<i>SMARCA4</i>	19:11191677	rs10402112	Regular Drink	LDL-C	<b>1.85E-15</b>	<b>5.75E-04<sup>a</sup></b>	<b>9.04E-04<sup>a</sup></b>	8.04E-06
Signals identified by $T_{MR\_GxE}$ ( $P < 5E-08$ ) and by $T_{Direct}$ ( $P < 7.35E-04$ ) but failed in UKBB replication								
<i>RPL5P26</i> *	10:71533084	rs11591480	Regular Drinking	HDL-C	<b>3.34E-08</b>	<b>1.11E-04<sup>a</sup></b>	5.23E-02 <sup>a</sup>	8.69E-05
<i>ZPR1</i> **	11:116662579	rs651821	Ever Smoking	TG	<b>7.34E-17</b>	<b>3.44E-05<sup>a</sup></b>	6.87E-01 <sup>a</sup>	1.73E-04

The  $P$ -values of  $T_{MR\_GxE}$  and  $T_{Direct}$  are two-sided  $P$ -values based on Z-scores. The  $P$ -values in the last column (combined GLI and UKBB direct Test  $P$ -value) are calculated from a chi-square test with 4 degrees of freedom. All  $P$ -values were not adjusted for multiple comparisons. The bold  $P$ -values represent significant variants after adjusting for multiple comparisons.

\*The locus has been reported to be associated with cigarette smoking.

\*\*The locus has been reported to be associated with both cigarette smoking and alcohol drinking.

<sup>a</sup>The interaction effect direction is the same in GLI and UKBB. Detailed effect sizes and standard errors are presented in Supplementary Data S3a.

<sup>b</sup>The interaction effect direction is opposite in GLI and UKBB. Detailed effect sizes and standard errors are presented in Supplementary Data S3a.

contributions from  $G \times E$  and mediation through the corresponding environment factors (Supplementary Data S4).

### $G \times E$ interaction and mediation to heterogeneity of genetic effect sizes across populations

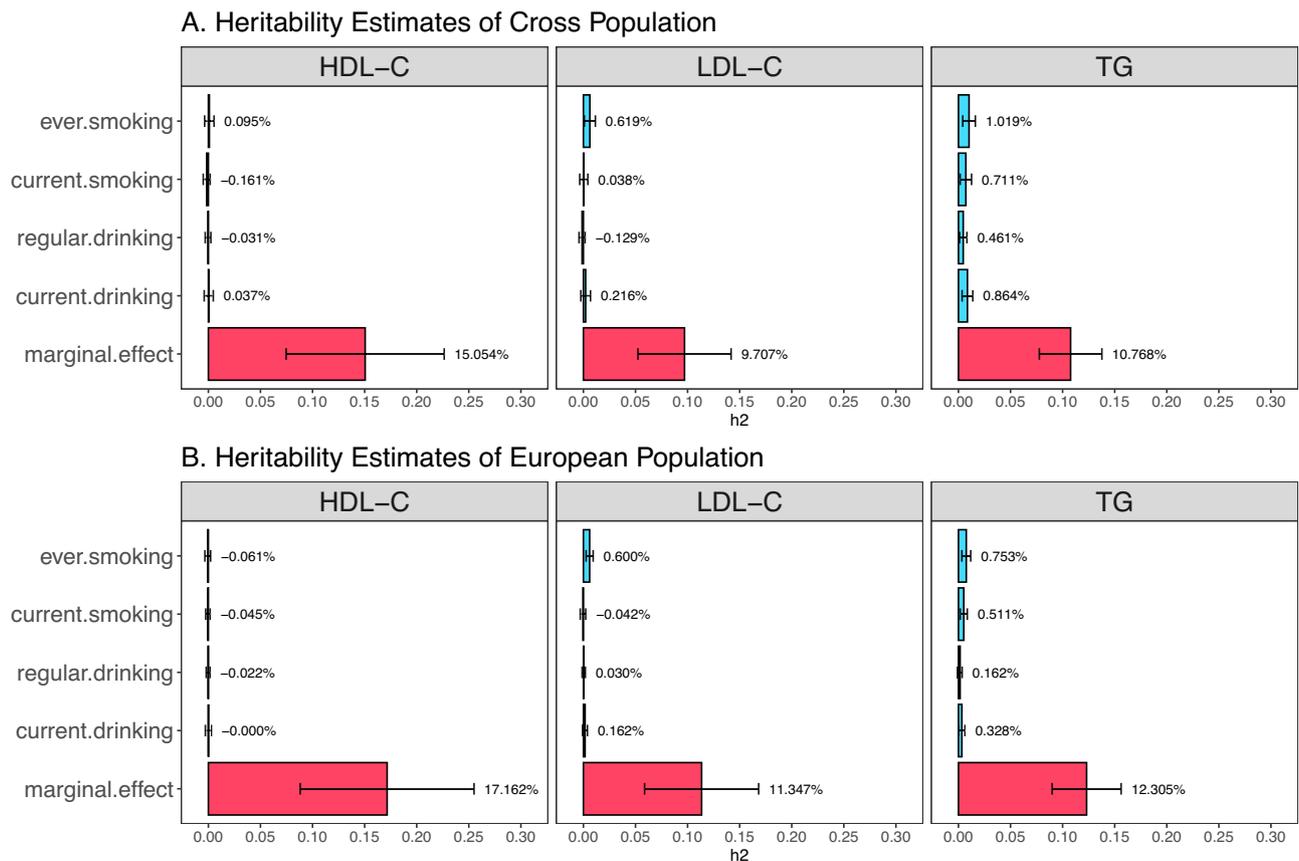
As noted in Eq. (3), the marginal effect estimate of a genetic variant in GWAS consists of the  $G \times E$  and mediation contribution when the  $G \times E$  and mediation occur. Because of the environment heterogeneity across populations, we expected that the marginal effect sizes of the variants will be less correlated across populations for the variants with than without  $G \times E$  interaction or mediation. We calculated the marginal effect size correlations between European, African, Hispanics, and Eastern Asian for these variants reported in Graham et al<sup>3</sup> after excluding the variants in Supplementary Data S3a where their  $G \times E$  interactions or mediations were observed in this study. Similarly, we calculated the marginal effect size correlations for the variants in Supplementary Data S3a. We compared the correlation and observed a median of 24.4% drop of the cross-population correlation coefficient (Fig. 5), strongly support that  $G \times E$  interactions or mediations contribute to the marginal effect size heterogeneity across populations.

