Genome-wide Interaction Studies by the CHARGE Gene-Lifestyle Interactions Working Group: what we've learned and what's coming next.

Vincent Laville[1], Timothy Majarian[2], Yun J Sung[3], Mary F Feitosa[4], Daniel Chasman[5], Amy R Bentley[6], Charles N Rotimi[6], L Adrienne Cupples[7,8], Paul S de Vries[9], Michael R Brown[9], Alanna C Morrison[9], Aldi T Kraja[4], Mike Province[4], Karen Schwander[3], James Gauderman[10], DC Rao[3], Alisa Manning[2,11], Hugues Aschard[1,12], on behalf of the CHARGE Gene-Lifestyle Interactions Working Group

[1]Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, 75724 Paris, France
[2] Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA, 02114, USA
[3]Division of Biostatistics, Washington University, St. Louis, MO, 63110, USA
[4]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO, 63108-221, USA
[5]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 2215, USA
[6]Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA
[7]Department of Biostatistics, Boston University School of Public Health, Boston, MA, 2118, USA
[8]Framingham Heart Study, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, 20982, USA
[9]Department of Epidemiology, Human Genetics, and Environmental Sciences, The University of Texas School of Public Health, Houston, TX, 77030, USA
[10]Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, 90032, USA
[11]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
[12]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115

Abstract

The Gene-Lifestyle Interactions Working Group has recently conducted series of multi-ancestry genome-wide association screenings (GWAS) involving up to 610,475 individuals for three lipids (total triglyceride, high-density lipoprotein, and low-density lipoprotein), and four blood pressure traits (diastolic blood pressure, systolic blood pressure, mean arterial pressure, and pulse pressure) while accounting for potential interaction effect with drinking and smoking exposures. These GWAS reported both a 1 degree-of-freedom (df) test of gene-by-exposure (GxE) interaction, and a 2 df test of main genetic effect and interaction effect. Here we synthetized these results and generated a number of cross-studies statistics, providing a global perspective on this unique initiative, and suggesting guidelines for future GxE studies. We first summarized the signals identified across all phenotype-exposure pairs and ancestries considered, highlighting similarity and specificity of gene-environment interaction effects across phenotypes and ancestries. We then performed several follow-up analyses using genome-wide summary statistics, focusing in particular on multi-SNPs approaches and the relationship between interactions effect and marginal genetic effect. Those analyses identified new associations and demonstrated potential alternatives approaches for future GxE screenings. Finally, we explored differences in heritability conditional on the exposures considered. Our work provides new evidences for the potential implication of exposure-specific pathways, opening new paths for future gene-environment interaction studies.

Introduction

The question of the role of gene-environment interaction (GxE) in human diseases remains mostly open. Despite the fact that genome-wide GxE studies have been conducted for many phenotypes, the number of identified GxE is very small as compared to the large number of genetic variants identified in marginal effect screenings. A number of issues related to the identification of GxE have been well described in the literature[1-3], in particular a very low power because of the collinearity induced between main and interaction parameters in standard interaction models[4]. As a result, the sample size needed to detect GxE has to be substantially larger than for GWAS of marginal genetic effect. Moreover, there have been very limited studies exploring potential difference in GxE association across populations, as well as studies assessing the contribution of GxE to the variance of human phenotypes, or studies exploring enrichment of GxE for specific functional mechanisms.

The Gene-Lifestyle Interactions Working Group[5] (GLIWG) is an international initiative that has the potential to address some of these challenges. It is a large-scale, multi-ancestry consortium that aims at systematically evaluating genome-wide gene–lifestyle interactions on cardiovascular disease related traits using genotypic data from up to 610,475 individuals. The consortium recently published several GxE genome-wide screenings focusing on four blood pressure phenotypes: diastolic blood pressure (DBP), systolic blood pressure (SBP), pulse pressure (PP), mean arterial pressure (MAP), and three lipids level: triglyceride (TG), high-density lipoprotein (HDL), and low-density lipoprotein (LDL). For each phenotype, a systematic genome-wide association study (GWAS) using the 1 degree of freedom (df) test for GxE interaction and the 2 df joint test of genetic and interaction effects[6] has been conducted. The results from these analyses have been published in four papers: SNP-by-alcohol interaction[7] and SNP-by-smoking interaction[8] on blood pressure, and SNP-by-alcohol interaction[9] and SNP-by-smoking interaction on lipids[10].

Here we first synthesize the GWAS results for all phenotype-exposure combinations. We highlight the importance of our large-scale initiative to discover new loci and identify loci involved in interaction effect, comparing our results against previous marginal effect screening on these phenotypes. We then show the importance of using diverse populations, providing qualitative data showing that interacting loci and variants might differ by genetic ancestry. We present power calculations for several testing procedures, highlighting both the challenges and opportunities that exist for detecting GxE interactions in a GWAS setting. Finally, we translate the effect size into variance explained of the outcomes under study, showing that in general, GxE does not explain a substantial amount of phenotypic variance for these traits. However, we show that the limited variance explained by interaction effect alone does not rule out the importance of GxE, and use the stratified *LDscore*[11] to show that the environmental exposure might indeed induce a specific and important genetic response.

Results

Overview

Our GWAS scans to date have focused on three lipid (LDL, HDL, and TG) and four blood pressure (DBP, SBP, MAP, and PP) phenotypes, each examining GxE interaction with two smoking (current smoking and ever smoking) and two alcohol (current drinking and drinking habits) exposures, for a total of 28 GWAS

(**Table 1**). All outcome-exposure pairs considered were analyzed using a two-stage approach involving up to 610,475 individuals. In stage 1, genome-wide interaction analysis was performed in up to 29 cohorts with a total of up to 149,684 individuals from multiple ancestries: European-Ancestry (EA), African-Ancestry (AA), Asian-Ancestry (ASA), and Hispanic-Ancestry (HA). In stage 2, involving up to an additional 71 studies with 460,791 individuals also from multiple ancestries, studies focused on the replication of a subset of variants from stage 1 with a *p*-value threshold of 1.0×10^{-6}. Note that the total sample size (discovery + replication) varied substantially across the trait analyzed, with an average of 330K for lipids and 460K for blood pressure traits. To ensure a fair comparison across all analyses, we re-processed all GWAS results using the same pipeline. Note that the results from the original studies and our analyses are highly concordant, but some minor differences might exist.

