

A new Approach to Identify Gene-Environment Interactions and Reveal New Biological Insight in Complex traits

Xiaofeng Zhu (✉ xxz10@case.edu)

Case Western Reserve University <https://orcid.org/0000-0003-0037-411X>

Yihe Yang

Case Western Reserve University <https://orcid.org/0000-0001-6563-3579>

Noah Lorincz-Comi

Case Western Reserve University <https://orcid.org/0000-0002-0517-2499>

Gen Li

Case Western Reserve University

Amy Bentley

National Institutes of Health, NHGRI <https://orcid.org/0000-0002-0827-9101>

Paul de Vries

Human Genetics Center

Michael Brown

University of Texas

Alanna Morrison

The University of Texas Health Science Center at Houston <https://orcid.org/0000-0001-6381-4296>

Charles Rotimi

National Institutes of health <https://orcid.org/0000-0001-5759-053X>

W. James Gauderman

Keck School of Medicine of USC

Dabeeru Rao

Washington University in St. Louis

Hugues Aschard



Institut Pasteur <https://orcid.org/0000-0002-7554-6783>

Article

Keywords:

Posted Date: October 5th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3338723/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

A new Approach to Identify Gene-Environment Interactions and Reveal New Biological Insight in Complex traits

Xiaofeng Zhu^{1*}, Yihe Yang¹, Noah Lorincz-Comi¹, Gen Li¹, Amy Bentley², Paul S de Vries³, Michael Brown³, Alanna C Morrison³, Charles Rotimi², W. James Gauderman⁴, DC Rao⁵, Hugues Aschard⁶, On behalf of the CHARGE Gene-lifestyle Interactions Working Group

Affiliations:

¹Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA,

²Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA,

³Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA,

⁴Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

⁵Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA

⁶Institut Pasteur, Université Paris Cité, Department of Computational Biology, F-75015 Paris, France

* Corresponding author. Xiaofeng Zhu, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA. E-mail: xxz10@case.edu

Abstract:

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 CHARGE Gene-Lifestyle Interactions Working Group 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 identified and confirmed 5 loci (6 independent signals) interacting with either cigarette smoking or alcohol consumption for serum lipids, and empirically demonstrated 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 contribution ranges from 1.76% to 14.05% of SNP heritability of serum lipids 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^{1,2}, and have identified a large number of genetic loci that, in some cases, explain large proportions of the trait's SNP-heritability³⁻⁵. While it is commonly agreed that complex traits are influenced by genetic and environmental factors and their interactions⁶⁻⁹, there is a long-standing disagreement about the magnitude of the $G \times E$ contribution to heritability because of different theoretical models and assumptions^{10,11}. As pointed out in¹², 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 standard⁵, these 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 models well^{10,13}. 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 $G \times E$ ¹⁴. 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. Instead, 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)^{9,15}, where only a modest number of genetic loci have been identified through testing for $G \times E$ alone¹⁶⁻¹⁹. 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^{20,21}. Although MR approaches have been used to explain $G \times E$ ²² and assess risk factor interactions²³, 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 the MR framework, illuminate their similarities and demonstrate that the test of horizontal pleiotropy in MR^{24,25} can be used for detecting $G \times E$. Based on this principle, one can identify novel $G \times E$ using existing summary statistics without needing costly and time-consuming new analyses from all cohorts. We applied this approach to the summary statistics from the Global Lipids Genetics Consortium study (GLGC, n=1.65M)³ and the summary statistics in the interaction analysis with cigarette smoking and alcohol drinking in the CHARGE GLI working group¹⁷, 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 β_1, β_2 and β_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 ϵ 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 this does 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 to 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 μ_E and σ_E^2 are the mean and variance of the environmental factor in the data¹⁴. 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 β_1 and β_3 . This covariance produces uncertainty (i.e. larger standard error) which by itself reduces power for testing either β_1 or β_3 , even if the underlying true model includes $G \times E$ alone (i.e., $\beta_1 = 0$ and $\beta_3 \neq 0$)^{10,14}.

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 α as the ‘marginal’ effect from a GWAS (in the absence of E) to differentiate it from the main effect β_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 ρ is the mediation contribution of G through E , μ_{E1} , σ_{E1} , and σ_{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 studies¹⁷, from which the summary statistics of the marginal, main, and interaction effects are available and the marginal effect was obtained after adjusting for E (**Fig S1**). 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^{26,27}. 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 θ 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 (**Online 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. This pattern will not be impacted by the systemic 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 θ 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 then estimate the causal effect θ of the pseudo exposure \tilde{X} on trait Y in the MR framework and the $G \times E$ effect or mediation through E is tested in the same way as testing for horizontally pleiotropic variants²⁴. In doing so, we first select a set of independent variants associated with trait Y and perform the inverse variance weighted analysis to estimate θ , denoted as $\hat{\theta}$. Second, we test the $G \times E$ or mediation of a genetic variant through E by the statistic $T_{MR_{GxE}} = \frac{(\hat{\alpha} - \hat{\theta} \hat{\beta}_1)^2}{\text{var}(\hat{\alpha} - \hat{\theta} \hat{\beta}_1)} \sim \chi_1^2$. This test can be performed by the iterative Mendelian randomization and pleiotropy (IMRP) approach^{24,28}. 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 applying $T_{MR_{GxE}}$ to screening the whole genome and then performing T_{direct} on the variants surviving the $T_{MR_{GxE}}$ screen. This two-step procedure can increase power at the screening step when there is interaction and mediation and increases power at the 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 $G \times E$. To mitigate this problem, we can exclude the variants identified through GWAS of E , which could have large mediation effects.