## Discussion

In this study, we utilize marginal effects from GWAS to search for  $G \times E$ . We conceptually demonstrated the deep connection between detecting  $G \times E$  and MR for causal inference. Although  $T_{MR\_GxE}$  tests for the combined effect of  $G \times E$  and mediation, the two-step procedure of  $T_{MR\_GxE}$  followed by  $T_{Direct}$  in fact tests for  $G \times E$ , and its statistical power is much improved because of the following reasons: (1) the step 1  $T_{MR\_GxE}$  can increase power when a genetic variant has a mediation effect through the environmental factor. In this case, we expect a larger difference between the marginal effect and the main effect (Eq. (3)) than no mediation. (2) The difference between the marginal and main effect can further increase when the environmental distributions between GWAS and GWIS cohorts are different (Eq. (3) and Fig. 1D). (3) At the two-step procedure, multiple comparison burden is significantly alleviated because only significant variants survived at step 1 need to be examined. As demonstrated in this study, the two-step procedure

identified 8 independent signals in comparing with two by the direct test in GWIS. This is also consistent with when comparing with the direct test in GWIS, the two-step procedure identified more  $G \times E$  signals for each of the three lipid traits and four environmental factors (Supplementary Data S3b). Detecting  $G \times E$  using direct tests can be biased by unmeasured confounders due to omitting covariates in the regression models<sup>35</sup>, but the two-step procedure is robust because  $T_{MR\_GxE}$  is not affected by confounders such as population structure. Considering the advantages of the two-step procedure, we view it as a complement rather than a replacement of the direct test. This perspective arises from the fact that the two-step test necessitates additional GWAS summary statistics and may be less powerful than the direct test in some situations (Fig. 1D).

Our study demonstrated that the current heritability estimates based on marginal effects could also include contributions from  $G \times E$  and mediation through the corresponding environment factors (Fig. 4 and Supplementary Data S4). We excluded cigarette smoking- or alcohol drinking-associated variants identified from a large cigarette smoking and alcohol consumption GWAS of 1.2 million individuals<sup>28</sup> in our analysis, which mitigates the potential mediation contribution in the  $T_{MR\_GxE}$  analysis. However, among the 15 loci identified by  $T_{MR\_GxE}$ , only three were not reported in the much larger recent cigarette smoking and alcohol consumption GWAS of 3.4 million individuals, suggesting mediation through cigarette smoking and/or alcohol consumption is still present but with modest effects. Among the six  $G \times E$  variants identified, 4 are associated with either cigarette smoking or alcohol drinking, suggesting that the  $G \times E$  variants are also likely to be mediated through E and the mediation improves power to detect  $G \times E$ . Furthermore, we demonstrated that the current SNP heritability estimates based on marginal effects could also include significant contributions from  $G \times E$  and mediation through the corresponding environment factors for LDL-C and TG (Fig. 4 and Supplementary Data S4). We did not observe significant contributions from  $G \times E$  and mediation to heritability for HDL-C, potentially attributable to the relatively small sample sizes in our GWIS. Since the LDSC regression<sup>34</sup> cannot be used to estimate  $G \times E$  heritability, our estimates reflect the low bound of the interaction and environmentally mediated



**Fig. 4 | The estimated heritability of HDL-C, LDL-C, and TG using LDSC regression. A** Cross-Population. **B** European population. X-axis represents heritability in percentage. Y-axis represents the corresponding heritability estimated in percentage (marginal.effect: marginal effect heritability; current.drinking: gene and current drinking interaction effect heritability; regular.drinking: gene and regular drinking interaction effect heritability; current.smoking: gene and current smoking interaction effect heritability; ever.smoking: gene and ever smoking interaction effect heritability). Marginal effect heritability refers to the heritability estimated