Summary of main results

We first synthesized the results obtained from the four studies, reporting all independent loci identified in the trans-ancestry meta-analysis in **Table1** and per ancestry in **Table S1**. Overall, the 2df joint tests trans-ancestry analyses identified a total of 5961 association signals in 1689 loci of approximately 1Mb (see ***Methods***). A total of 52% (*N*=874) loci harbored a single association signal. For the other remaining loci, the number of independent signals ranged from 2 to 72 (**Figure S1**), and the average over all regions equaled 3.5. However, as the 2df approach tests jointly the main genetic effect and interaction with the specific exposure considered, most loci overlapped across the exposures tested. Merging overlapping loci identified by different exposure scans, our studies found a total of 113 loci for HDL, 97 loci for LDL, 79 loci for TG, 75 loci for SBP, 76 loci for DBP, 76 loci for MAP, and 58 loci for PP. Interestingly, there was variability in the identification of loci per exposure considered. On average, 13% of the loci were identified by a single exposure scan, while conversely, 42% were identified by all four exposure association studies for each phenotype (**Figure 1a**). Looking at each GWAS exposure scan separately, current drinking captured 80% of all loci, drinking habits captured 61% loci, while both smoking scans identified approximately 70% of all loci (**Figure 1b**). Note that the lower number of signals for drinking habits might be partly explained by the smaller sample size on average.

The standard 1df interaction trans-ancestry test identified a much smaller number of associations. There were 23 independent association signals at 20 loci. All except two were found for HDL and LDL and drinking exposures (**Table S2**), although the sample sizes for these lipid phenotypes tended to be slightly smaller than for BP traits. The interaction effects for the top variants tended to be in the opposite direction compared to the main genetic effects (60% opposite, P_{bin}=0.1). Most of these loci were also identified by the 2df test, which is expected as the latter approach is supposed to have a much higher power to detect variants displaying both main and interaction effects (**Table S2**). We found two loci where this was not the case, although in both cases the joint test provided suggestive genome-wide significant.

A major novelty of the GLIWG is the inclusion of a large proportion of non-European individuals. More precisely, over the two stages there was 64% (*N*=390,757) of European (EA), 27% (*N*=162,369) of Asian (ASA), 5% (*N*=33,606) of African (AA) and 4% (*N*=22,612) of Hispanic (HA) ancestries. For the 2df test, the total number of significant associations per ancestry was proportional to the available sample size (**Table S1**). There was 1570, 254, 152, and 150 loci identified by this approach in EA, ASA, AA, and HA ancestries, respectively. Deriving the overlap across ancestries for the significant loci, we found that almost all ASA and HA loci were also identified by the larger EA studies (**Figure 2a**). The only exception was for AA, for

which half of the loci were exclusively identified in this population. To account for sample size differences and assess whether top variants were persistent across populations, we also looked at independent associated single nucleotide polymorphism (SNP). We extracted for each ancestry-specific signal the *p*-value at the same top SNP for the other ancestries, and assume replication if that *p*-value was smaller than 0.05. **Figure 2b** shows this overlap over all phenotypes and per-phenotype, and confirm the enrichment for ancestry specific variants in African-ancestry population. For the 1df interaction test, and only a few loci reach genome-wide significance, with 3, 30, 13, and 0 loci identified in EA, ASA, AA, and HA ancestries, respectively. The number of significant loci was not correlated with sample size per ancestry and indicates the potential presence of ancestry specific interaction effects.

Comparison against marginal effect screening

We retrieved from the literature loci exhibiting significant marginal genetic effect on blood pressure traits[12-14] and lipid traits[15-18], and compared those associations against both 1df and 2df test from our stage 1 analysis (as some SNPs were not available at stage 2). Description of these references are provided in **Table S3**, and the list of SNPs used in **Table S4**. We derived in terms of regions defined as ±500kb around the top associated variant. Overall, GLIWG identified 239 new loci-outcome associations. Among the 360 associations retrieved from the literature, 230 were also found in our studies, while 130 associations were not replicated at genome-wide significance. Most of the new association results for lipids were identified when accounting for interaction with drinking exposures, while the majority of new blood pressure associations were identified when accounting for interaction with smoking exposures (**Table 2**). For example, 94% (N=46) of the 49 new associations with HDL were found in the gene-by-current drinking GWAS, when only 45% (N=22) were identified in the gene-by-current smoking GWAS. Conversely 92% (N=44) of the 48 new associations with PP were found in the gene-by-ever smoking GWAS, while 31% (N=15) were identified in the gene-by-drinking habits GWAS.

We also used these known variants derived in external data to assess potential enrichment of interaction effects. The distribution of the interaction effect at those variants did not indicated any strong enrichment (**Figs S1-S2**). We next considered three approaches previously described[19]: an omnibus test that combines chi-squared statistics to form a K degree of freedom statistics, where K is the number of SNPs tested jointly; an unweighted genetic risk score (uGRS)-by-exposure interaction, and a weighted genetic risk score (wGRS) where SNPs are weighted by the marginal effect reported in the original studies. Our analysis did not find any enrichment for the majority of studies except for TG and current smoking (**Table S5**). Our multi-SNP test showed a strong overall enrichment with the Omnibus approach (P=2.6x10^{-8}) for this exposure. Out of the 28 variants considered, 8 were nominally significant, with the smallest *p*-value achieved for rs4810479 (*P*=4.8e-4) (**Table S6**). This variant is 4kb upstream the gene *PLTP*, that encode a phospholipid transfer protein. Interestingly, several studies showed an effect of smoking on the expression of this gene[20,21]. More precisely, they highlighted an increased PLTP activity in smokers.

Power

We plotted in **Figure 3** the number of identified loci for both the 1df interaction test and the 2df joint test as a function of the sample size at discovery stage for both ancestry-specific GWAS and the trans-ancestry meta-analysis. The pattern for the 2df test clearly shows the expected increased number of detected signals as sample size increases, similar to trend observed for marginal effect GWAS[22]. The 1df

interaction test also showed an increasing trend. However, the gain in power was minimal, highlighting the need for even larger sample size to further increase statistical power. Looking at power per analysis, we observed several interesting features. First, there are substantial differences in performance across phenotypes-exposure pairs. Drinking and lipids appear to be the best candidates for an increase in power for both 1df and 2df tests for future studies, as opposed to *e.g.* lipids and smoking that show a smaller increase rate. Second, while we found that trans-ancestry meta-analysis to be beneficial for the 2df test, this was not the case for the 1df test. For example, the 1df test meta-analysis for HDL and LDL identified respectively 6 and 8 GxE interactions with current drinking, when the Asian-ancestry analyses identified 5 and 19 interactions for the same outcome-exposure pair, respectively. In general, some of these interactions seem to be ancestry-specific, and increased sample size in specific populations can potentially lead to new discoveries that might be missed by trans-ancestry analyses. Some of the ancestry-specific results might be explained by differences in allele frequencies across populations (**Table S2**). However, other factors might be involved, and, for example, most of the SNP identified in the Asian-specific lipids-drinking interaction appear to be common in other populations.