Type I error rate and power of $T_{MR_{GxE}}$ and the two-stage 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 ($\rho = 0$). 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 our theoretical prediction (**Fig. 2A and Fig. S2**). The interaction effect β_3 can be estimated directly by model (1) or by $(\hat{\alpha} - \hat{\beta}_1 \hat{\theta})/\mu_E$ when $\rho = 0$. We observed that both ways are unbiased (**Fig. 2B**), 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 $T_{MR_{GxE}}$ and the direct test are well controlled (**Fig. 2C and Fig. S2c**). The power of $T_{MR_{GxE}}$ depends on multiple parameters,

including μ_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 **Fig. S2d**). 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 **Fig. S3-S5**.

We next investigated the performance of T_{Direct} , $T_{MR_{GxE}}$ 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 **Fig. S6**). When mediation is present, the type-I error rate was still well controlled, although inflation can be observed for the two-step procedure and $T_{MR_{GxE}}$ when E contributes to 5% of the outcome variation and the samples between GWAS and GWIS are completely overlapped (**Fig S6**). This inflation was caused by mediation and quickly disappeared when the overlapping rate between GWAS and GWIS subjects was reduced. The statistical power of $T_{MR_{GxE}}$ and the two-step procedure for testing $G \times E$ was much more improved than T_{Direct} when mediation was present (**Fig. 2F** and **Fig. S6**).

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=1.65M) and the CHARGE GLI (n=134K). 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 drinking²⁹, 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 (**Table 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 $T_{MR_{GxE}}$, with the genomic control λ values being close to 1 (range 0.93-1.05, **Table S2**).

Using $T_{MR_{GxE}}$ 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, Table S3**). 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 samples³⁰, 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 study²⁹, these detected variants should represent modest mediation effects. Locus specific plots of all significantly associated loci are presented in **Fig. S7**, 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 **Fig. S8**, which are also consistent with the Cross-population results, although the main contribution comes from the European population.

By applying the two-stage 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**, $P < 7.35 \times 10^{-4}$). Tissue enrichment analyses using GWAS-based pathway analysis tools, MAGMA³¹ and FUMA³², suggest that these loci are enriched in liver, hippocampus, small intestine, and stomach tissues (**Fig. S9**). 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)³³ (**Fig. S10**).

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 were replicated in the UKBB after adjusting for 32 tests ($P < 1.56 \times 10^{-3}$), 5 of which were genome wide significant by the T_{Direct} test in combined CHARGE GLI and UK Biobank data (**Table 1**). All 8 independent signals had the same interaction direction in CHARGE GLI and UKBB except *LPL*, which is not significant in UKBB (**Table S3**). The *CETP* and *SMARCA4* loci were the only two loci with no reported mediation evidence through either cigarette smoking or alcohol drinking.

We next aimed to understand if the interaction evidence observed in this study had an alternative explanation³⁴ 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 (**Fig S11**), 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 can use $\hat{\alpha} - \hat{\beta}_1 \hat{\theta}$ to estimate the heritability contributed by the interaction and mediation through the LD score (LDSC) regression³⁵. 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. We observed significant interaction and mediation heritability ($P < 1.93 \times 10^{-3}$) with cigarette smoking and alcohol consumption to lipids traits, ranging from 1.76% (LDL-C, regular drinking) to 14.05% (TG, regular drinking) of SNP heritability estimated by the marginal effects (**Fig. 4, Table S4**).

$G \times E$ Interaction and Mediation to heterogeneity of genetic effect sizes across populations

As noted in equation (3), the marginal effect estimate of a genetic variant in GWAS consists of the $G \times E$ and mediation contribution when the $G \times E$ or mediation occurs. 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 $G \times E$ interaction or

mediation than without. We calculated the marginal effect size correlations between European, African, Hispanic, and Eastern Asian populations for these variants reported in Graham et al.³ after excluding the variants in **Table S3** 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 **Table S3** using the effect sizes reported in Graham et al.³. We compared the correlations and observed a median of 24.4% drop of the cross-population correlation coefficient (**Fig 5**), strongly suggesting that $G \times E$ interactions or mediations contribute to the marginal effect size heterogeneity across populations.

Discussion

To the best of our knowledge, this study is the first to 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 $T_{MR_{G \times E}}$ tests for the combined effect of $G \times E$ and mediation, the two-step procedure of $T_{MR_{G \times E}}$ followed by T_{Direct} in fact tests for $G \times E$, and its statistical power is much improved when mediation is present and type I error rate is reasonably controlled. Detecting $G \times E$ using direct tests can be biased by unmeasured confounders due to omitting covariates in the regression models³⁶, but the new two-stage procedure is robust because $T_{MR_{G \times E}}$ is not affected by confounders such as population structure.