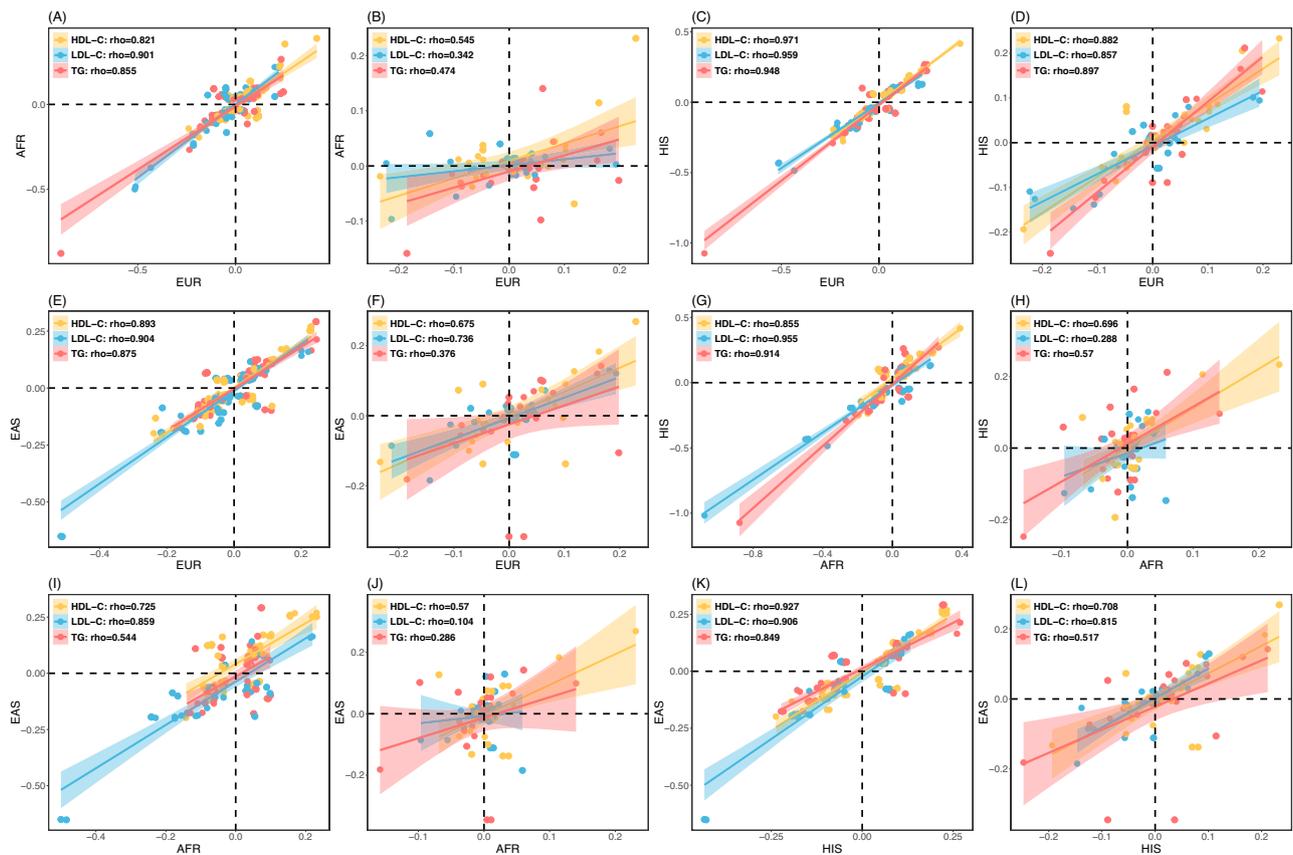
through the marginal effect  $\hat{\alpha}$ , and interaction effect heritability refers to the heritability estimated through  $\hat{\alpha} - \theta\beta_1$ . The percentage number displayed on the right side of each bar represents the estimated heritability, and the corresponding 95% confidence interval shown as horizontal error bars. For the marginal effect analysis, the sample size is 1.65 M and 1.32 M for cross-population and European population analysis, respectively. For the interaction effect analysis, the sample size is 134 K and 80 K for cross-population and European population analysis, respectively.

heritability. We therefore suggest that the current SNP heritability estimates based on the marginal genetic effects be called marginal SNP heritability, to differentiate it from narrow-sense heritability<sup>36</sup> that is defined by additive genetic actions without the inclusion of  $G \times E$  or mediation contributions. We believe this differentiation is important for correctly interpreting the current heritability estimates and understanding the genetic architecture of complex traits.

The 5 (6 independent signals) replicated loci interacting with cigarette smoking and alcohol consumption contain genes that are enriched in liver tissue, possibly reflecting the effect of alcohol drinking on aspartate amino transferase, alanine aminotransferase and  $\gamma$ -glutamyl transferase activities via the actions of numerous ingredients that alter the activities of enzymes found in the liver<sup>37</sup>. Among them, the interaction between alcohol consumption and cholesteryl ester transfer protein (*CETP*) has been reported for HDL-C and coronary artery disease<sup>38–40</sup>. The interaction between alcohol consumption and *APOE* on LDL-C has also been reported in a Mediterranean Spanish population<sup>41</sup>, while the interactions between *APOA5* and cigarette smoking and alcohol drinking status associated with elevated TG and reduced HDL-C were observed in the Chinese and Korean populations<sup>42,43</sup>. However, our study is the only well-powered study demonstrating significant evidence at the genome-wide level and the interaction loci are replicable. *SMARCA4* was reported to be associated with LDL-C in the lipids GWAS in Africans<sup>44</sup>

but not in the recent largest lipids GWAS which is predominantly European ancestry<sup>3</sup>. Overall, the marginal effect sizes of the variants are less correlated across populations for the variants with than without  $G \times E$  interaction or mediation (Fig. 5), empirically verified that  $G \times E$  and mediation contribute to marginal effect differences across different populations<sup>45</sup>. We expect that including  $G \times E$  interactions should improve polygenic risk score prediction across populations.