Power of future studies might also be potentially increased using new methods. In particular, several 2-step approaches have been proposed in the literature[23-25]. In brief, these approaches screen SNPs in Step 1 based on some test and pre-defined Step-1 significance threshold, and then test GxE interaction in Step 2 for only those SNPs that pass the screen (**Figs S3**). A key requirement of these approaches is the use of independent tests for Step 1 and Step 2. For example, one can screen in Step 1 based on the standard marginal (G only) test at the α_1 significance threshold, and for those SNPs that pass the screen apply the independent GxE test in Step 2. A Bonferroni correction to the desired overall significance level is based on only the number of SNPs tested in Step 2, which can translate into greater power to detect GxE interaction. Most models of GxE interaction induce a marginal G effect (ref.), which gives this G followed by GxE 2-step approach the potential to identify novel signals that may be missed by a standard GxE scan of all SNPs. As an example, Figure X shows a comparison of power to detect GxE interaction in a genomewide scan using the standard 1-df GxE test and the abovementioned 2-step procedure with marginal G screening in Step 1. The power of the 2-step approach is substantially greater than the standard 1-df GxE test, and is nearly as high as power for the 2-df joint test of G and GxE. The main limitation is that power relies on the relevance of the criteria used at the first step.

To assess for the potential of this 2-step method based on our data, we quantified the enrichment for variants nominally significant (*i.e. P*<0.05) for GxE interaction effect after filtering based on marginal G effect *p*-value derived both within CHARGE, and from external GWAS studies (**Figure 4**). Some phenotype/exposure pairs show evidence of enrichment for significant GxE interaction, particularly as the threshold for marginal G significance becomes more strict. The most extreme examples include TG/smoking, and DBP/current drinking. Results for the former pair is in strong agreement with our multi-SNP test of interaction using genome-wide hits (see previous section), indicating a strong correlation between marginal genetic effect and interaction effect with this exposure. The absence of enrichment for other phenotype-exposure pairs does not rule out the relevance of two-step approach, but suggests that alternative metrics might be used at step 1.

Variance explained

We used a tool we recently developed[26] to estimate the variance explained by marginal genetic effects, the joint G and GxE interaction effects, and the interaction effect only, which corresponds to the difference between the two latter estimates. Here, we included all genome-wide significant variants with an r-squared lower than 0.2 and derive our estimate across all phenotype-exposure-ancestry analysis sets (**Table S7**). Overall, marginal genetic effects explain between 0.11% and 13.12% of the total phenotypic variance with an average of 3.12%. On the other hand, the interaction terms explain between 0% and 1.14% of the phenotypic variance, corresponding to an average increase of 9.11% of the total variance explained as compared to the marginal model. The largest amount of variance explained were observed for lipids traits, (average of 4.5%, as compared to 0.65% for blood pressure phenotypes). We did not observe any major qualitative differences across populations for the marginal genetic effect and the joint parameters. Conversely, the contributions of the interaction effect display substantial heterogeneity for some phenotype-exposure pairs. In particular, the variance explained by SNP-by-current drinking interaction on LDL was markedly higher for Asian ancestry as compared to other ancestries (1.14% for ASA, 0.20% for AA, 0.14% for EA, and 0.11% for HA). This is in agreement with the results from previous sections indicating an enrichment for interaction signal in that specific analysis.

Our results are in agreement with previous studies, showing that the contribution of GxE terms on top of marginal genetic effect is relatively modest[27], and confirm the likely limited impact of discovering GxE for prediction purposes in general. However, further work would be needed to assess special cases, such as the prediction performances in strata defined by environmental exposure, which might in some situation lead to gain in predictive power[28]. Importantly, we derived the contribution of GxE following standard[29] –*i.e.* as the additional variance explained on top of the marginal genetic effect. This does not necessarily reflect the actual biological contribution of potential interaction effect, as GxE effect might be projected on the marginal genetic effect and thus accounted for by the marginal terms. Alternative approaches might address this limitation by assessing *e.g.* differences in genome-wide genetic effects conditional on the exposures considered (see next section about *Enrichment for specific annotation*).

Heritability analyses

As discussed in the previous section, the relative importance of gene-by-environment (GxE) interactions to multifactorial traits might not be well characterized by the variance explained, as the latter parameter only expresses the deviation of genetic effect relative to the mean of the exposure. Understanding the role of GxE in human traits and diseases can be challenging, because human genetic studies relies on observational data. Here, we suggest that differences in genetic mechanisms conditional on the individual exposome can be partly assessed by fine phenotypic heritability

We first estimated the heritability of the three lipids and two blood pressures (DBP and SBP) traits across all individuals and in subset defined by exposure, using the *LDscore* approach[30] applied to summary statistics from the trans-ancestry analyses (**Figure 5**). The heritabilities across the three groups (all, exposed and unexposed) were similar for ever smoking and current drinking for all phenotypes. Conversely, we found substantial differences for the two other exposures. Heritability among current smokers was systematically smaller than among unexposed, especially for lipids phenotypes (11%, 12% and 2%, on average for all, unexposed and exposed, respectively). On the other hand, heritability among high drinkers was also smaller than among low drinkers for both DBP and SBP. Both patterns might seems

surprising as several of our previous analyses indicated a higher detection rate for gene-by-smoking interaction on blood pressure and gene-by-drinking interaction on lipids. However, this pattern is concordant with our results for TG and current smoking, which shows a negative correlation of effects between marginal and interaction terms. Indeed, one naive explanation for these differences is a systematic decrease in genetic effect size in exposed individual. Previous work showed that the detection of such signal can be challenging, but future larger studies with higher power might confirm this hypothesis for TG, and other traits.

We next perform a second heritability analysis, partitioning the genetic contribution by functional annotation, also using the *LDscore* approach[11,31]. We focused here on off-the-shelf baseline annotations, but we also considered the *Genoskyline*[32], a recently proposed annotation set integrating a rich collection of epigenomic data from the Roadmap Epigenomics Project. We derived for the same three lipids, and the two blood pressure (SBP and DBP) traits the enrichment score for all annotations (**Figs S4**). Overall, we did not observe any major difference across the phenotype-exposure pairs considered, expect for TG and drinking habits. For that pair, numerous *Genoskyline* annotations were strongly enriched only in exposed individuals. This pattern suggests a potential change in the genetic architecture of TG conditional on drinking habits (i.e. the variants involved differs depending on the exposure). If valid, this hypothesis could explain the increase in power observed in our GWAS result for this exposure with TG, while at the same time the absence of correlation between marginal genetic effect and interaction effects.

Discussion

In this study, we assembled and synthesized the results from 28 gene-by-environment interaction GWAS on lipid and blood pressure phenotypes performed across four ancestries, which were recently published by the Gene-Lifestyle Interactions Working Group[5,7-10]. This transversal analysis highlighted a number of features regarding large-scale GxE analysis and trans-ancestry studies. Overall, we found the trans-ancestry 2df test to be the most powerful approach for SNP discovery, although power depended substantially on the exposure considered. Conversely, for interaction effect analysis, all our analyses pointed toward more ancestry-specific patterns, which might be due to differences in allelic frequencies at causal variants, but also to other unmeasured factors. For example, African-ancestry analyses displayed several interaction effects across all phenotypes-exposure pairs, not detected by the 2df test, and involving variants almost absent in other populations. On the other hand, Asian-ancestry cohorts showed a very specific enrichment for interaction effect between lipids –and LDL in particular– and drinking, involving variants common in all populations. Our study also found differences across the exposures considered. We noted a greater increase in detection for lipid associated variants when accounting for interaction with drinking, and a greater increase in detection for blood pressure associated variants when accounting for interaction with smoking, thus stressing the potential importance of these phenotype-exposure pairs.