Our study demonstrated that the current heritability estimates based on marginal effects also include significant contributions from $G \times E$ and mediation through the corresponding environment factors (**Fig. 4** and **Table S4**). We excluded cigarette smoking- or alcohol drinking-associated variants identified from a large cigarette smoking and alcohol consumption GWAS of 1.2 million individuals²⁹ in our analysis, which mitigates the potential mediation contribution in the $T_{MR_{G \times E}}$ analysis. However, among the 15 loci identified by $T_{MR_{G \times E}}$, 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 also include significant contributions from $G \times E$ and mediation through the corresponding environment factors. We therefore suggest that the current SNP heritability estimates based on the marginal genetic effects be called marginal or broad-sense SNP heritability, to differentiate it from narrow-sense heritability³⁷ 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 replicated loci (6 independent signals) interacting with cigarette smoking or alcohol consumption contain genes that are enriched in liver tissue, possibly reflecting the effect of alcohol drinking on aspartate amino transferase, alanine aminotransferase and γ -glutamyl transferase activities via the actions of numerous ingredients that alter the activities of enzymes found in the liver³⁸. Among them, the interaction between alcohol consumption and cholesteryl ester transfer protein (*CETP*) has been reported for HDL-C and coronary artery disease³⁹⁻⁴¹. The interaction between alcohol consumption and *APOE* on LDL-C has also been reported in a Mediterranean Spanish population⁴², 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^{43,44}. 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 Africans⁴⁵ but not in the recent largest lipids GWAS which is predominantly European ancestry³. Overall, the marginal effect sizes of the variants are less correlated across populations for the variants with $G \times E$ interaction or mediation than without (**Fig 5**), empirically verifying that $G \times E$ and mediation contribute to marginal effect differences across different populations⁴⁶. We expect that including $G \times E$ interactions should improve polygenetic risk score prediction across populations.

It is well known that causal effect estimates in the MR framework can be biased when any of the three IV assumptions are violated. However, the MR-based $G \times E$ approach 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 satisfied²⁴ and, in particular, it is expected that most of the IVs do not interacted with the environmental factor E .

In summary, our novel $G \times E$ 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.

URLs

GLGC summary statistics, <http://csg.sph.umich.edu/willer/public/glgc-lipids2021>,
CHARGE GLI summary statistics, [https://www.ncbi.nlm.nih.gov/projects/gap/study ID: phs000930.v10.p1](https://www.ncbi.nlm.nih.gov/projects/gap/studyID:phs000930.v10.p1)
UK Biobank, <https://biobank.ctsu.ox.ac.uk>.
1000G reference, https://privefl.github.io/bigsnpr/reference/download_1000G.html
Bigsnpr, https://privefl.github.io/bigsnpr/reference/snp_id_scores.html
LDSC regression, <https://github.com/bulik/ldsc>
FUMA, <https://fuma.ctglab.nl/>
ezQTL (<https://analysistools.cancer.gov/ezqtl/#/home>)
IMRP, <https://github.com/XiaofengZhuCase/IMRP>
GTEx, <http://www.gtexportal.org>

Acknowledgments

This work was supported by grant HG011052 and HG011052-03S1 (to XZ) from the National Human Genome Research Institute (NHGRI) and HL156991 (D.C.) 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).

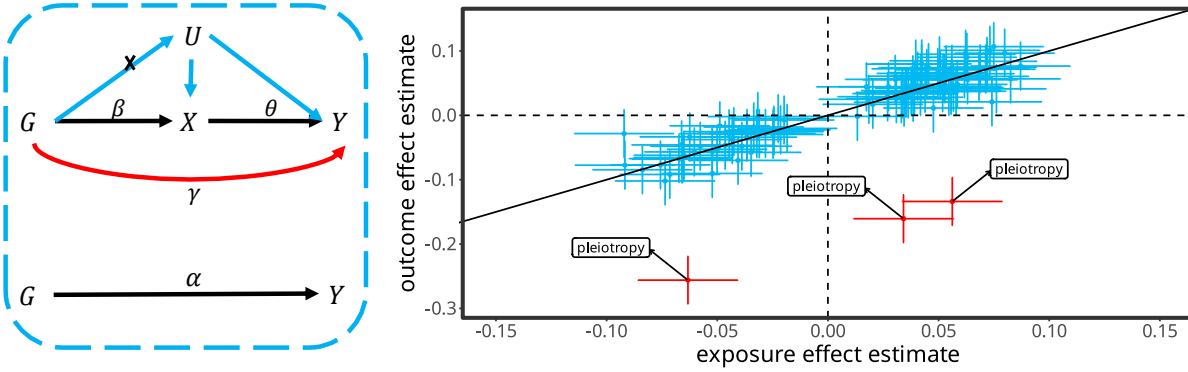
Author contributions: X.Z. conceived and designed the study. X.Z., Y.Y. , N. L. and G.L. performed analysis. X.Z. drafted the initial manuscript with inputs from others. All authors critically revised and approved the manuscript.

Competing interests: No

Data and materials availability: The marginal summary statistics of HDL-C, LDL-C, and TG from the Global Lipids Genetics Consortium study (GLGC, n=1.65M) 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=1.32M); African or admixed African (N=99k); East Asian (N=146k); Hispanic (N=48k); and South Asian (N=41k). 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.