It is well known that causal effect estimate in MR framework can be biased when the three IV assumptions are violated. However, our goal is to detect  $G \times E$  rather than to estimate the causal effect. Detecting  $G \times E$  based on MR is less likely to be biased for these reasons: (1) the effect sizes of IVs on the pseudo exposure are all highly significant in GWAS, which represent strong IVs. (2) It is less likely to have a confounding effect between a trait and its pseudo exposure, i.e., a polygenic score. (3) The iterative Mendelian randomization and pleiotropy test is a powerful method to detect pleiotropy when the two IV conditions are satisfied<sup>23</sup>, in particular, it is expected that most of the IVs are not interacted with E. (4) Although the causal effect estimate can be affected if population structure is not well corrected, testing  $G \times E$  by  $T_{MRG \times E}$  is not. The reason is that  $T_{MRG \times E}$  can be viewed as a weighted linear regression of the effect size of GWAS on the effect size of GWIS followed by searching the outliers of variants departing from the regression line. While the regression line (equivalent to causal



**Fig. 5 | Cross-population comparison of the LDL-C, HDL-C, and TG marginal effect sizes of the variants reported in Graham et al.<sup>3</sup>** **A** EUR vs AFR, no  $G \times E$  interaction or mediation. **B** EUR vs AFR,  $G \times E$  interaction or mediation. **C** EUR vs HIS, no  $G \times E$  interaction or mediation. **D** EUR vs HIS,  $G \times E$  interaction or mediation. **E** EUR vs EAS, no  $G \times E$  interaction or mediation. **F** EUR vs EAS,  $G \times E$  interaction or mediation. **G** AFR vs HIS, no  $G \times E$  interaction or mediation. **H** AFR vs HIS,  $G \times E$  interaction or mediation. **I** AFR vs EAS, no  $G \times E$  interaction or mediation. **J** AFR vs EAS,  $G \times E$  interaction or mediation. **K** HIS vs EAS, no  $G \times E$  interaction or mediation. **L** HIS vs EAS,  $G \times E$  interaction or mediation. The variants with no  $G \times E$  interaction

or mediation are those not included in Supplementary Data S3a. The variants with  $G \times E$  interaction or mediation are those in Supplementary Data S3a. We only included independent variants. The shadow error bands represent the 95% confidence intervals. Clearly the variants without  $G \times E$  interactions or mediations have substantially larger cross-population correlations than the variants with  $G \times E$  interactions or mediations, suggesting that  $G \times E$  interactions or mediations contribute the marginal effect size heterogeneity across populations. (European (EUR), African (AFR), Hispanics (HIS), Eastern Asian (EAS)).

effect estimate in MR) can be affected by population structure, the outlier detection is not.

In summary, our  $G \times E$  approach is powerful and able to detect genetic loci interacting with environments that account for significant phenotypic variability. Our findings indicate that the contribution of  $G \times E$  in lipids is not ignorable. Our study only focuses on the interactions of genes with cigarette smoking and alcohol consumption in lipids. The cumulative interaction contribution with many environmental factors can even be greater. Detecting individual genetic loci with environmental interactions facilitates a better understanding of the genetic architecture of complex traits and can improve phenotype prediction.

## Methods

### Summary statistics data

The marginal summary statistics of high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and triglycerides (TG) from the Global Lipids Genetics Consortium study (GLGC,  $n = 1.65\text{M}$ )<sup>3</sup> were downloaded at <http://csg.sph.umich.edu/willier/public/glgc-lipids2021>.

GLGC consists of GWAS results from 1.65 M subjects representing five genetic ancestry groups: European ( $N = 1.32\text{M}$ ); African or admixed African ( $N = 99\text{K}$ ); East Asian ( $N = 146\text{K}$ ); Hispanic ( $N = 48\text{K}$ ); and South Asian ( $N = 41\text{K}$ ). We did not perform South Asian specific

analysis because there was no corresponding GWIS in the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium. The GWIS summary statistics from CHARGE gene-lifestyle (GLI) working group in this study are available via dbGaP (accession number phs000930). The CHARGE GWIS consists of 60 GWIS summary datasets: (LDL-C, HDL-C, and TG)-current smoking, (LDL-C, HDL-C, and TG)-ever smoking, (LDL-C, HDL-C, and TG)-current alcohol drinking, and (LDL-C, HDL-C, and TG)-regular alcohol drinking, for European, African or Admixed African, East Asian, Hispanic and multi-ancestry.

### QCs for performing $T_{MR_{G \times E}}$ analysis

To perform MR analysis, we aligned the GWAS summary statistics HDL-C, LDL-C, and TG from the GLGL with the corresponding GWIS summary statistics from the CHARGE gene-lifestyle consortium. We flipped the effect size if the corresponding reference allele did not match. We dropped a genetic variant if the two alleles were either {A, T} or {C, G}. We also excluded any variants with minor allele frequency (MAF) difference larger than 0.15 between GWAS and GWIS study. If multiple variants fall on the same chromosome position, we required the matched variants with MAF difference less than 0.01. We further excluded any variants with the effective sample size in GLGC trans-ethnics or European less than 100 K and the other populations (African, Hispanic, East Asian) less than 30 K. To reduce the effect by mediations through

the smoking and alcohol drinking, we excluded all loci with  $P$ -value  $< 5E-7$  identified by the GWAS of smoking status or alcohol drinking<sup>28</sup>.

