Principal Component Analysis Reduces Collider Bias in Polygenic Score Effect Size Estimation

Nathaniel S. Thomas*[1], Peter Barr[5], Fazil Aliev[1], Mallory Stephenson[3], Sally I-Chun Kuo[4], Grace Chan[6,7], Danielle M. Dick[1,2], Howard J. Edenberg[8,9], Victor Hesselbrock[6], Chella Kamarajan[5], Samuel Kuperman[7], Jessica E. Salvatore[4]

*Corresponding author.

Mailing Address: Psychology, Virginia Commonwealth University, Box 842018, Richmond, VA 23284-2018.

Phone: (804) 828-1193, Fax: 804-828-2237,

Email: thomasns@vcu.edu

[1] Department of Psychology, Virginia Commonwealth University, Richmond, VA

[2] Department of Human & Molecular Genetics, Virginia Commonwealth University, Richmond, Virginia

[3] Virginia Institute for Psychiatric and Behavioral Genetics, Richmond, Virginia

[4] Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, Piscataway, New Jersey

[5] Department of Psychiatry & Behavioral Sciences, SUNY Downstate Health Sciences University, Brooklyn, New Jersey

[6] Department of Psychiatry, University of Connecticut School of Medicine, Farmington, Connecticut

[7] Department of Psychiatry, University of Iowa, Carver College of Medicine, Iowa City, Iowa

[8] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana

[9] Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana

SUGGESTED RUNNING HEAD: PCA Reduces Collider Bias

Abstract

In this study, we test principal component analysis (PCA) of measured confounders as a method to reduce collider bias in polygenic association models. We present results from simulations and application of the method in the Collaborative Study of the Genetics of Alcoholism (COGA) sample with a polygenic score for alcohol problems, DSM-5 alcohol use disorder as the target phenotype, and two collider variables: tobacco use and educational attainment. Simulation results suggest that assumptions regarding the correlation structure and availability of measured confounders are complementary, such that meeting one assumption relaxes the other. Application of the method in COGA shows that PC covariates reduce collider bias when tobacco use is used as the collider variable. Application of this method may improve PRS effect size estimation in some cases by reducing the effect of collider bias, making efficient use of data resources that are available in many studies.

Keywords: Polygenic Scores, Collider Bias, Principal Component Analysis

Introduction

Genome-wide polygenic scoring is a popular method to test for associations between genetic liability and specific phenotypes (Barr et al., 2020; Duncan et al., 2019; Martin et al., 2019), and characterize environmental mediators through which that liability is realized (Domingue et al., 2020; Pasman et al., 2019; Uher & Zwicker, 2017). Oftentimes in polygenic association analyses, covariates are entered into the model to evaluate whether an association with the polygenic risk scores (PRS) of interest is robust to potential confounders; however, the effects of PRS may be biased by the inclusion of heritable covariates when the covariate is influenced by unmeasured confounding variables (Akimova et al., 2021). This bias is generally referred to as collider bias. For example, the estimated effect of a polygenic score for alcohol consumption may be biased in a model that also includes educational attainment as a covariate if the polygenic score for alcohol consumption is correlated with educational attainment. Previous work demonstrates that alcohol consumption is genetically correlated with educational attainment (Sanchez-Roige et al., 2019; Walters et al., 2018; Zhou et al., 2020). Furthermore, educational attainment has been shown to be correlated with a wide variety of other unmeasured variables, such as personality (Mõttus et al., 2017), internalizing behavior, and externalizing behavior (Veldman et al., 2014). There are many possible mechanisms that give rise to the correlations between a polygenic score, the target phenotype from its corresponding discovery GWAS, a heritable environment that might be used as a covariate, and the wide variety of unmeasured variables that can be correlated with the target phenotype and environment. Regardless, they share a common consequence in polygenic association studies: biasing PRS effect size estimation. The partial effect size of the PRS (β) and the variance accounted for by the PRS (R^2) are reduced, while the estimate of the variance accounted for by the PRS and

environment together is inflated (Akimova et al., 2021). If the effect of unmeasured confounding variables can be approximated and accounted for, bias in polygenic associations may be reduced.

Most large cohort studies collect measures on a wide variety of constructs to allow a broad range of research hypotheses to be tested, but few research designs make use of the correlation structure of all of these data in aggregate. In this project, we aim to leverage this common feature of large cohort studies to approximate the effect of unmeasured confounding variables. Principal component analysis (PCA) is a common data reduction technique that aims to explain the maximum amount of variance in a set of variables using as few variables as possible. Under complementary assumptions about (1) the proportion of confounding data that is measured and (2) the correlation structure of the measured and unmeasured confounding data, PCA of measured data may provide some insight into the effects of measured and unmeasured confounders in aggregate. Specifically, the principal components (PCs) are assumed to be constructed from observed confounders that act as proxies for the correlated error structure in the model driven by both measured and unmeasured factors. Under these assumptions, inclusion of the PCs of measured confounders as covariates may reduce the effect of unmeasured confounders. We propose this solution to reduce collider bias when the correlation between PRS and environment is not driven by passive rGE (i.e., correlation between the genotypes that parents transmit to their children that are also associated with the type of rearing environment parents provide). In cases of passive rGE, directly controlling for confounders that drive this correlation is feasible and sufficient.