A. Mendelian Randomization Pleiotropy Test.



B. Gene-Environment Interaction Test.

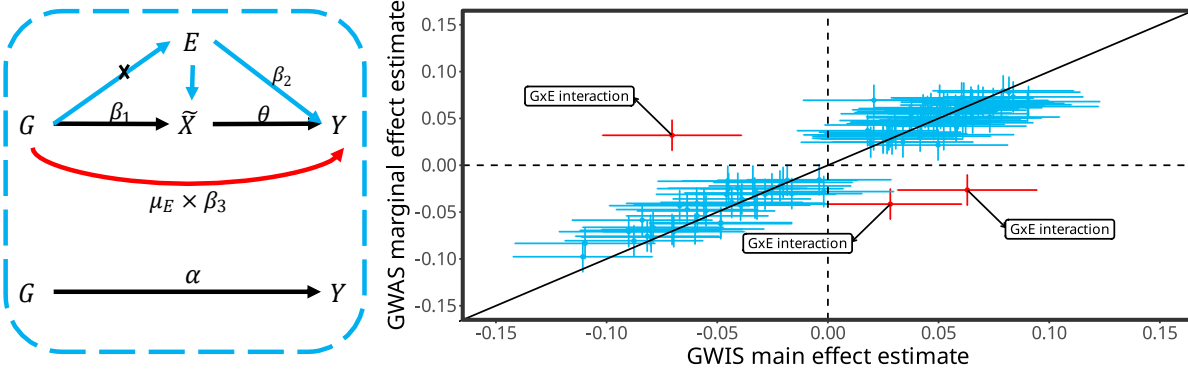


Fig. 1. Illumination of Mendelian Randomization and $G \times E$. **A.** Left panel: the path diagram of the MR, where U refers to all confounders. Genetic variants (G) contributing to outcome Y through mediation of exposure X are often selected as the valid genetic instrumental variables (black paths). Genetic variants contributing to Y through both black and red paths independently are horizontal pleiotropic variants. Right panel: the horizontal pleiotropic variants depart from the regression line and can be detected. **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 the total contribution of genetic variants to trait Y . 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: like the horizontal pleiotropic variants in the MR framework, $G \times E$ variants depart from the regression line and can be detected assuming no medication.

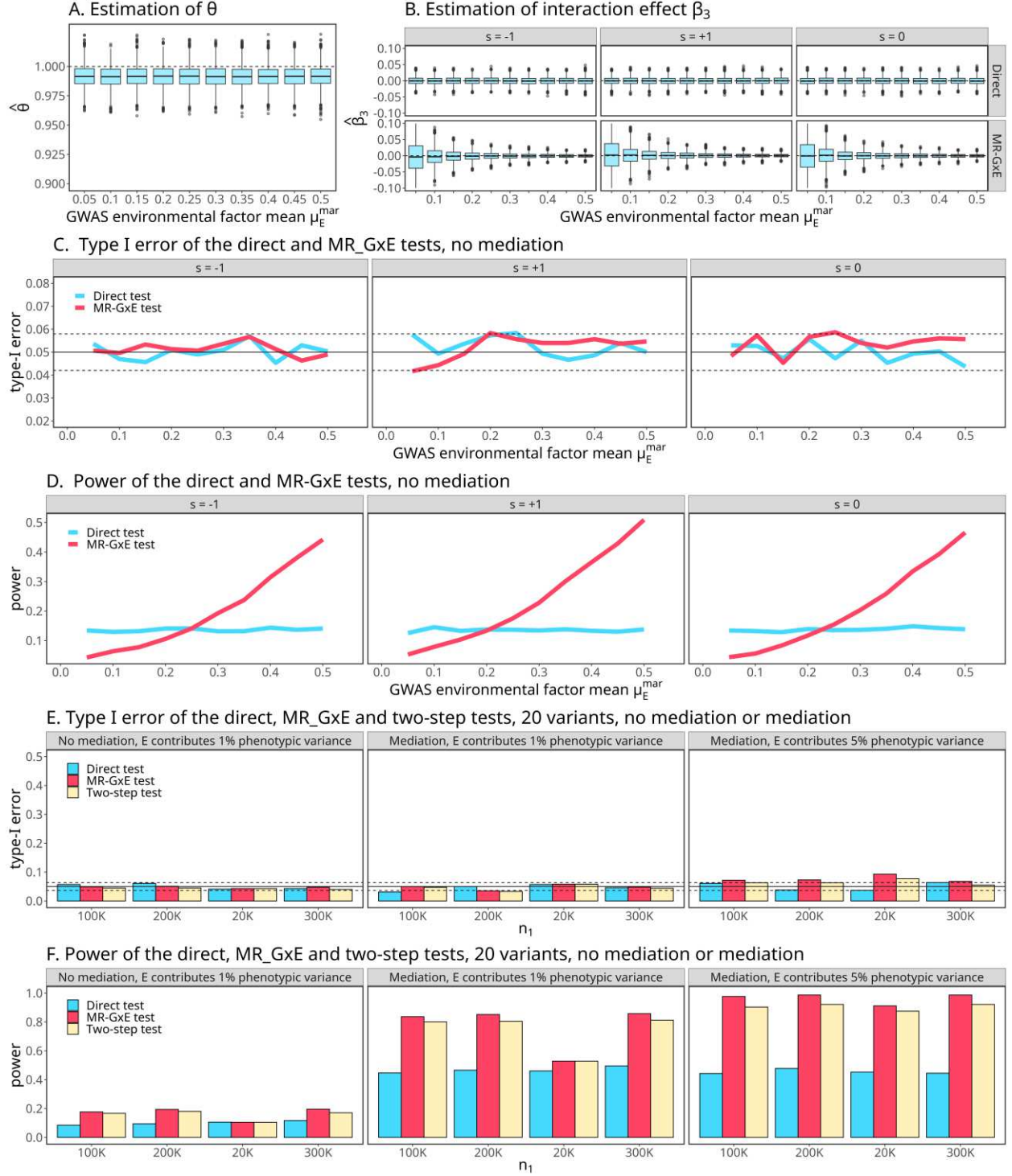


Fig. 2. Simulation performance of $T_{MR_{GxE}}$ and the two-step procedure. A-D: No medication was present. The simulation details were described in **Online Methods**. **A.** $E(\hat{\theta})$ is close to 1 as expected. **B.** The direct estimate of β_3 in GWIS or by $(\hat{\alpha} - \hat{\beta}_1 \hat{\theta})/\mu_e$ through MR analysis are both 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 **Online Methods**. **E.** Type I error comparison for T_{Direct} , $T_{MR_{GxE}}$ and two-step procedure. **F.** Power comparison for T_{Direct} , $T_{MR_{GxE}}$ and two-step procedure.