### $T_{MRGE}$ analysis

To perform  $T_{MRGE}$ , we applied the Mendelian randomization (MR) software IMRP<sup>23</sup> to estimate the causal effect by considering the main effect sizes from the GWIS of the CHARGE gene-lifestyle consortium as the exposure effects, and the marginal effects from the GLGC as the outcome effects, respectively. To identify instrument variables, we first selected all the variants with the  $P$ -value  $< 5E-8$  after GC-correction in the GLGC, and then pruned them using the window size 500 KB and  $r^2$  value 0.1 by the Plink software<sup>46</sup>. We standardized the effect sizes as in<sup>27</sup>. IMRP requires the input of the correlation coefficient to account for the effect of sample overlapping between GWAS and GWIS cohorts and this correlation was calculated based on the insignificant variants ( $P$ -value  $> 0.05$ ) across the genome. After estimate the causal effect, we performed  $T_{MRGE}$ , which is equivalent to the pleiotropy test in the IMRP, to all the genetic variants across the genome.

### Independent locus definition

Independent loci were defined as the regions within 1 Mb of the most significant variants by the  $T_{MRGE}$  test. Independent signals were defined as the variants in a locus with  $r^2 < 0.1$ . The 1000 G data was used as the reference genetic data for LD calculation.

### Choosing independent variants for replication in UK Biobank

By applying  $T_{MRGE}$ , we observed that 15 genome-wide significant loci consisting of 17 independent signals ( $P$ -value  $< 5E-8$ ), including 4 and 5 loci for LDL-C, 7 and 5 loci for HDL-C, and 5 and 6 loci for TG, interacted with alcohol drinking and cigarette smoking, respectively (Supplementary Data S3a). At a locus with the  $T_{MRGE}$  significant ( $P$ -value  $< 5E-8$ ) for a lipid trait (LDL-C, HDL-C, or TG) and environment (smoking or alcohol drinking), we searched the variant with the smallest  $P$ -value of the direct test  $T_{Direct}$  among the significant variants by the  $T_{MRGE}$ . The variants with  $T_{Direct}$   $P$ -value  $< 7.35E-4$ , which correct for the 17 tests and 4 environmental factors, were considered as significant for  $G \times E$  interaction (two-step procedure). We obtained 8 independent variants in 6 loci among these 17 independent signals survived the threshold  $P$ -value =  $7.35E-4$  and these variants were further tested for the replication of the interaction effects in UK Biobank using  $T_{Direct}$  test.

### LD score regression

We applied the LD score regression to estimate heritability contributed by  $G \times E$  interaction and mediation through the environment factor  $E$ . We estimated heritability by combining all chromosomes rather than chromosome specifically. We used the R package bigsnpr<sup>47</sup> to estimate LD scores in the corresponding populations from 1000 G reference data with default settings.

### Functional mapping and annotation

We performed overall enrichment tests using the residual  $\hat{\alpha}_j - \hat{\beta}_j \hat{\theta}$  as the effect size and  $se(\hat{\alpha}_j - \hat{\beta}_j \hat{\theta})$  as the corresponding standard error. We used MAGMA<sup>30</sup> (Multi-marker Analysis of GenoMic Annotation) and DEPICT<sup>48</sup> (Data-driven Expression Prioritized Integration for Complex Traits) to identify tissues and cells that are highly expressed at genes within the  $G \times E$  loci. We also used DEPICT to test for enrichment in gene sets associated with gene ontology (GO) ontologies, mouse knockout phenotypes and protein-protein interaction networks. In addition, we reported significant enrichments with a false discovery rate 0.05. Analysis was done using the online platform FUMA GWAS.

### Colocalization

We performed colocalization analysis by using the software ezQTL (<https://analysistools.cancer.gov/ezqtl/#/home>). We chose the

public genotype-tissue expression (GTEx) v7 with eQTL<sup>32</sup> as the QTL data and chose the public European reference panels for calculating the LD data. We performed colocalization analysis between GWIS and QTL results within a locus using eCAVIAR (eQTL and GWAS Causal Variant Identification in Associated Regions)<sup>49</sup>, where the Colocalization Posterior Probability (CLPP) is used to describe the significance level of colocalization. We only recorded colocalization with CLPP  $> 0.01$ , as suggested by the authors of eCAVIAR.

### UK Biobank individual level data for replication

The UK Biobank (UKBB)<sup>50</sup> individual-level data used for replications were available through Application ID: 81097. Quality Controls Participants in the UKBB were genotyped using a custom Affymetrix UK Biobank Axiom array<sup>51</sup>. Genotypes were imputed by the UKBB using the Haplotype Reference Consortium reference panel<sup>52</sup> with imputation  $r^2$  value greater than 0.3. Related individuals with pairwise kinship coefficient greater than 0.0884 (suggested by UKBB) were removed from analysis, resulting in  $N = 445,424$  individuals of European, African, and East Asian ancestries. The principal components were calculated by UKBB with genotype data within each ancestry to account for population structure. These data were independent of GLI cohorts and consisted of European, African, and Asian individuals (race determined using UKBB field ID 21000-0.0) in UKBB who were unrelated (genetic kinship coefficient less than 0.0884; 22021-0.0). Linear regression model (1) in main text was performed. Covariates included age at assessment (21003-0.0), age<sup>2</sup>, sex (31-0.0), the first 10 PCs (22009-0.1 to 22009-0.10), and environment exposure, a genetic variant and their interaction. Environmental exposures included ever/never smoking status (20116-0.0), current/non-current smoking status (20116-0.0), and alcohol intake frequency (1558-0.0).