Comparing our results against previous GWAS of marginal genetic effect, we found strong concordance of effects for lipid analyses, with most of the previously identified phenotype-loci association being replicated in GLIWG, and over 190 new association identified in GLIWG. Results for blood pressure were more heterogeneous, with approximately half of the known associations being replicated, and as many associations being only found either in GLIWG or in previous GWAS. Our data did not highlight any

specific explanation, but the heterogeneity of results for blood pressure might potentially involve heterogeneity of genetic effect across population. Indeed, the vast majority of unreplicated loci for DBP and SBP were found in the UK Biobank cohorts. Using top associated SNPs from these GWAS, we next performed series of multi-SNP analyses to assess the presence of interaction effect at top genetic variants with marginal effect. We did not observe much enrichment for interaction effects among those variants, except for TG and current smoking. While the univariate GLIWG studies did not found genome-wide significant interaction in this analysis, the joint analysis of the 30 top TG SNPs was highly significant. This highlight the presence of likely small, but persistent interactions between many TG associated variants and current smoking. An additional analysis exploring the correlation between marginal genetic effects and SNP-by-TG interaction confirmed this enrichment, highlighting a particular relation between genetics and smoking on this phenotype.

We also used the GLIWG summary data to derive more global GxE parameters. We first estimated the overall phenotypic variance explained by marginal genetic effect and interaction effect across all identified variants. In agreement with the previous literature, we found a relatively limited contribution of interaction effect, confirming that GxE would be of limited interest for prediction purposes in the general population. However, correlation between main and interaction, as observed for example for TG and smoking and DBP and current drinking, highlights the potential for developing improved polygenic risk score in strata defined by exposures for those phenotypes. Importantly, a modest contribution of GxE to phenotypic variance does not rule out the potential importance role of GxE in the etiology of these traits –for example, marginal model can simply capture most of the variance explained by interaction effect, thus masking more complex biological mechanism[4].

To address this possibility, we derived heritability estimates conditional on the exposure. Our analysis suggested the presence of two different patterns. First, some exposures might affect the overall effect size of the genetic variants of some phenotypes (e.g. drinking habits and blood pressure, or smoking and lipids). Such interactions are difficult to detect using agnostic screening, but other approaches such as the 2-step method might substantially boost power. Second, we found one example (TG and drinking habits) where the genetic variants involved are likely different depending on the individual exposure, suggesting heterogeneity in the biological mechanism involved in the phenotype. In this alternative model, large differences in genetic effect between exposure strata imply larger power –which is in agreement with our power plot for that specific phenotype-exposure pair–, and at the same time, the absence of correlation between marginal genetic effect and interaction effects. Future studies with larger sample size and individual level data might help validating these hypotheses and assess the relevance of others competing models. For example, differences in phenotypic variance between exposure strata might potentially explain some of the observed differences in heritability estimates.

Finally, we performed several analyses using GxE summary statistics, including multi-SNPs analysis, estimation of variance explained, and assessment of enrichment for cell-type specific genetic effect depending on the exposure. The summary data provided by the GLIWG consortium provide opportunities for numerous additional follow-up analyses. Our research group has started several multi-traits and multi-exposures analyses that can help detecting new associated variants missed by univariate analyses. Future studies, extending methodologies developed for marginal genetic effect GWAS can be used to gain further knowledge on GxE, using fine-mapping[33], co-heritability[34], or conditional analyses[35] approaches.

The Gene-Lifestyle Interactions Working Group is a unique initiative that aims at understanding better the interplay between genetics and lifestyle on human phenotypes across various ancestries. Here we presented an overview of the early GxE screenings involving SNP by drinking and smoking exposures interactions on lipid and blood pressure traits. The characteristics we identified across phenotypes, exposures, and ancestries will provide the community with new working hypothesis and guidelines for future GxE studies.

Methods

Data pre-processing

For each outcome-exposure, we had access to meta-analyses summary statistics of genome-wide GxE screenings in both the discovery and the replication stages for different populations (European, African, Asian and Hispanic ancestries). Details regarding the conduct of each study can be found elsewhere (REFS). In the discovery stage, we excluded SNPs with a MAF below 1% and with significant heterogeneous effects across individual cohorts ($P < 10^{-6}$). SNPs present in only one ancestry were excluded from trans-ancestries analyses. Variants with a p-value below $P < 10^{-6}$ were further considered in the replication stage. Trans-ancestry summary statistics in the replication stage were filtered similarly to the discovery stage. Finally, we computed meta-analyses results for the combined analyses (discovery stage + replication stage) in each individual ancestry and trans-ancestry. For each ancestry and each phenotype-exposure combination, only SNPs included in both stages were retained in the final combined dataset. All meta-analyses were computed using the METAL software (REF)

Identification of independent signals and loci

For each outcome-exposure and in each ancestry, we identified genome-wide significant variants in the combined meta-analyses ($P < 5x10^{-8}$). Independent signals were defined using the clumping framework implemented in the PLINK software (REF), using a LD threshold of 0.2 and a maximum physical distance from the top SNP of 500 kb. We used 1000 Genomes individual data as a reference panel accounting for ancestry. We kept the EUR, AFR, EAS and AMR samples as proxy for the European, African-American, Asian and Hispanic individuals. We merged all those individuals to be used as a reference panel for the trans-ancestry analyses. In addition, we also defined loci by clustering SNPs located less than 500 kb upstream or downstream the lead SNP.

Univariate and multivariate interaction test at known loci

We aimed to test whether SNPs previously identified at genome-wide significance level in marginal effect GWAS tend to display interaction effects. We retrieved from the literature loci exhibiting significant marginal genetic effect on blood pressure traits[12-14] and lipid traits[15-18]. For each locus, we selected only the most associated variants so that all candidate SNPs considered are independent from each other. We

first extracted single SNP interaction *p*-values and regression coefficients in order to assess the presence of nominally significant interactions after correction for multiple testing based on the number of selected candidates.