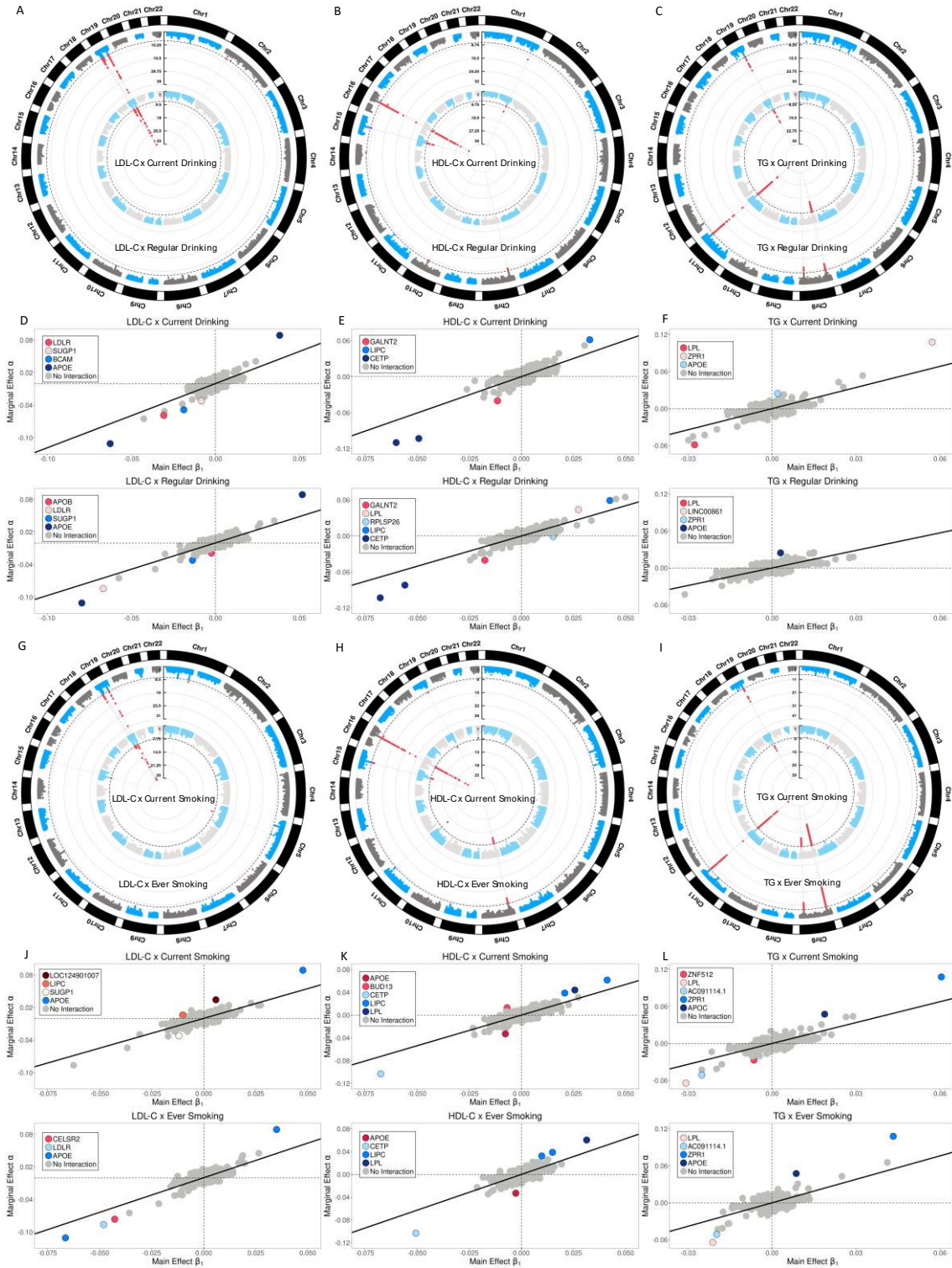


Fig. 3. Manhattan plots for interactions, marginal and main effect size comparisons. A-C: The circle Manhattan plots of gene \times alcohol drinking interactions for LDL-C, HDL-C, and TG. The genome wide significant loci are presented in red dots. **D-F:** The marginal and main effect size comparisons for LDL-C, HDL-C, and TG. The colored circles represent the genome-wide

significant loci and gray circles represent the insignificant loci by T_{MR_GXE} test. **G-I:** The counterparts of gene \times cigarette smoking for LDL-C, HDL-C, and TG, respectively.