Analogous to the  $G \times E$  analysis in ref. 17, HDL-C (30760-0.0) and TG (30870-0.0) measurements were natural log transformed and LDL-C measurements (30780-0.0) were converted from mmol/L to mg/dl then multiplied by a factor of 0.7 if there was a history of lipid-lowering medication (6177-0.0) present. LDL-C measurements were therefore considered no medication if there were missing values for medication history. This introduced missing values in LDL-C for 248,419 individuals.

### Theoretical properties of $T_{MRGE}$

In MR analysis, the instrumental variables are independent and are genome-wide significant variants selected from GWAS. Let  $\hat{\beta}_{1,j}, \hat{\beta}_{2,j}, \hat{\beta}_{3,j}$  and  $\hat{\alpha}_j, j = 1, \dots, m$ , be the corresponding effect size estimates in GWIS (model 1) and GWAS (model 2) for the  $m$  instrument variables.

The causal effect  $\theta$  of the inverse variance weighted (IVW) is estimated by

$$\hat{\theta} = \arg \min_{\theta} \left\{ \frac{1}{m} \sum_{j=1}^m \frac{(\hat{\alpha}_j - \hat{\beta}_{1,j} \theta)^2}{\text{var}(\hat{\alpha}_j)} \right\}. \quad (4)$$

It is much simply to work on  $\hat{\theta}$  by standardizing the IVs and this procedure does not change the conclusion. Thus, we let  $\sigma_{G,j}^2 = 1, j = 1, \dots, m$ , in both GWAS and GWIS data. Further, we let the phenotype residue variance  $\sigma^2 = 1$ . By equation (S15) in Supplementary Note, we

$$\text{have } \text{var}(\hat{\alpha}_j) = n_1^{-1} j = 1, \dots, m, \text{ and } \hat{\theta} = \frac{\sum_{j=1}^m \hat{\alpha}_j \hat{\beta}_{1,j}}{\sum_{j=1}^m \hat{\beta}_{1,j}^2}.$$

Since only the variants without either  $G \times E$  interaction or mediation are valid in the MR analysis, we assume  $\rho = 0$  (no mediation) and

$\beta_{3,j} = 0$  (no interaction). We have

$$\hat{\alpha}_j = \hat{\beta}_{1,j} + \mu_{E1} \hat{\beta}_{3,j} \tag{5}$$

By applying the Slutsky's theorem, and let  $\beta_{3,j} = 0$ , we have:

$$E(\hat{\theta}) = \frac{\frac{1}{m} \sum_{j=1}^m E(\hat{\alpha}_j \hat{\beta}_{1,j})}{\frac{1}{m} \sum_{j=1}^m E(\hat{\beta}_{1,j}^2)} = \frac{\frac{1}{m} \sum_{j=1}^m \beta_{1,j}^2 + \frac{n_0}{n_1 n_2} \left(1 + \frac{\mu_{E0}^2}{\sigma_{E0}^2}\right)}{\frac{1}{m} \sum_{j=1}^m \beta_{1,j}^2 + \frac{1}{n_2} \left(1 + \frac{\mu_{E2}^2}{\sigma_{E2}^2}\right)} \tag{6}$$

Because  $\sigma_{G_j}^2 = 1$ ,  $\frac{1}{m} \sum_{j=1}^m \beta_{1,j}^2$  is the average phenotypic variance accounted by an IV. Define  $\sigma_\beta^2 = \frac{1}{m} \sum_{j=1}^m \beta_{1,j}^2$ , we have:

$$E(\hat{\theta}) = \frac{\sigma_\beta^2 + \frac{n_0}{n_1 n_2} \left(1 + \frac{\mu_{E0}^2}{\sigma_{E0}^2}\right)}{\sigma_\beta^2 + \frac{1}{n_2} \left(1 + \frac{\mu_{E2}^2}{\sigma_{E2}^2}\right)} \tag{7}$$

which converges to 1 when  $n_1$  and  $n_2 \rightarrow \infty$ . However, when  $\sigma_\beta^2$  is small (weak instrument in MR analysis), the converge of  $E(\hat{\theta})$  to 1 is slow. We also note that  $E(\hat{\theta}) \leq 1$ .

**Simulation settings without medication contribution (Fig. 2A–D, Supplementary Figs. 6–9)**

For  $i$ th individual, we generated  $m = 102$  number of independent variants through for  $j = 1, \dots, m$  by  $G_{ij}^* \sim \text{Binom}(2, p_j)$ , where  $p_j \sim \text{unifom}(0.05, 0.5)$ . We standardized genotypes by  $G_{ij} = \frac{G_{ij}^*}{2p_j(1-p_j)}$ . For the environment factor in the GWAS model, we generated  $E_{i1} \sim \mathcal{N}(\mu_{E1}, 1)$ . For the environment factor in the GWIS model, we generated  $E_{i2} \sim \mathcal{N}(\mu_{E2}, 1)$ . For the samples overlapped between the GWAS and GWIS, we generated their environment values through  $\mathcal{N}(\mu_{E2}, 1)$ . We varied the values of  $\mu_{E1}$ ,  $\mu_{E2}$  and the proportion of overlapped samples.

The main effect size of the  $j$ th variant was generate by  $\beta_{1j} \sim \mathcal{N}(0, \sigma_\beta^2)$ , where  $\sigma_\beta^2$  is the trait variance accounted for by the IVs. For the first variant, we added its interaction effect with E. The phenotype  $Y_i$  by generated by

$$Y_i = \sum_{j=1}^m G_{ij} \beta_{1j} + 0.1E_i + 0.05(G_{i1} \times E_{i1}) + \epsilon_i, \tag{8}$$

where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ . The causal effect  $\theta$  was estimated using the last 100 variants as the IVs. The power and type I error rate for  $T_{Direct}$  and  $T_{MR\_GxE}$  were calculated based on the first and second variants, respectively.