Our goal in this paper is to test PCA as a method to use information from measured covariates in order to construct principal components that reduce collider bias in polygenic association studies. We present results from a simulated implementation of the method alongside

complementary application in observed data. Specifically, we examine two complementary assumptions required for phenotypic PC covariates to reduce collider bias related to the proportion of confounding data that is measured and the correlation structure of the measured and unmeasured confounding data. We provide evidence that these assumptions are complementary, such that meeting one assumption relaxes the requirements for the other. We further provide two examples of applications of the method in observed data to demonstrate the utility of this method to reduce collider bias in polygenic association studies, which deflates the partial effect size of the PRS (ß) and the variance accounted for by the PRS (R^2), while inflating the combined R^2 of the PRS and heritable collider. Finally, we provide some suggestions about the practical utility of this method and directions for future applications.

The target phenotype of the applied analysis is DSM-5 alcohol use disorder clinical criterion counts (AUD Sx). Collider bias occurs if a covariate is an outcome of two different variables; for example, if the covariate is associated with (1) the PRS and (2) an unmeasured confounding variable. We examine tobacco use and educational attainment as heritable collider variables. We selected these two variables as heritable collider variables for three reasons. First, tobacco use and educational attainment are genetically correlated with AUD , suggesting that a PRS for AUD may be associated with tobacco use and educational attainment (Kranzler et al., 2019; Walters et al., 2018; Zhou et al., 2020). Second, preliminary results indicate differing strengths of correlation with polygenic liability for AUD, providing a useful range of circumstances for examining the method. Third, tobacco use (Cheng & Furnham, 2021; Green et al., 2018) and educational attainment (Esch et al., 2014; Krapohl et al., 2014) are endogenous to a wide variety of other predictor variables. Given the wide variety of variables that predict tobacco use and educational attainment (Cheng & Furnham, 2021; Esch et al., 2014; Green et al.,

2018; Krapohl et al., 2014), we expect that tobacco use and educational attainment are associated with unmeasured confounding variables. Furthermore, we expect that an array of measured variables will provide indirect insight into a wide variety of constructs beyond what is explicitly measured, provided that the observed confounders are proxies for the correlated error structure in the model driven by unmeasured factors. For example, if an unmeasured personality construct happens to be correlated with the measured variables included in the phenotypic PCs, the phenotypic PCs would index some amount of variance in this unmeasured construct, proportional to the correlations between measured variables and the unmeasured personality construct.

Methods

This study uses both simulated data and observed data from the Collaborative Study of the Genetics of Alcoholism (COGA) sample. Details for each aspect of the study are described below.

Simulation

We conducted a simulation study in R (R Core Team, 2017) to test principal component analysis (PCA) of a series of measured confounders as a correction for collider bias in tests of polygenic association. The R script used to conduct this simulation is available on GitHub (https://github.com/thomasns0/PCA_Collider.git). We sampled data from a model which tested different values for three parameters of interest: the correlation structure of the confounding data, the effect of the PRS on the heritable environment, and the proportion of confounding variables that was available for use in the PCA correction. We tested multiple values for these parameters to provide information about the empirical requirements for this method to provide adequate

correction for collider bias at different levels of gene-environment correlation (rGE). The sample size was 1000 in all simulations.

First, we sampled 100 variables from a multivariate normal distribution with mean set to 0. We defined the correlation structure of the multivariate normal distribution by drawing individual values from a uniform distribution for each cell of the symmetric correlation matrix. The range of the uniform distribution varied across simulation iterations to model confounders that are correlated at a range of different levels (0.05 – 0.1, 0.2 – 0.3, 0.5 – 0.6, 0.8 – 0.9). We chose to use a set of 100 confounding variables in order to be able to test the incremental difference in effect size correction that results from decreasing the proportion of confounders in the correction PC. Starting with 100 variables allowed us to test random sets of cofounders ranging from 10 variables (10% of the confounders) to 100 variables (100%) at intervals of 1% in order to determine how PC covariates perform when only a subset of the total confounding data is measured. This provides a more detailed picture of how the method can work in practice. A PRS variable was generated from a standard normal distribution. We generated a heritable collider environment variable as a function of the PRS and a single PC, which was derived from all 100 confounding variables. The effect of the polygenic risk score on the heritable environment (rGE) was fixed at different values in different iterations of the simulation (0, 0.1, 0.2, 0.3, 0.4, 0.5). We generated a target phenotype as a function of the PRS, the first PC of 100 confounders, and the heritable collider environment. We set the true value of the effect of PRS on the target phenotype to 0.1 . This effect size magnitude is comparable to previous studies that use PRS to predict complex traits in observed data sets (Barr et al., 2020). Observed datasets will generally not include all relevant confounding variables. Therefore, we calculated a second PC from randomly selected subsets of the confounding variable, ranging from 10% (10 variables) to