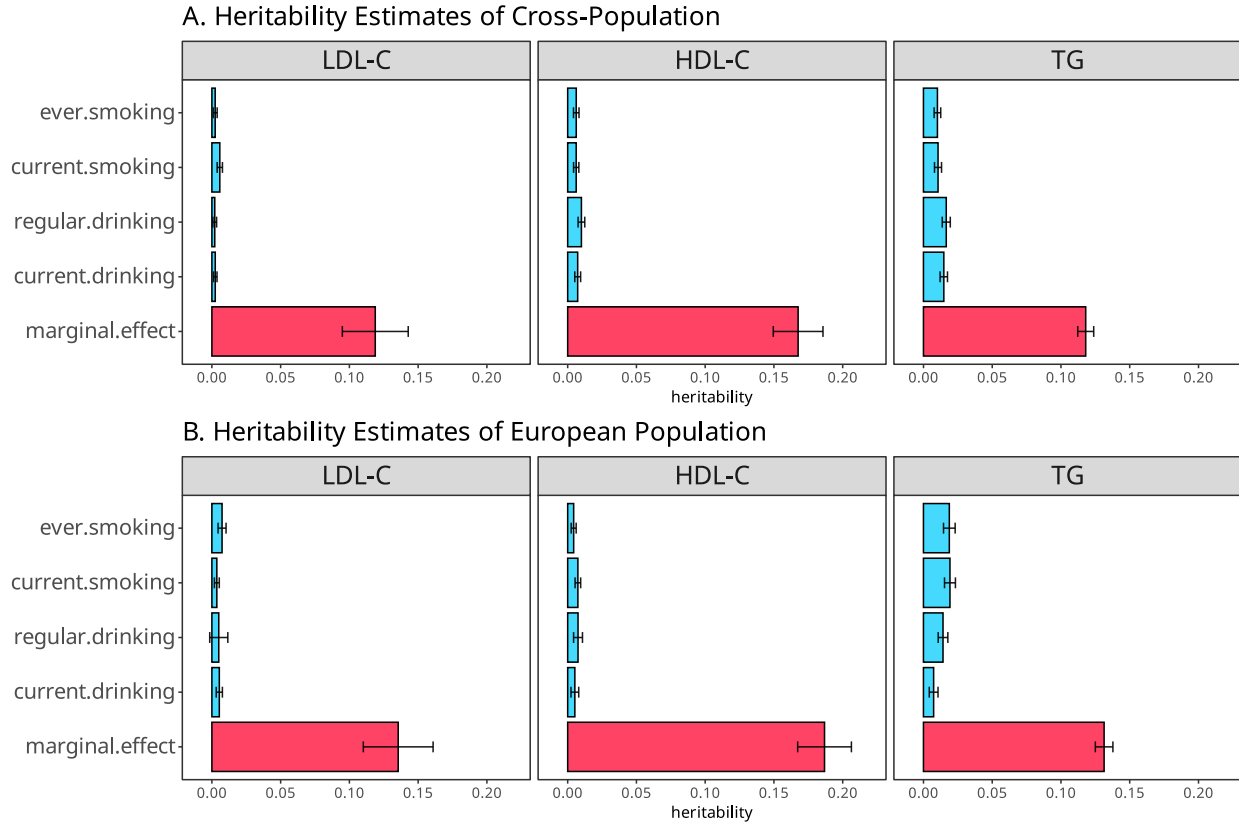


Fig. 4. The estimated heritability of Cross-Population (A) and European population (B) using the LDSC regression. Marginal effect heritability refers to the heritability estimates through the marginal effect $\hat{\alpha}$ (red bars), and interaction effect heritability refers to the heritability estimated through $\hat{\alpha} - \hat{\theta}\hat{\beta}_1$ (blue bars). Only cross-population and European population are presented. The rest population specific heritability is presented in **Table S4**.