We next assessed the presence of potential enrichment for interaction that would be missed by the univariate approach. We considered three complementary approaches to test jointly interaction effects from multiple SNP: an omnibus test, an unweighted genetic risk score (uGRS) test, and a weighted genetic risk score (wGRS). Consider a vector of L single SNP interaction coefficient $\boldsymbol{S} = [\hat{\gamma}_1, \dots \hat{\gamma}_{\mathrm{L}}]$ and their corresponding variance-covariance matrix $\mathbf{Q}$, which off-diagonal term equal 0, and diagonal equals to the variance of each estimate $\boldsymbol{\Gamma} = \left[\sigma^2_{\hat{\gamma}_1}, \dots\ \sigma^2_{\hat{\gamma}_L}\right]$. In the standard omnibus test, all interaction effects, are tested jointly, by forming the statistics $\boldsymbol{S^T}\mathbf{Q}^{-1}\boldsymbol{S}$, which follows a chi-square distribution with L degree of freedom. For GRS-based interaction tests, we assume the GRS are built as the (weighted or unweighted) sum of risk alleles of the L candidate SNPs. Explicitly, $uGRS = \sum_{i=1\dots L} G_i$, and $wGRS = \sum_{i=1\dots L} w_i \times G_i$, where w_i is defined as marginal genetic risk estimates of the SNP. We aim at testing the significance of the $uGRS \times E$ and $wGRS \times E$. As previously demonstrated[4], the corresponding statistical tests can be respectively approximated using interaction summary statistics as $\left(\sum_{i=1\dots L} \frac{\hat{\gamma}_{\mathrm{i}}}{\sigma^2_{\hat{\gamma}_{\mathrm{i}}}}\right)^2 \Big/ \sum_{i=1\dots L} \frac{1}{\sigma^2_{\hat{\gamma}_{\mathrm{i}}}}$, and $\left(\sum_{i=1\dots L} \frac{w_i\hat{\gamma}_{\mathrm{i}}}{\sigma^2_{\hat{\gamma}_{\mathrm{i}}}}\right)^2 \Big/ \sum_{i=1\dots L} \frac{w_{\mathrm{i}}^2}{\sigma^2_{\hat{\gamma}_{\mathrm{i}}}}$, and both follow a 1 degree of freedom chi-square under the null.

Power calculation

We first explored the number of identified locus as a function of sample size. We extracted both the number of identified loci and the sample size at the discovery stage from both **Table 1** (for the trans-ancestry meta-analysis) and **Table S1** (for ancestry-specific analyses). Note that we used sample size at discovery because of the strong heterogeneity of sample size at the replication stage. We then derived the proportion of interaction effect significant at an alpha threshold of 0.05, after filtering the SNPs based on their marginal effect. In this analysis we considered only individual of European ancestry, in order to maximize the sample size while limiting issue due to genetic heterogeneity, where the top variants might differ across populations. We used two panels to determine marginal genetic effects: i) in-sample estimates derived from The Gene-Lifestyle Interactions Working Group dataset, and ii) GWAS summary statistics from external consortium data which are partly independent (*i.e.* some of the individuals from these consortia might come from CHARGE).

Variance explained

We estimated the fraction of phenotypic variance explained by the main effects, the interaction effects and those effects jointly f_G, f_I, f_J respectively using the R package VarExp [PMID=29726908]. Considering a joint regression model including interaction terms fitted in a sample of N individuals, the fraction of phenotypic variance explained by the genetic main effects of a set of SNPs, their interaction effect and jointly can be estimated using only summary statistics from the joint model by $f_G = \frac{N\left(\alpha_G'^{\,T}\Sigma^{-1}\alpha_G'\right)-q}{(N-q)var(Y)}$, $f_I = \frac{N\left(\alpha_{INT}'^{\,T}\Sigma^{-1}\alpha_{INT}'\right)-q}{(N-q)var(Y)}$ and $f_J = \frac{N\left(\alpha_G'^{\,T}\Sigma^{-1}\alpha_G' + \alpha_{INT}'^{\,T}\Sigma^{-1}\alpha_{INT}'\right)-q}{(N-q)var(Y)}$ respectively; where α_G' and

α'_{INT} denote the standardized main genetic and interaction effects and q is the rank of the SNP correlation matrix Σ. In this analysis, we estimated, for each ancestry and each phenotype-exposure combination, the fraction of genotypic variance explained by independent SNPs identified in the combined meta-analyses. The SNP correlation matrix was derived using relevant unrelated individuals from the 1000 Genomes data as a reference panel. Individuals from the EUR, AFR, EAS and AMR samples were included in the reference panel for analyses in European, African, Asian and Hispanic ancestries respectively. To avoid dealing with singular SNP correlation matrices, we only considered independent SNPs ($r < 0.2$) in the analyses, each SNP with an absolute value of correlation greater than 0.2 with another SNP was removed.

Reference

1. McAllister, K. *et al.* Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am J Epidemiol* **186**, 753-761 (2017).
2. Gauderman, W.J. *et al.* Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am J Epidemiol* **186**, 762-770 (2017).
3. Ritchie, M.D. *et al.* Incorporation of Biological Knowledge Into the Study of Gene-Environment Interactions. *Am J Epidemiol* **186**, 771-777 (2017).
4. Aschard, H. A perspective on interaction effects in genetic association studies. *Genet Epidemiol* **40**, 678-688 (2016).
5. 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).
6. Kraft, P., Yen, Y.C., Stram, D.O., Morrison, J. & Gauderman, W.J. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* **63**, 111-9 (2007).
7. Feitosa, M.F. *et al.* Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. *PLoS One* **13**, e0198166 (2018).
8. 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).
9. de Vries, P.S. *et al.* Multi-Ancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. *Am J Epidemiol* (2019).
10. Bentley, A. *et al.* Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat Genet* **In press**.
11. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
12. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet* **47**, 1282-1293 (2015).
13. Liang, J. *et al.* Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. *PLoS Genet* **13**, e1006728 (2017).
14. Warren, H.R. *et al.* Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet* **49**, 403-415 (2017).
15. Below, J.E. *et al.* Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. *Sci Rep* **6**, 19429 (2016).

16. Prins, B.P. *et al.* Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Sci Rep* **7**, 11008 (2017).
17. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet* **47**, 589-97 (2015).
18. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
19. Aschard, H. *et al.* Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. *Int J Epidemiol* **46**, 894-904 (2017).
20. Cheung, M.C. *et al.* Phospholipid transfer protein activity is associated with inflammatory markers in patients with cardiovascular disease. *Biochim Biophys Acta* **1762**, 131-7 (2006).
21. Dullaart, R.P., Hoogenberg, K., Dikkeschei, B.D. & van Tol, A. Higher plasma lipid transfer protein activities and unfavorable lipoprotein changes in cigarette-smoking men. *Arterioscler Thromb* **14**, 1581-5 (1994).
22. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
23. Gauderman, W.J., Zhang, P., Morrison, J.L. & Lewinger, J.P. Finding novel genes by testing G x E interactions in a genome-wide association study. *Genet Epidemiol* **37**, 603-13 (2013).
24. Kooperberg, C. & Leblanc, M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* **32**, 255-63 (2008).
25. Zhang, P., Lewinger, J.P., Conti, D., Morrison, J.L. & Gauderman, W.J. Detecting Gene-Environment Interactions for a Quantitative Trait in a Genome-Wide Association Study. *Genet Epidemiol* **40**, 394-403 (2016).
26. Laville, V. *et al.* VarExp: estimating variance explained by genome-wide GxE summary statistics. *Bioinformatics* **34**, 3412-3414 (2018).
27. Aschard, H. *et al.* Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am J Hum Genet* **90**, 962-72 (2012).
28. Aschard, H., Zaitlen, N., Lindstrom, S. & Kraft, P. Variation in predictive ability of common genetic variants by established strata: the example of breast cancer and age. *Epidemiology* **26**, 51-8 (2015).
29. 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).
30. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
31. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
32. Lu, Q., Powles, R.L., Wang, Q., He, B.J. & Zhao, H. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS Genet* **12**, e1005947 (2016).
33. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722 (2014).
34. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
35. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).