100% (100 variables) at intervals of 1% in order to model the effectiveness of the method under different assumptions about the proportion of confounding data that is measured. We fixed other model parameters across all simulation iterations, as shown in Figure 1. We extracted estimates of the effect of PRS on the target phenotype from the following models:

Model A1. Target Phenotype ~ PRS

Model A2. Target Phenotype ~ PRS + Environment

Model A3. Target Phenotype ~ PRS + Environment + Incomplete Phenotypic PC

Model A4. Target Phenotype ~ PRS + Environment + Complete Phenotypic PC

We repeated this procedure 1250 times for each combination of parameters. New simulated data was generated at each iteration of the simulation. Estimates were plotted using the ggplot2 (Wickham, 2009) package in R. Note that in this simulation we refer to the collider variable as an environment. In the application of this method to real data, we use the more general term "collider variable" to identify the covariate that may induce collider bias.

Application in Real Data

Sample

Participants came from the Collaborative Study on the Genetics of Alcoholism (COGA), a diverse, multi-site, family-based study whose objective is to identify genetic variants associated with AUD and related psychiatric disorders (Begleiter, 1995; Bucholz et al., 2017; Reich et al., 1998). Probands were identified through alcohol treatment centers across seven sites in the United States. Probands along with their families were invited to participate if the family was sufficiently large (usually sibships greater than 3 with parents available), with two or more members in the COGA catchment area. Comparison families were recruited from the same

communities. The Institutional Review Board at all data collection sites approved the study, and written consent was obtained from all participants.

In the present study, we focused on all COGA participants of European ancestry (EA) with genome-wide association (GWAS) data. The analytic sample of the current study varied between two heritable collider variables of interest: tobacco use and educational attainment, described in the measures section below. In our analysis of tobacco use, the total sample size was 7,270, the mean age was 37.67 years (SD = 14.44), and 53% of the sample was female. In our analysis of educational attainment, the total sample size was 7,286, the mean age was 37.69 years (SD = 14.45), and 53% of the sample was female.

Measures

Alcohol Use Disorder Symptoms. Maximum lifetime alcohol use disorder criterion count in any interview (AUD Sx) were assessed based on Diagnostic and Statistical Manual of Mental Disorder (5^{th} edition; American Psychiatric Association, 2013) using the reliable and validated SSAGA or adolescent version of the SSAGA interviews (Bucholz et al., 1994; Hesselbrock et al., 1999; Kuperman et al., 2013).

Tobacco Use (TOB). We operationally defined *tobacco use* (TOB) as someone having smoked a total of 100 cigarettes over lifetime and having a >0 score on the Fagerström Test for Nicotine Dependence (FTND; Heatherton et al., 1991), and no tobacco use was defined as someone with <100 lifetime cigarettes with a zero on FTND. Because FTND was not administered during the early phase of COGA data collection, a separate yes/no question from the SSAGA ("Have you ever smoked cigarettes daily for a month or more?") was substituted to define tobacco use for those without the FTND data.

Educational Attainment (EDU). We used participants' self-reported highest level of education (EDU). Participants responded to the question "What is the highest grade in school you completed?" Scores were converted to the number of years typically required to complete that level of education and ranged from 0 to 17 years (primary or secondary school = actual year; technical school/ 1 year college = 13 years; 2 years college = 14 years; 3 years college = 15 years; 4 years college = 16 years; any graduate degree = 17 years).

Genotyping and Ancestry PCs. Participants' DNA samples were genotyped using the Illumina Human1M array (Illumina, San Diego, CA), the Illumina Human OmniExpress 12V1 array (Illumina), the Illumina 2.5M array (Illumina) or the Smokescreen genotyping array (Biorealm LLC, Walnut, CA). A full description of data processing, quality control, and imputation is available elsewhere (Lai et al., 2019). Briefly, data were imputed to Haplotype Reference Consortium (HRC). Single nucleotide polymorphisms (SNPs) with a genotyping rate < 0.95, that violated Hardy-Weinberg equilibrium ($p < 10^{-6}$), or had minor allele frequency (MAF) < 0.01 were excluded from analysis. SNPrelate (Zheng et al., 2012) was used to estimate principal components from GWAS data. These principal components are distinct from the phenotypic principal components and are referred to as ancestry PCs throughout the rest of the manuscript.