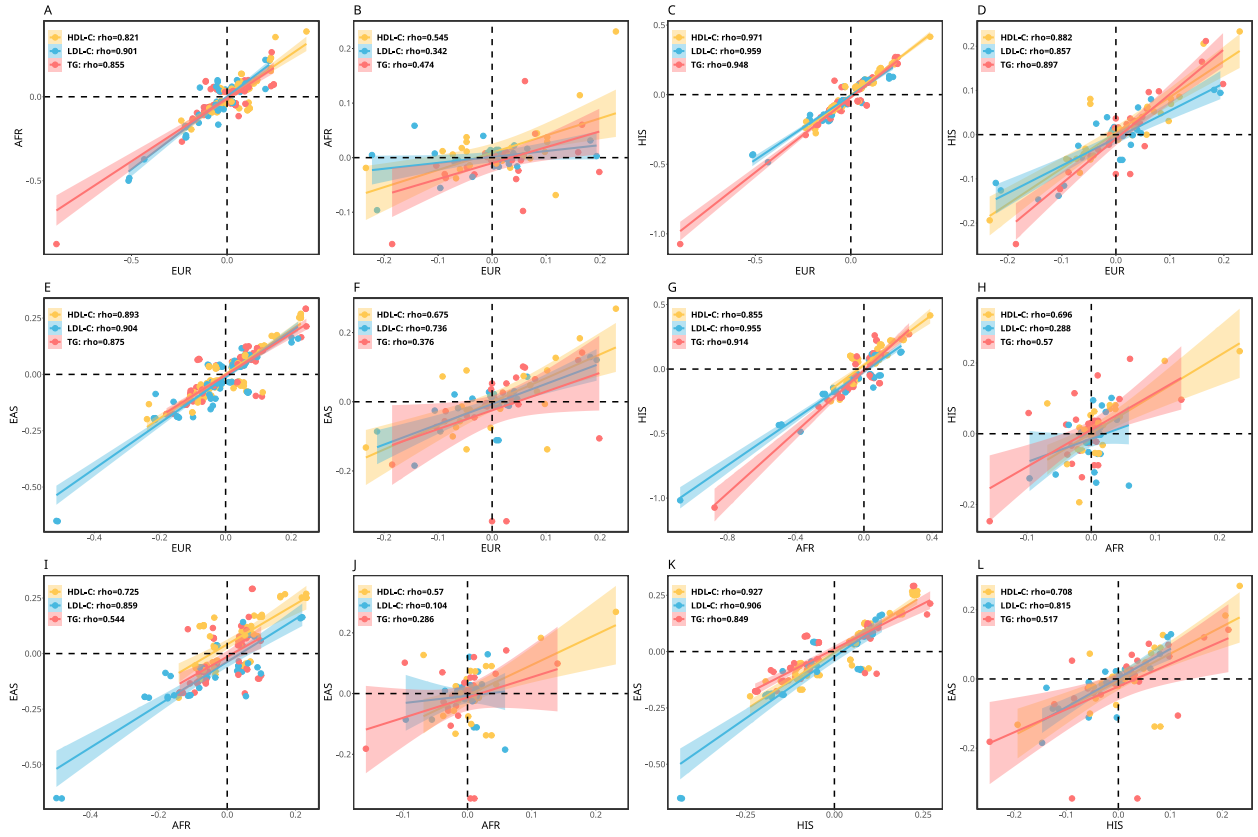


Fig. 5. Cross population comparison of the LDL-C, HDL-C and TG marginal effect sizes of the variants reported in Graham et al ³ and the independent variants in **Table S3** where their $G \times E$ interactions or mediations were observed in this study. The correlations in the first and the third columns represent the variants reported in Graham et al while the correlations in the second and fourth columns represent the variants in **Table S3**. Clearly the variants without $G \times E$ interactions or mediations have substantially larger cross population correlations than the variants with, 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)).

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	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	2.47E-18	3.61E-06^a	1.97E-04^a	2.14E-08
	11:116657561	rs3741298	Current Smoking	TG	2.80E-13	1.16E-10^a	4.24E-01 ^a	6.99E-10
<i>CETP</i>	16:57000696	rs8045855	Current Drinking	HDL-C	6.12E-24	1.85E-07^a	4.97E-07^a	4.05E-12
	16:57006829	rs289713	Regular Drinking	HDL-C	5.01E-19	3.63E-07^a	3.16E-06^a	4.62E-11
<i>BCAM</i> *	19:45392254	rs6857	Regular Drinking	LDL-C	4.02E-12	1.28E-06^a	2.95E-04^a	8.57E-09
<i>NECTIN2</i> * <i>TOMM40</i> <i>APOE</i> <i>APCO1</i>	19:45422946	rs4420638	Regular Drinking	LDL-C	6.55E-36	4.41E-05^a	1.95E-06^a	2.08E-09
<i>LPL</i> *	8:19830170	rs1569209	Current Smoking	TG	4.77E-10	1.01E-13^b	3.49E-02 ^b	1.04E-13
<i>SMARCA4</i>	19:11191677	rs10402112	RegDrink	LDL-C	1.85E-15	5.75E-04^a	9.04E-04^a	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	3.34E-08	1.11E-04^a	5.23E-02 ^a	8.69E-05
<i>ZPR1</i> **	11:116662579	rs651821	Ever Smoking	TG	7.34E-17	3.44E-05^a	6.87E-01 ^a	1.73E-04

* 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

^aThe interaction effect direction is the same in GLI and UKBB. Detailed effect sizes and standard errors are presented in **Table S3**.

^bThe interaction effect direction is opposite in GLI and UKBB. Detailed effect sizes and standard errors are presented in **Table S3**.

Online 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, $n=1.65M$)³ 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 ($N = 1.32M$); African or admixed African ($N = 99K$); East Asian ($N = 146K$); Hispanic ($N = 48K$); and South Asian ($N = 41K$). We did not perform South Asian specific analysis because there was no corresponding GWIS in the CHARGE consortium. The GWIS summary statistics from CHARGE GLI working group in this study are available via dbGaP (accession number phs000930).

QCs for performing T_{MR_GxE} analysis.

To perform MR analysis, we merged the GWAS summary statistics for HDL-C, LDL-C, and TG from the GLGC with the corresponding GWIS summary statistics from the CHARGE GLI consortium. We flipped the effect size if the corresponding reference allele did not match. We dropped the genetic variant if its 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 GLGC and CHARGE GLI studies. If multiple variants fell 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-ethics or European less than 100K and the other populations (African, Hispanic, East Asian) less than 30K. 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 drinking²⁹.

T_{MR_GxE} analysis

To perform T_{MR_GxE} , we applied the Mendelian randomization (MR) software IMRP²⁴ 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 instrumental 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 software⁴⁷. We standardized the effect sizes as in²⁸. 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 variants with P-value > 0.05 across the genome⁴⁸. After estimating the causal effect, we performed T_{MR_GxE} , 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 1Mb of the most significant variants by the T_{MR_GxE} test. Independent signals were defined as the variants in a locus with $r^2 < 0.1$. The 1000G data was used as the reference genetic data for LD calculation.

Choosing independent variants for replication in UK Biobank

By applying $T_{MR_{G \times E}}$, 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 (Table S3). At a locus with significant $T_{MR_{G \times E}}$ test 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 $T_{MR_{G \times E}}$. 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 in the two-step procedure. We observed 8 independent signals in 6 loci among all 17 independent signals surviving the threshold P-value = 7.35E-4. 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 regression³⁵ to estimate heritability contributed by $G \times E$ interaction and mediation through the environment factor E . To account for potential heterogeneity, we first estimated chromosome specific heritability and then summed across chromosomes. We used the R package bigsnpr⁴⁹ to estimate LD scores with the 1000G Phase 3 reference data and then estimated the chromosome-specific heritability with default settings. We observed that some of the chromosome specific heritability estimates were negative. We only summed the non-negative chromosome specific heritability estimates.