**Simulation settings without medication contribution (Fig. 2E–F), Supplementary Fig. 10**

We generated 20 independent variants by  $G_j \sim \text{Binom}(2, 0.3)$  and standardized it but without mean correction. We simulated environment E according to mediation present or not present. If no mediation, E is generated from  $\mathcal{N}(1, 1)$ . If there is mediation,  $E \sim 0.05G + \mathcal{N}(2, 0.9975)$ , or G contributes 0.25% variation of E. The phenotype is generated according to the following models:

1. No mediation and no interaction:  $Y \sim 0.1G + \gamma E + N(0, 10)$ , where  $E \sim \mathcal{N}(1, 1)$
2. Mediation but no interaction:  $Y \sim 0.1G + \gamma E + N(0, 10)$ , where  $E \sim 0.05G + \mathcal{N}(1, 0.9975)$ .
3. Mediation and interaction:  $Y \sim 0.1G + \gamma E + 0.1^* G^* E + N(0, 10)$ , where  $E \sim 0.05G + \mathcal{N}(1, 0.9975)$ .

We let  $\gamma$  take values of 1 and  $\sqrt{5}$ . We also simulated data with environment mean 0.5 (Supplementary Fig. 10). We first simulated  $n_2 = 20,000$  subjects for GWIS cohort (or main effect estimation). The sample size for marginal effect estimation varied from  $n_1 = 20,000$  to 300,000, with the 20,000 subjects in GWIS cohort was always included. For the non-overlapped subjects, we let the environment mean to be 1.5 times of the environment mean in GWIS cohort. The type I error and power for  $T_{Direct}$  and  $T_{MR\_GxE}$  were calculate by correcting for 20 tests using the Bonferroni correction. For the two-step procedure, we first applied  $T_{MR\_GxE}$  and Bonferroni correction. The variants survived after  $T_{MR\_GxE}$  were further tested by  $T_{Direct}$  and Bonferroni correction was further applied.

**Reporting summary**

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

**Data availability**

The marginal summary statistics of HDL-C, LDL-C, and TG from the Global Lipids Genetics Consortium study (GLGC,  $n = 1.65$  M) were downloaded at <http://csg.sph.umich.edu/willer/public/glgc-lipids2021>. GLGC consists of GWAS results from 1.65M subjects representing five genetic ancestry groups: European ( $N = 1.32$  M); African or admixed African ( $N = 99$ k); East Asian ( $N = 146$ k); Hispanic ( $N = 48$ k); and South Asian ( $N = 41$ k). We did not perform South Asian specific analysis because there was no corresponding GWIS in the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium. The GWIS summary statistics from CHARGE gene-lifestyle (GLI) working group in this study are available via dbGaP (accession number phs000930). The UKBB individual-level data for replications were available through Application ID: 81097.

**Code availability**

$T_{MR\_GxE}$  test was used the software IMRP which is available in the Github repository with the following link, <https://github.com/XiaofengZhuCase/IMRP><sup>23</sup>. Heritability analysis was performed by Bigsnpr, [https://privefl.github.io/bigsnpr/reference/snp\\_id\\_scores.html](https://privefl.github.io/bigsnpr/reference/snp_id_scores.html)<sup>47</sup> and LDSC regression, <https://github.com/bulik/ldsc><sup>47</sup>. FUMA: <https://fuma.ctglab.nl/><sup>31</sup>. Software ezQTL: <https://analysisstools.cancer.gov/ezqtl/#/home>. MAGMA: <https://ctg.cncr.nl/software/magma><sup>30</sup>. DEPICT: <https://github.com/perslab/depict><sup>48</sup>. Plink: <https://www.cog-genomics.org/plink2/><sup>46</sup>. The codes for performing simulations and analyzing  $G \times E$  interaction of lipids data were deposited in the Zenodo database<sup>53</sup>.

**References**

1. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
2. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
3. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
4. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
5. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
6. Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
7. Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
8. Wang, X., Elston, R. C. & Zhu, X. The meaning of interaction. *Hum. Hered.* **70**, 269–277 (2010).