Alcohol problems polygenic scores (PRS). Genetic risk for alcohol problems was indexed using genome-wide polygenic scores (PRS), which are an aggregate measure of the number of risk alleles individuals carry, weighted by effect sizes from GWAS summary statistics (Wray et al., 2014). We calculated PRS using PRS-CS "auto" (Ge et al., 2019), which employs a Bayesian regression and continuous shrinkage method to correct for the non-independence among nearby SNPs in the genome (i.e., linkage disequilibrium, or LD). We derived the alcohol

problems PRS using meta-analyzed GWAS weights (detailed in Barr et al., 2020) from the EA subset of the Psychiatric Genomic Consortium's (PGC) GWAS of alcohol dependence (Walters et al., 2018) and a GWAS of the problem subscale from the Alcohol Use Disorders Identification Test (AUDIT-P) in UK Biobank's (Sanchez-Roige et al., 2019). Higher polygenic scores indicated higher polygenic liability for alcohol problems.

Candidate confounders. Candidate confounding data were selected as part of a larger study on marital status and substance use (Thomas et al., 2021). The candidate confounding data available from this study include typical covariates such as sex, generational cohort, and age, as well as measures of externalizing and internalizing behavior, romantic relationship behaviors, parental marital quality, and parental alcohol use. A summary of the candidate confounding variables that were considered and retained in the PCA for TOB and EDU are available in Supplemental Table I, Supplemental Table II, and Supplemental Table III (TOB) and Supplemental Table V, Supplemental Table VI, and Supplemental Table VII (EDU). Lifetime measures were calculated where multiple observations over time were available by taking the maximum value, or, for age of onset variables, the minimum value.

Analyses

We established a pipeline in R for generating phenotypic PCs from the candidate confounding variables. First, we calculated zero-order correlations between the target phenotype (DSM-5 AUD Sx) and the heritable environment (EDU/TOB) using the hetcor function from the polycor package (Fox, 2019). We calculated Pearson correlations for pairs of continuous variables, polychoric correlations for pairs of binary variables, and polyserial correlations for pairs where one variable was continuous and the other was binary. Confounder variables that

were associated with either the target phenotype or the heritable collider variable at $p<0.05$ were retained for further analysis.

We used K-nearest-neighbors (KNN) imputation to account for missing confounder data using the kNN function from the VIM package (Kowarik & Templ, 2016) in R. We conducted KNN imputation with 5 neighbors, mean aggregation (mean option) for continuous variables, and modal aggregation (maxCat option) for binary variables. We used all variables in the set of confounder variables to identify neighbors. Next, we calculated a mixed correlation matrix of Pearson, polychoric, and polyserial correlations from the imputed phenotype data. Variables were removed if they caused errors in the correlation matrix of imputed phenotype data. We calculated eigenvectors from this correlation matrix using eigen function in R. We post multiplied the imputed data by the eigenvectors to generate principal components and used parallel analysis to determine the number of principal components to retain as covariates using the paran function from the paran package (Dinno, 2018) in R. We conducted the parallel analysis using the mixed correlation matrix, n set to the number of rows in the imputed data, and 1000 iterations. We constructed PCs to be orthogonal to the PRS by extracting the residual from the regression of each PC on the polygenic score (resid(PRS ~ PC)). The residuals from this series of linear models were used as covariates in subsequent analyses and are referred to as "phenotypic PCs" throughout the rest of this text.

We then fit a series of linear models to test the impact of the phenotypic PCs on the estimate of the effect of PRS on AUD Sx. Models were tested with 10 ancestry PCs for a total of 3 linear models for each collider variable.

Model B1. AUD Sx ~ PRS + Ancestry PCs
Model B2. AUD Sx ~ PRS + Ancestry PCs + Collider Variable
Model B3. AUD Sx ~ PRS + Ancestry PCs + Collider Variable + Phenotypic PCs

The presence of collider bias is inferred from the decrease in the partial effect size of the PRS in the presence of the heritable collider variable; for example, if the PRS effect decreases from model B1 to B2. A correction for this bias is identified when the partial effect size of the PRS increases in the presence of PC covariates; for example, if the PRS effect increases from model B2 to B3. We present change in β and R^2 between models in their original scale and as a percentage of the model B1 effect size.

Results

Simulation

Results from the most extreme parameters tested (rGE = 0.1 / 0.5; Confounder Correlations ~ 0.05-0.1 / 0.8 - 0.9) are presented in Figures 2 and 3. The left panel displays a series of regression lines that summarize the relationship between the PRS beta and the proportion of confounding data that was included in the Incomplete PC. The regression line has slope equal to 0 where the Incomplete PC is not included in the analysis. The right panel displays the distribution of PRS betas from each model. The complete results, which include the intermediate levels of rGE, demonstrate similar patterns as those presented here and are available on GitHub (https://github.com/thomasns0/PCA_Collider.git). We present results throughout this section as the average percentage of change from the true value of the PRS effect (i.e. (Model A1 – Model A2) / 0.1).