Figures

Figure 1. Loci identified by the 2df joint test across the four exposures

We assessed the relative performance of the joint 2df test across the four exposures. Panel a) shows overlapping loci for the 2df test across the four exposures GWAS. Panel b) further decomposes these results. It shows for each of the four exposures (highlighted by four different colors), the proportion of loci identified per phenotype divided by the total number of loci identified for that same phenotype.

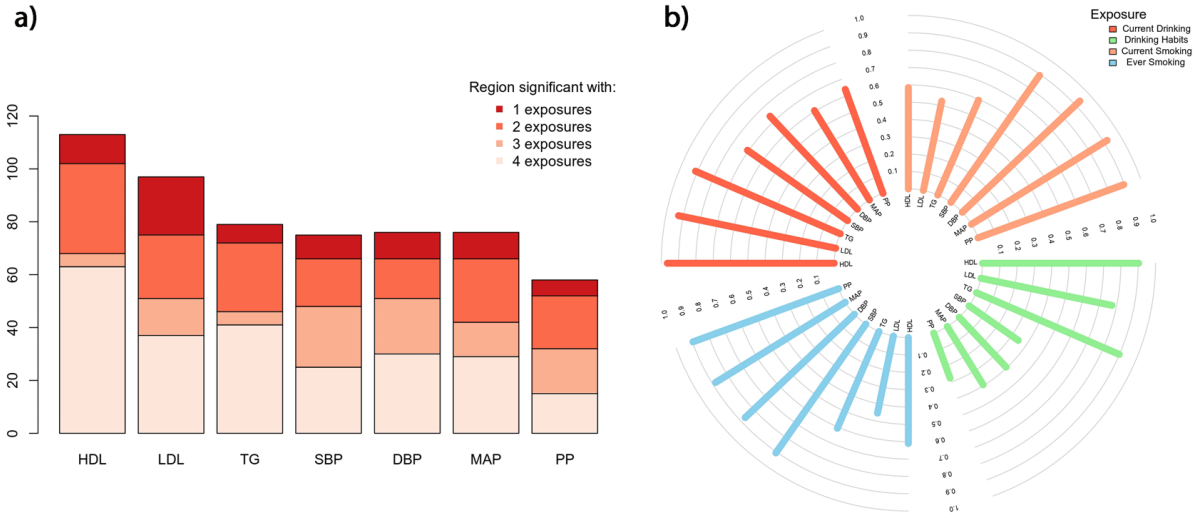


Figure 2. Overlapping associations for the 2df test across ancestries

We derived the overlap in association signal for the joint 2df test of main and interaction effects across the four ancestries: Asian (ASA), African American (AA), European (EA), and Hispanic (HA). In panel a) we focused on loci found at genome-wide significance level. In panel b) we extracted genome-wide significant and independent SNPs per ancestry (*i.e.* reference population) and extracted the *p*-value for those SNPs in other population (*i.e.* the matching population). The barplot shows for each reference population, the proportion of SNPs in the matching population that achieve a *p*-value below 0.05.

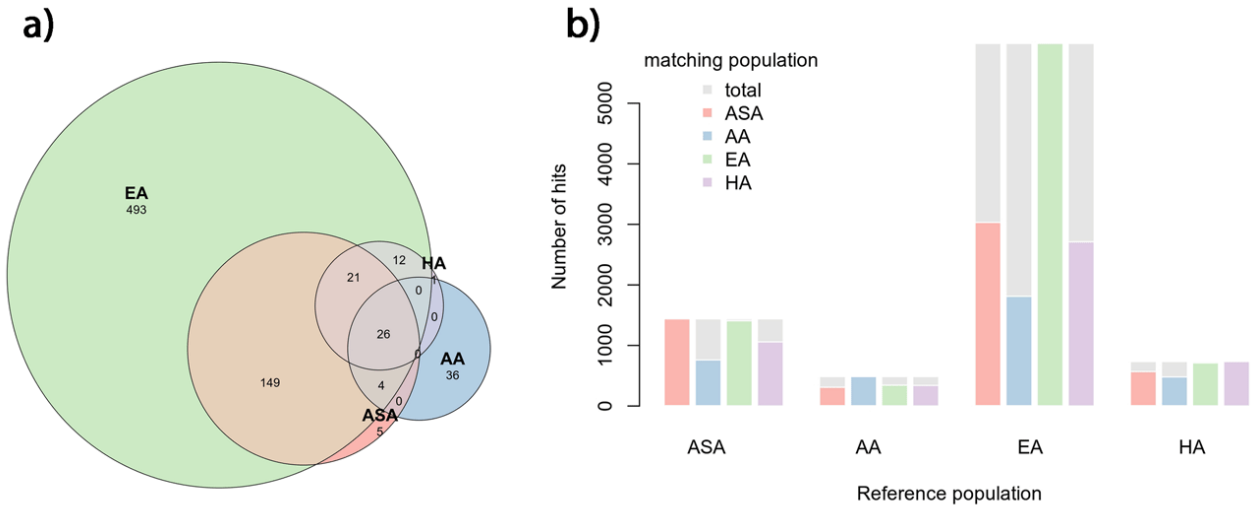


Figure 3. Observed number of hits as a function of sample size

We plotted the number of loci identified at genome-wide significance level as a function of the sample size for both each ancestry-specific analysis and for the trans-ancestry meta-analyses, and for the 1df interaction test (right panels), and the 2df joint test (left panels). Exposures were merged in smoking (bottom panels), and drinking (top panels), and phenotypes are highlighted by different colours.