9. Rao, D. C. et al. Multiancestry study of gene-lifestyle interactions for cardiovascular traits in 610 475 individuals from 124 cohorts: design and rationale. *Circ. Cardiovasc Genet.* **10**, e001649 (2017).
10. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
11. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
12. Huang, W. & Mackay, T. F. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genet.* **12**, e1006421 (2016).
13. Crow, J. F. On epistasis: why it is unimportant in polygenic directional selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 1241–1244 (2010).
14. Aschard, H. A perspective on interaction effects in genetic association studies. *Genet. Epidemiol.* **40**, 678–688 (2016).
15. Laville, V. et al. Gene-lifestyle interactions in the genomics of human complex traits. *Eur. J. Hum. Genet.* **30**, 730–739 (2022).
16. Sung, Y. J. et al. A large-scale multi-ancestry genome-wide study accounting for smoking behavior identifies multiple significant loci for blood pressure. *Am. J. Hum. Genet.* **102**, 375–400 (2018).
17. Bentley, A. R. et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat. Genet.* **51**, 636–648 (2019).
18. de Vries, P. S. et al. Multiancestry genome-wide association study of lipid levels incorporating gene-alcohol interactions. *Am. J. Epidemiol.* **188**, 1033–1054 (2019).
19. Kilpelainen, T. O. et al. Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. *Nat. Commun.* **10**, 376 (2019).
20. Smith, G. D. & Ebrahim, S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
21. Smith, G. D. et al. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med.* **4**, e352 (2007).
22. Gage, S. H., Davey Smith, G., Ware, J. J., Flint, J. & Munafò, M. R. G = E: what GWAS can tell us about the environment. *PLoS Genet* **12**, e1005765 (2016).
23. Zhu, X., Li, X., Xu, R. & Wang, T. An iterative approach to detect pleiotropy and perform Mendelian Randomization analysis using GWAS summary statistics. *Bioinformatics* **37**, 1390–1400 (2021).
24. Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
25. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, e39725 (2019).
26. Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* **8**, e39702 (2019).
27. Zhu, X., Zhu, L., Wang, H., Cooper, R. S. & Chakravarti, A. Genome-wide pleiotropy analysis identifies novel blood pressure variants and improves its polygenic risk scores. *Genet Epidemiol.* **46**, 105–121 (2022).
28. Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
29. Saunders, G. R. B. et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, **612**, 720–724 (2022).
30. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
31. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
32. Consortium, G. T. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
33. Wood, A. R. et al. Another explanation for apparent epistasis. *Nature* **514**, E3–E5 (2014).
34. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
35. Keller, M. C. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol. Psychiatry* **75**, 18–24 (2014).
36. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* (Longman, 1996).
37. Wannamethee, S. G. & Shaper, A. G. Cigarette smoking and serum liver enzymes: the role of alcohol and inflammation. *Ann. Clin. Biochem* **47**, 321–326 (2010).
38. Fumeron, F. et al. Alcohol intake modulates the effect of a polymorphism of the cholesteryl ester transfer protein gene on plasma high density lipoprotein and the risk of myocardial infarction. *J. Clin. Invest.* **96**, 1664–1671 (1995).
39. Dachet, C., Poirier, O., Cambien, F., Chapman, J. & Rouis, M. New functional promoter polymorphism, CETP/-629, in cholesteryl ester transfer protein (CETP) gene related to CETP mass and high density lipoprotein cholesterol levels: role of Sp1/Sp3 in transcriptional regulation. *Arterioscler Thromb. Vasc. Biol.* **20**, 507–515 (2000).
40. Williams, P. T. Quantile-dependent expressivity and gene-lifestyle interactions involving high-density lipoprotein cholesterol. *Lifestyle Genom.* **14**, 1–19 (2021).
41. Corella, D. et al. Environmental factors modulate the effect of the APOE genetic polymorphism on plasma lipid concentrations: ecological studies in a Mediterranean Spanish population. *Metabolism* **50**, 936–944 (2001).
42. Lin, E. et al. Association and interaction of APOA5, BUD13, CETP, LIPA and health-related behavior with metabolic syndrome in a Taiwanese population. *Sci. Rep.* **6**, 36830 (2016).
43. Park, S. & Kang, S. Alcohol, carbohydrate, and calcium intakes and smoking interactions with APOA5 rs662799 and rs2266788 were associated with elevated plasma triglyceride concentrations in a cross-sectional study of Korean adults. *J. Acad. Nutr. Diet.* **120**, 1318–1329 e1 (2020).
44. Bentley, A. R. et al. GWAS in Africans identifies novel lipids loci and demonstrates heterogenous association within Africa. *Hum. Mol. Genet.* **30**, 2205–2214 (2021).
45. Patel, R. A. et al. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am. J. Hum. Genet* **109**, 1286–1297 (2022).
46. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet* **81**, 559–575 (2007).
47. Prive, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).
48. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
49. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet* **99**, 1245–1260 (2016).
50. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
51. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

52. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
53. Zhu, X. et al. Code for the manuscript A new approach to identify gene-environment interactions and reveal new biological insight in complex traits. <https://zenodo.org/records/10815731> (2024).

## Acknowledgements

This work was supported by grant R01HG011052 and R01HG01152-03S1 (to X.Z.) from the National Human Genome Research Institute (NHGRI) and R01HL118305 and R01HL156991 (to D.R.) from the National Heart, Lung and Blood Institute. This work was also supported in part by the Intramural Research Program of NHGRI through the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is also supported by the National Institute of Diabetes and Digestive and Kidney Diseases and the Office of the Director of the National Institutes of Health (Z01HG200362 to A.R.B.).

## Author contributions

X.Z. conceived and designed the study. X.Z., Y.Y., N.L., and G.L. performed analysis. X.Z. drafted the initial manuscript with inputs from others. X.Z., Y.Y., N.L., G.L., A.R.B., P.S.d.V., M.B., A.C.M., C.N.R., W.J.G., D.R., and H.A. critically revised and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-47806-3>.

**Correspondence** and requests for materials should be addressed to Xiaofeng Zhu.

**Peer review information** *Nature Communications* thanks Wei Pan and Doug Speed for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

---

## the CHARGE Gene-lifestyle Interactions Working Group

---

Xiaofeng Zhu<sup>1</sup>✉, Amy R. Bentley<sup>2</sup>, Paul S. de Vries<sup>3</sup>, Michael Brown<sup>3</sup>, Alanna C. Morrison<sup>3</sup>, Charles N. Rotimi<sup>2</sup>, W. James Gauderman<sup>4</sup>, Dabeeru C. Rao<sup>5</sup> & Hugues Aschard<sup>6,7</sup>