Five noteworthy patterns emerge from the comparison of these extreme conditions. First, estimates from model A1 (~PRS) are inflated relative to the true effect size. Inflation of the model A1 effect size increases when the correlation between PRS and environment (rGE) is higher. The A1 PRS effect is inflated approximately 150% in the rGE=0.1 conditions and approximately 350% in the rGE=0.5 conditions.

Second, the PRS effect size decreases more in the presence of the environmental covariate when rGE is higher. Note that the estimates from simulation model A2 (~ PRS+Env; shown in red) are further from the true value of 0.1 in both rGE = 0.5 conditions. This replicates previous results reported in Akimova et al. (2021). The model A2 effect size also decreases more when confounder correlations are smaller. The model A2 effect is deflated by 92% with rGE=0.1 and confounder correlations between 0.05-0.1, 511% with rGE=0.5 and confounder correlations between 0.05-0.1, 146% with rGE=0.1 and confounder correlations between 0.8-0.9, and 736% with rGE=0.5 and confounder correlations between 0.8-0.9.

Third, PRS effect size estimates from the model that uses the incomplete phenotypic PC as a correction for collider bias approach the true value of 0.1 as the proportion of complete data included in the PC increases. Note that the estimates from simulation model A3 (~PRS+Env+IncompletePC*;* shown in green) are summarized best by a positive slope that approaches 0.1. This suggests that PCA will provide a better correction for collider bias when more of the confounding data is measured. The relationship between the proportion of confounding data that is measured and the magnitude of the correction varies as a function of the magnitude of the rGE parameter (the magnitude of collider bias). Here, we present average change in Model A3 in four bins for the proportion of confounding data that is measured in the correction: (1) 10% - 25%, (2) 26% - 50%, (3) 51% - 75%, and (4) 76% - 99%.When rGE = 0.1 and confounder correlations range between 0.05-0.1 the model A3 PRS effect increases by 5%, 14%, 25%, and 37% for bins 1 through 4, respectively. When rGE = 0.5 and confounder correlations range between 0.05-0.1 the model PRS A3 effect increases by 25%, 75%, 146%, and 220% for bins 1 through 4.

Fourth, the intercept of the simulation model A3 (~PRS+Env+Incomplete Phenotypic PC) regression line is higher when confounder correlations are higher. Again, we present average change in Model A3 in four bins for the proportion of confounding data that is measured in the correction: (1) 10% - 25%, (2) 26% - 50%, (3) 51% - 75%, and (4) 76% - 99%. When rGE=0.1 and confounder correlations range between 0.8-0.9 the model A3 PRS effect increases by 58%, 78%, 89%, and 94%. When rGE=0.5 and confounder correlations range between 0.8-0.9 the model A3 effect increases by 253%, 367%, 435%, and 472%. The correction performs better with less of the confounding data included in the PCA relative to the results reported above with confounder correlations between 0.05-0.1. This suggests that required assumptions about the proportion of confounding data that is measured and the correlation structure of the confounding data are complementary. If a larger proportion of the confounders is measured, the required assumptions about the correlation structure of the confounders are relaxed. If the confounders are highly correlated, the assumptions about the proportion of confounders that are measured is relaxed. This pattern aligns with derivations in Akimova et al. (2021) which indicate that confounders are less influential in the collider bias expression when rGE is smaller. In this simulation, the PCA corrected model outperforms an uncorrected model, even with as few as 10% of the confounders measured, if the confounders are highly correlated.

Fifth, the dispersion of the simulated sampling distribution of the PRS effect size, depicted in the violin plots in Figure 2 and Figure 3, vary between models. Most notably, the estimates from the fully corrected simulation model A4 (~PRS+Env+Complete Phenotypic PC*)* demonstrate lower variance than the estimates from simulation model A1 (~PRS). This suggests that correcting PRS effect size estimates via PCA in this way may increase power to detect small PRS effects by reducing the standard deviation of the sampling distribution (standard error) of

the estimate. We note that this finding may be specific to ordinary least squares regression models with a continuous outcome variable.

In summary, modeling the first PC of measured confounders as a covariate recovers the PRS effect size estimate under reasonable assumptions about the proportion of the confounding data that is measured and the correlation structure of the confounding data. These assumptions are complementary, such that meeting one assumption more robustly relaxes the other assumption. Required assumptions become stricter as rGE (and the magnitude of bias) increases.

Application in Observed Data

The following section presents results from application of PCA as a correction for collider bias in the COGA sample, examining tobacco use (TOB) and educational attainment (EDU) as the two heritable collider variables. Descriptive statistics for the analytic sample are reported in Table I.