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 MAGMA³¹ (Multi-marker Analysis of GenoMic Annotation) and DEPICT⁵⁰ (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 ezQTL⁵¹. We chose the public genotype-tissue expression (GTEx) v7 with eQTL³³ 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)⁵², where the Colocalization Posterior Probability (CLPP) was 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)⁵³ 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 ⁵⁴. Genotypes were imputed by the UKBB using the Haplotype Reference Consortium reference panel ⁵⁵ 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. For a pair of related individuals, one was randomly excluded. 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). Linear regression model (1) in the main text was performed. Covariates included age at assessment (21003-0.0), age², sex (31-0.0), the first 10 PCs (22009-0.1 to 22009-0.10), the environmental 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 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 therefore did not consider medication if there were missing values for medication history. This introduced missing values in LDL-C for 248,419 individuals.

Theoretical properties of $T_{MR G \times E}$

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 θ of the inverse variance weighted (IVW) is estimated by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{m} \sum_{j=1}^m \frac{(\hat{\alpha}_j - \hat{\beta}_{1,j}\theta)^2}{\operatorname{var}(\hat{\alpha}_j)} \right\}.$$

It is much simpler 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 have

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

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

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

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

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

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:

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

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

We propose $T_{MR_{GxE}}$,

$$T_{MR_{GxE}} = \frac{(\hat{\alpha} - \hat{\theta} \hat{\beta}_1)^2}{\text{var}(\hat{\alpha} - \hat{\theta} \hat{\beta}_1)}, \quad (\text{S30})$$

for testing the combined effect of $G \times E$ interaction and mediation: $\frac{\rho_{\sigma E}}{\sigma_G} \beta_2 + \left(\mu_E + \frac{\rho_{\sigma E}}{\sigma_G}\right) \beta_3 = 0$. $T_{MR_{GxE}}$ can be performed by the IMRP method in MR analysis²⁴.

Our procedure to test for the $G \times E$ interaction is: 1) We apply $T_{MR_{GxE}}$ to search variants with joint effect of mediation and interaction effect; 2) we apply T_{Direct} for the variants detected by $T_{MR_{GxE}}$. To reduce the effect by the mediation, we can exclude the variants associated with E.

Simulation settings without medication contribution (Fig 2A-D and Fig. S2-S5).

For the i th individual, we generated $m = 102$ independent variants 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 μ_{E1} , μ_{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 σ_β^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

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

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The causal effect θ 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 and Fig S6).

We generated 20 independent variants by $G_j \sim \text{Binom}(2, 0.3)$ and standardized them but without mean correction. We simulated environment E according to mediation present or not present. If no mediation, E was generated from $\mathcal{N}(1, 1)$. If there was mediation, $E \sim 0.05G + \mathcal{N}(2, 0.9975)$, or G contributes 0.25% the variation of E . The phenotype was 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: $Y \sim 0.1G + \gamma E + 0.1G * E + N(0, 10)$, where $E \sim 0.05G + \mathcal{N}(1, 0.9975)$.

We let γ take values of 1 and $\sqrt{5}$. We also simulated data with environment mean 0.5 (Fig. S6). 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 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 calculated 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 surviving after $T_{MR_{GxE}}$ were further tested by T_{Direct} and Bonferroni correction was further applied.

Regerences

1. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-76 (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**, 109-18 (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**, 287-98 (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**, 2463-8 (2002).
8. Wang, X., Elston, R.C. & Zhu, X. The meaning of interaction. *Hum Hered* **70**, 269-77 (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**(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 U S A* **109**, 1193-8 (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**, 1241-4 (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. Rees, J.M.B., Foley, C.N. & Burgess, S. Factorial Mendelian randomization: using genetic variants to assess interactions. *Int J Epidemiol* **49**, 1147-1158 (2020).
24. 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).
25. 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).
26. Berg, J.J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**(2019).
27. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* **8**(2019).
28. 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).
29. 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).
30. Saunders, G.R.B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* (2022).
31. 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).
32. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
33. Consortium, G.T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
34. Wood, A.R. *et al.* Another explanation for apparent epistasis. *Nature* **514**, E3-5 (2014).
35. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
36. 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).
37. Falconer, D.S. & Mackay, T.F.C. *Introduction to quantitative genetics*, (Longman, Harlow, 1996).
38. Wannamethee, S.G. & Shaper, A.G. Cigarette smoking and serum liver enzymes: the role of alcohol and inflammation. *Ann Clin Biochem* **47**, 321-6 (2010).
39. 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**, 1664-71 (1995).
40. Dacet, 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**, 507-15 (2000).

41. Williams, P.T. Quantile-Dependent Expressivity and Gene-Lifestyle Interactions Involving High-Density Lipoprotein Cholesterol. *Lifestyle Genom* **14**, 1-19 (2021).
42. Corella, D. *et al.* Environmental factors modulate the effect of the APOE genetic polymorphism on plasma lipid concentrations: ecogenetic studies in a Mediterranean Spanish population. *Metabolism* **50**, 936-44 (2001).
43. 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).
44. 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).
45. 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).
46. 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).
47. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
48. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet* **96**, 21-36 (2015).
49. 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).
50. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015).
51. Zhang, T., Klein, A., Sang, J., Choi, J. & Brown, K.M. ezQTL: A Web Platform for Interactive Visualization and Colocalization of QTLs and GWAS Loci. *Genomics Proteomics Bioinformatics* **20**, 541-548 (2022).
52. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**, 1245-1260 (2016).
53. 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).
54. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
55. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTablesV18.xlsx](#)
- [SupplementaryMaterial.docx](#)