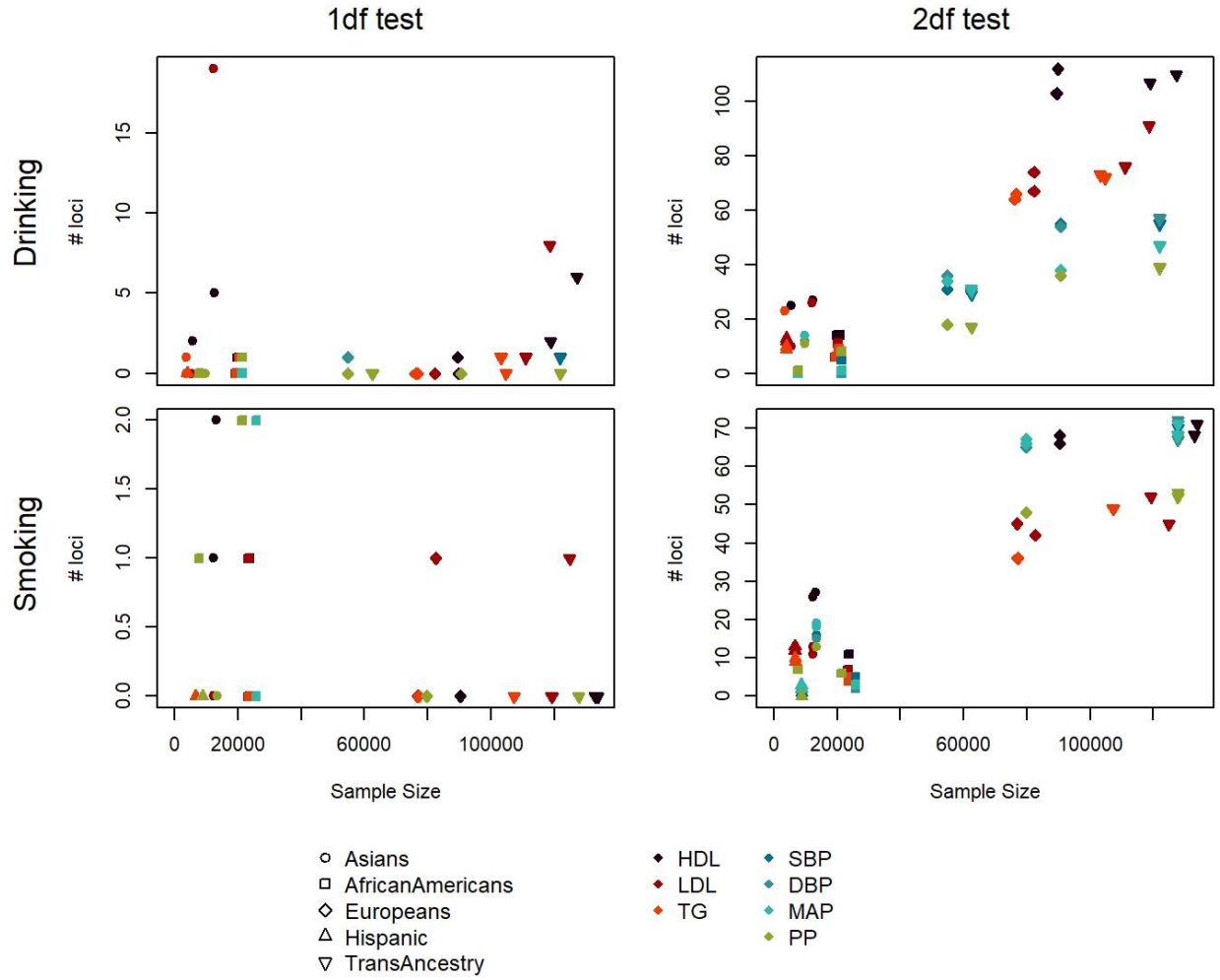


Figure 4. Potential power for 2-step approach

We plotted the proportion of SNPs displaying an interaction p-value below 0.05 in the CHARGE summary statistics after stepwise filtering based on the marginal genetic effect extracted from the CHARGE data (a) and from an independent analysis (b). Under the null, we expect that proportion to be close to 0.05, independently of the threshold consider.

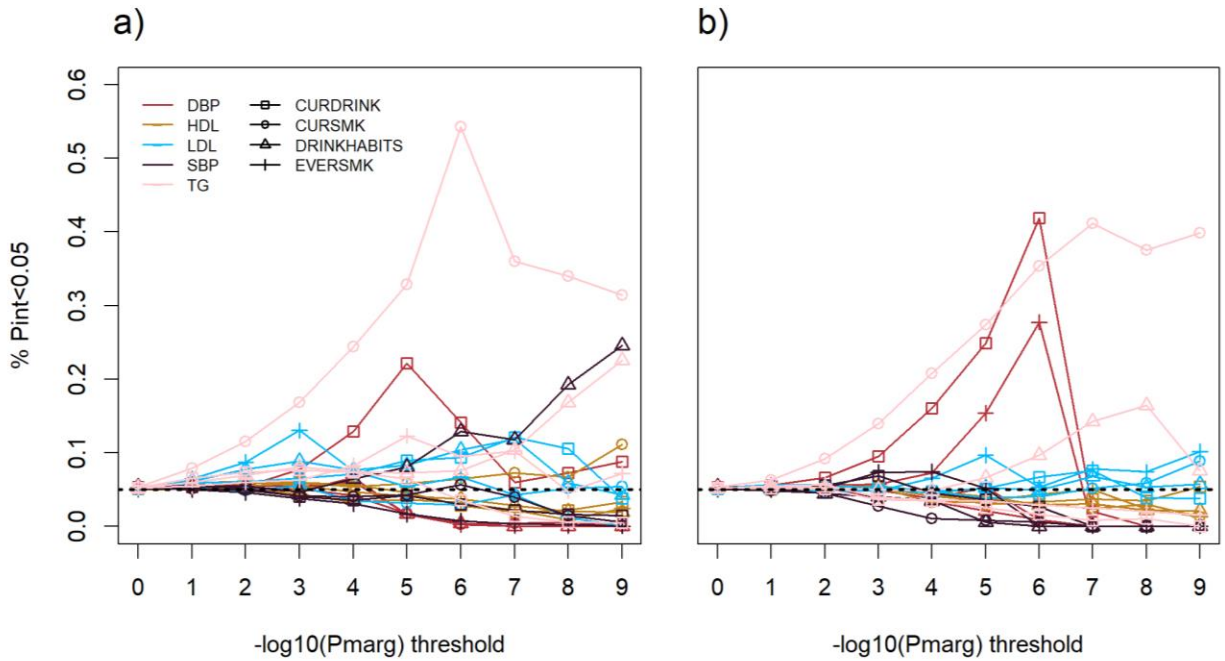
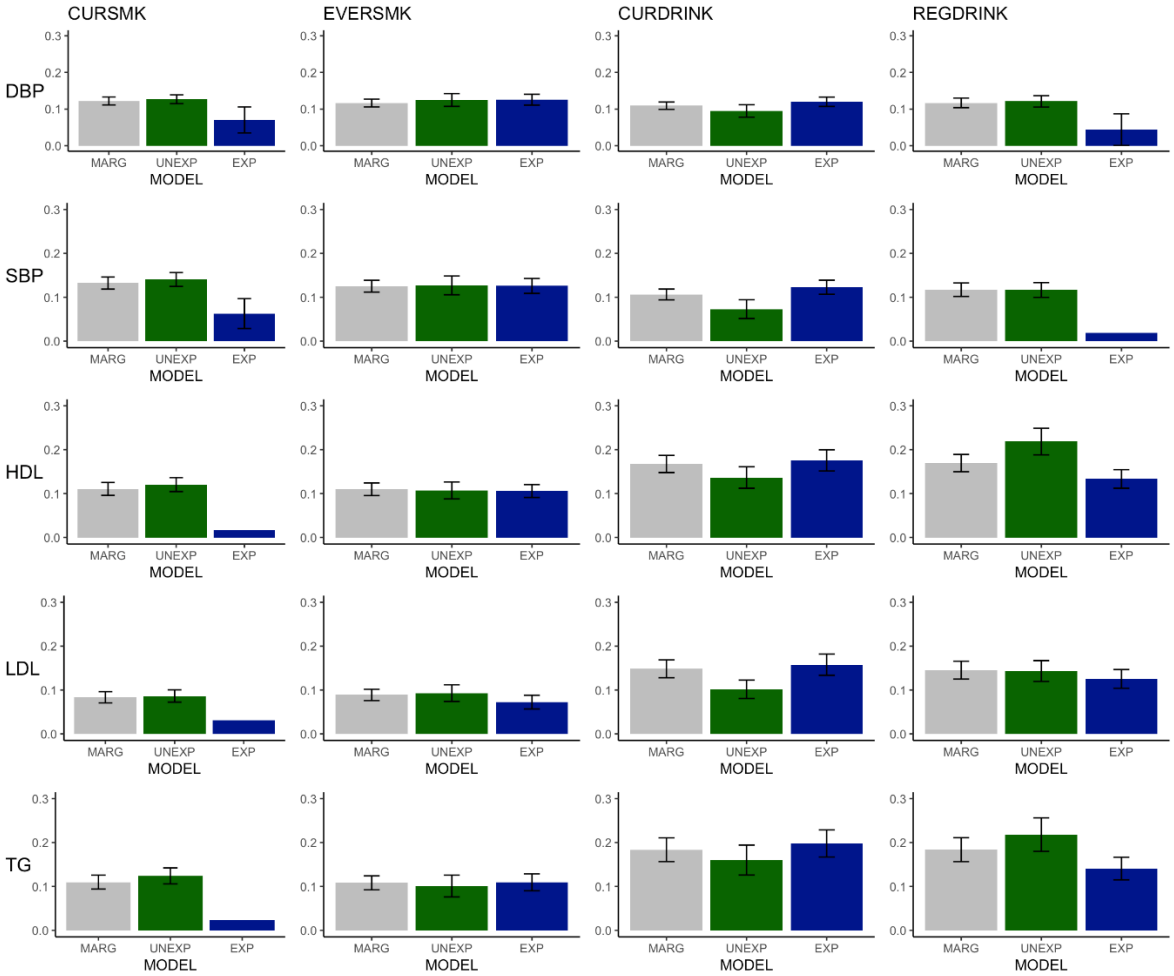