Tobacco Use (TOB) Results

8 phenotypic PCs were retained in the parallel analysis. Eigenvalues and a Scree plot of the retained components are available in Supplemental Table IV and Supplemental Figure 1, respectively. Change R^2 values for the PRS from each model are presented in Table II. Standardized coefficient estimates for PRS and TOB from each model are presented in Table III. Figure 4 displays change in the standardized coefficient estimates for PRS.

When TOB was added to the model, the standardized coefficient estimates for the PRS decreased by 0.041 (28%) (B1 to B2). PRS Change R^2 decreased by 0.010 (50%) (B1 to B2). When the phenotypic PCs were added to the model, the standardized coefficient estimate for the PRS increased by 0.029 (20%) (B2 to B3). PRS Change R^2 increased by 0.007 (35%) (B2 to B3).

These results suggest that the phenotypic PCs provide a modest correction for the collider bias that results from the correlation between PRS and TOB (r = 0.140). The decrease in PRS effect from models B1 to B2 suggests some magnitude of collider bias may be present. The subsequent increase in PRS effect from models B2 to B3 represents a correction for this bias under the assumption that the phenotypic PCs are proxies for the correlated error structure driven by both observed and unobserved factors.

Educational Attainment (EDU) Results

9 phenotypic PCs were retained in the parallel analysis. Eigenvalues and a Scree plot of the retained components are available in Supplemental Table VIII and Supplemental Figure 2. Change R^2 values for the PRS from each model are presented in Table IV. Standardized coefficient estimates for PRS and EDU from each model are presented in Table V. Figure 4 displays change in the standardized coefficient estimates for PRS.

When EDU was added to the model, the standardized coefficient estimates for the PRS decreased by 0.004 (3%) (B1 to B2). PRS Change R^2 decreased by 0.001 (5%) (B1 to B2). When the phenotypic PCs were added to the model, the standardized coefficient estimate for the PRS increased by 0.005 (3%) (B2 to B3). PRS Change R^2 increased by 0.001 (5%) (B2 to B3). Although these results demonstrate the same general pattern as the example above using TOB as the heritable environment, the small magnitude of beta and change R^2 suggest that EDU does not induce a substantial collider bias in this example, possibly due to the modest correlation between PRS and EDU (r = -0.051). Alternatively, confounder correlations with different directions of effect may reduce the magnitude of the observed bias. Importantly, the phenotypic PCs do not appear to increase the PRS effect size in the absence of an indication of robust collider bias.

Discussion

Polygenic association analyses often use phenotypic covariates to test whether the PRS of interest is robust to potential confounders, but the effects of PRS may be biased by the inclusion of heritable covariates when the covariate is influenced by unmeasured confounding variables. In this work, we conducted a simulation to test PCA as a potential correction for this bias and subsequently applied the method in observed data. The results of the simulation suggest that using phenotypic PCs as covariates may correct or reduce collider bias under complementary assumptions about the proportion of confounding data that is measured and the correlation structure of the confounding data. When a larger proportion of confounding data is measured, the assumptions about the correlation structure of the confounding data are relaxed. When the correlations between confounding variables are higher, the assumptions about the proportion of confounding data that needs to be measured are relaxed.

We then examined the effect of a PRS for alcohol problems on alcohol use disorder clinical criteria in our application of the method in observed data. We tested two heritable environments as sources of collider bias: tobacco use (TOB) and educational attainment (EDU). Inclusion of the phenotypic PCs in the TOB models increased the PRS beta and change R^2 modestly. The same pattern was observed in the EDU models, but the differences were very modest. This likely reflects the difference in correlation between the PRS and heritable environment, which influences the magnitude of collider bias. The correlation of -0.051 between PRS and EDU was likely too small to suppress the PRS effect and induce a measurable bias. On the other hand, the correlation between PRS and TOB ($r = 0.140$) was high enough to cause a detectable decrease in PRS effect and subsequent correction via PCA.

Our results should be considered in the context of the limitations of the study. Foremost, this method assumes that the observed confounders included in the PCA adequately capture the underlying mechanism of collider bias. If observed confounders do not account for this, either directly or as proxies of unmeasured confounders, bias due to these factors may remain. The observed changes in PRS beta and R-squared in the applied analysis in COGA were modest, and in all examples, the 95% confidence intervals for PRS beta estimates overlapped. This may be attributable to COGA being a high-risk sample with participants from extended families enriched for AUD. Accordingly, thresholding in phenotype and genotype may have reduced observed associations. Additionally, we generated a normally distributed outcome variable in our simulation and modeled AUD-Sx as a continuous variable with ordinary least squares regression in the applied examples presented here. Thus, this approach to addressing collider bias may not extend to outcome variables with other distributions. An extension of our simulation pipeline to accommodate logistic regression for binary outcomes performed poorly. The addition of PCs increased the variance of the PRS estimate slightly and the distribution of corrected PRS effect sizes was not reliably centered on the true value. The complete results of these simulations are available on the GitHub page associated with this work (https://github.com/thomasns0/PCA_Collider.git).

Furthermore, some parameterizations of our simulation with a normally distributed outcome variable did not demonstrate the expected increase in PRS change R-squared in models that include the PC correction. This unanticipated result likely reflects a ceiling effect in the estimation of R-squared; in these simulations, Environment and PC accounted for large amounts of variance on their own (Total model R-squared: Supplemental Figure 3; PRS change R-squared: Supplemental Figure 4). Conclusions from our series of simulations should be limited to

the estimation of PRS effect sizes, rather than parameters with an explicit boundary such as R-squared. Finally, our approach to PCA uses single imputation via the K-nearest neighbors algorithm, rather than multiple imputation. We chose single imputation to improve the accessibility and flexibility of the method, but recognize that performance may be improved by the use of multiple imputation with pooled results.

Future directions include replication of these results in general population samples and with other complex phenotypes. Additionally, future work may investigate the application of this correction to polygenic gene-by-environment interaction (GxE) analyses. Spurious GxE effects may be detected if the effect of the heritable environment on the target phenotype is moderated by unmeasured confounding variables (Akimova et al., 2021; Keller, 2014). Phenotypic PC covariates could be applied to correct these spurious results by the approach recommended in Keller et al. (2014), computing interactions between each phenotypic PC and the heritable environment of interest. Adequate correction of collider bias in polygenic association analyses may improve estimates of the influence of genetic risk on complex phenotypes in the presence of heritable covariates across a wide range of research designs.

In summary, principal component analysis reduces collider bias in polygenic risk score effect size estimation under particular statistical assumptions about missingness and correlations in the confounding data. Although the changes in beta and R^2 we observed here were modest, PRS effect sizes for complex phenotypes in general are usually small. Correlations between PRS and heritable environments are likely to increase as discovery GWAS become larger and PRS become more powerful. The magnitude of collider bias and the importance of adequately accounting for this bias will increase in turn. Efficient use of existing data resources should be treated as a high priority in complex trait genetics, where data collection is costly and polygenic

effect sizes are often small. Application of this method may improve PRS effect size estimation in some cases by reducing the effect of collider bias, making efficient use of data resources that are immediately available in many studies.

Figure 1. Diagram of the simulation model

.

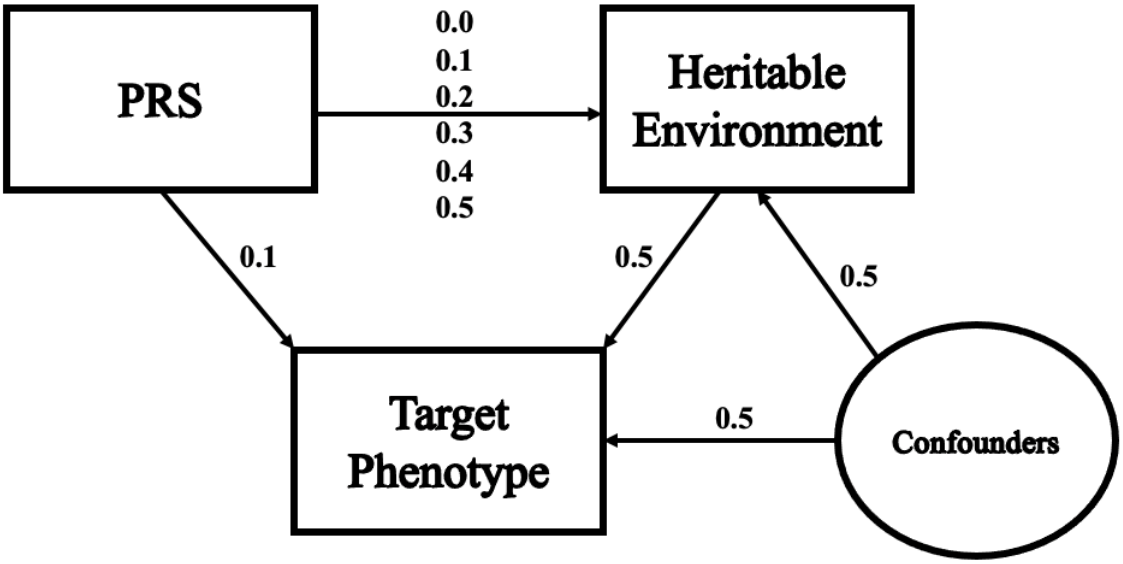

Figure 1 caption

The effect of PRS on the heritable covariate is varied between 0.0 to 0.5 at intervals of 0.1. The effect of PRS on the target phenotype is set to 0.1. The effect of the first PC of all confounders on the heritable covariate and the target phenotype is set to 0.5. The effect of the heritable covariate on the target phenotype is also set to 0.5.