Figure 5. Heritability by exposure group

Heritability of the three lipids and two blood pressure phenotypes (DBP and SBP) derived using the *LDscore* applied to summary statistics from the trans-ancestry meta-analysis. Heritability was derived for all individuals (MARG, grey bar) and for subset of unexposed (UNEXP, green bar) and exposed (EXP, blue bar) individuals. Vertical dark lines represent the 95% confidence intervals.



Tables

Table 1. Summary of GWAS results for 2df joint and 1df interaction.

Outcome		Exposure	#variants	Sample Size[a] (disc)	Sample Size[a](rep)	#hit 2df[b]	#hit 1df[b]
Lipids	HDL	Current Drinking	7,505,310	127,252	231,043	110 *(583)*	6 *(8)*
		Regular Drinking	6,848,811	118,899	217,468	107 *(536)*	2 *(3)*
		Current Smoking	6,306,314	133,508	253,467	68 *(343)*	0 *(0)*
		Ever Smoking	7,269,995	133,816	251,711	71 *(373)*	0 *(0)*
	LDL	Current Drinking	7,448,913	118,654	171,142	91 *(502)*	8 *(8)*
		Regular Drinking	6,834,699	111,093	155,280	76 *(461)*	1 *(1)*
		Current Smoking	6,261,354	125,629	188,109	52 *(262)*	0 *(0)*
		Ever Smoking	7,251,615	125,638	186,230	45 *(165)*	1 *(1)*
	TG	Current Drinking	7,410,534	104,716	221,722	72 *(408)*	0 *(0)*
		Regular Drinking	6,839,760	103,214	210,623	73 *(362)*	1 *(1)*
		Current Smoking	7,122,377	111,900	241,140	49 *(213)*	0 *(0)*
		Ever Smoking	8,438,564	111,909	238,972	49 *(219)*	0 *(0)*
Blood Pressure	SBP	Current Drinking	7,489,960	121,948	426,121	55 *(108)*	1 *(1)*
		Heavy Drinking	10,639,279	62,479	114,058	29 *(44)*	0 *(0)*
		Current Smoking	6,849,695	127,730	474,475	68 *(136)*	0 *(0)*
		Ever Smoking	7,928,860	127,733	458,034	70 *(140)*	0 *(0)*
	DBP	Current Drinking	7,490,269	121,947	426,177	57 *(101)*	0 *(0)*
		Heavy Drinking	10,639,829	62,479	114,111	30 *(41)*	0 *(0)*
		Current Smoking	6,784,799	127,730	474,531	72 *(140)*	0 *(0)*
		Ever Smoking	7,930,829	127,730	458,089	67 *(142)*	0 *(0)*
	MAP	Current Drinking	7,489,903	121,947	426,112	47 *(72)*	0 *(0)*
		Heavy Drinking	10,639,231	62,479	113,287	31 *(47)*	0 *(0)*
		Current Smoking	6,848,964	127,730	474,465	71 *(144)*	0 *(0)*
		Ever Smoking	7,932,503	127,730	458,024	68 *(140)*	0 *(0)*
	PP	Current Drinking	7,489,921	121,947	420,767	39 *(70)*	0 *(0)*
		Heavy Drinking	10,639,279	62,479	114,111	17 *(26)*	0 *(0)*
		Current Smoking	7,934,402	127,730	473,514	53 *(92)*	0 *(0)*
		Ever Smoking	7,934,402	127,730	457,073	52 *(91)*	0 *(0)*

Abbreviation: HDL, High Density Lipoprotein; LDL, Low Density Lipoprotein, TG, Triglycerides; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; MAP, Mean Arterial Pressure; PP, Pulse Pressure; 1df, 1 degree of freedom interaction test; 2df, 2 degrees of freedom joint test; disc, Discovery stage; rep, Replication stage.

[a] Maximum sample size

[b] Independent loci

Table 2. Association signal overlap between the 2df test and previous GWAS of marginal genetic effect.

Phenotype	Overall			CHARGE only, per exposure			
	External GWAS only	*both*	*CHARGE only*	*CurDrink*	*DrinkHabits*	*CurSMK*	*EverSMK*
HDL	2	55	49	46	41	22	22
LDL	2	41	52	45	31	15	14
TG	2	37	35	30	32	15	14
SBP	37	43	29	16	2	21	22
DBP	47	46	26	16	4	22	17
MAP	-	-	-	-	-	-	-
PP	40	8	48	33	15	45	44
all	130	230	239	186	125	140	133