Figure 2. Simulation results for rGE = 0.1 and rGE = 0.5 with confounder correlations ranging between 0.05 and 0.1

.

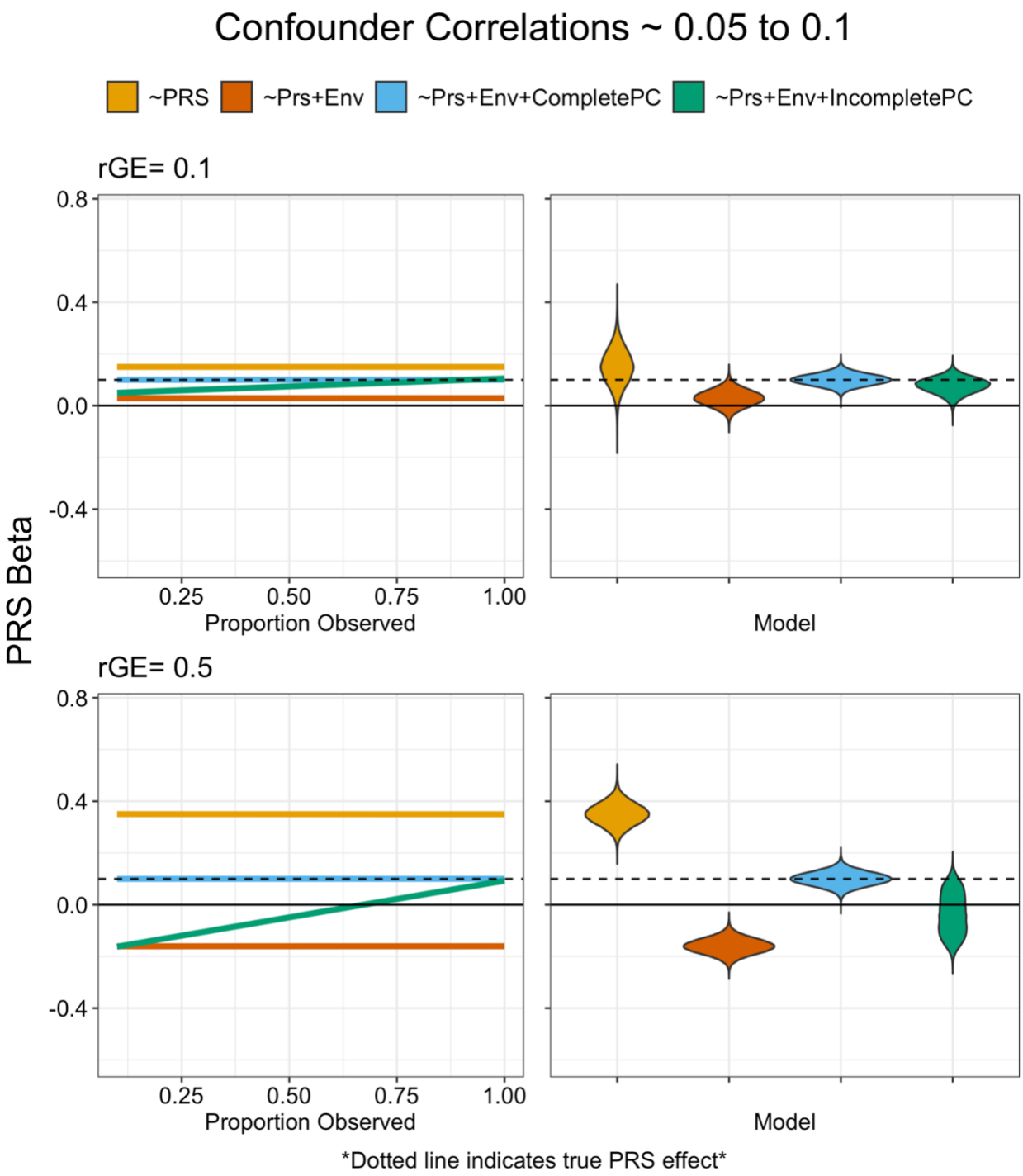

Figure 2 caption

PRS = polygenic risk score; Env = heritable covariate; PCs = principal components; rGE = gene-environment correlation

The magnitude of collider bias is larger when rGE is higher. The PRS beta estimate that is corrected by the incomplete PC approaches the true value of 0.1 as the proportion of confounders included in the PC increases.

Figure 3. Simulation results for rGE = 0.1 and rGE = 0.5 with confounder correlations ranging between 0.8 and 0.9.

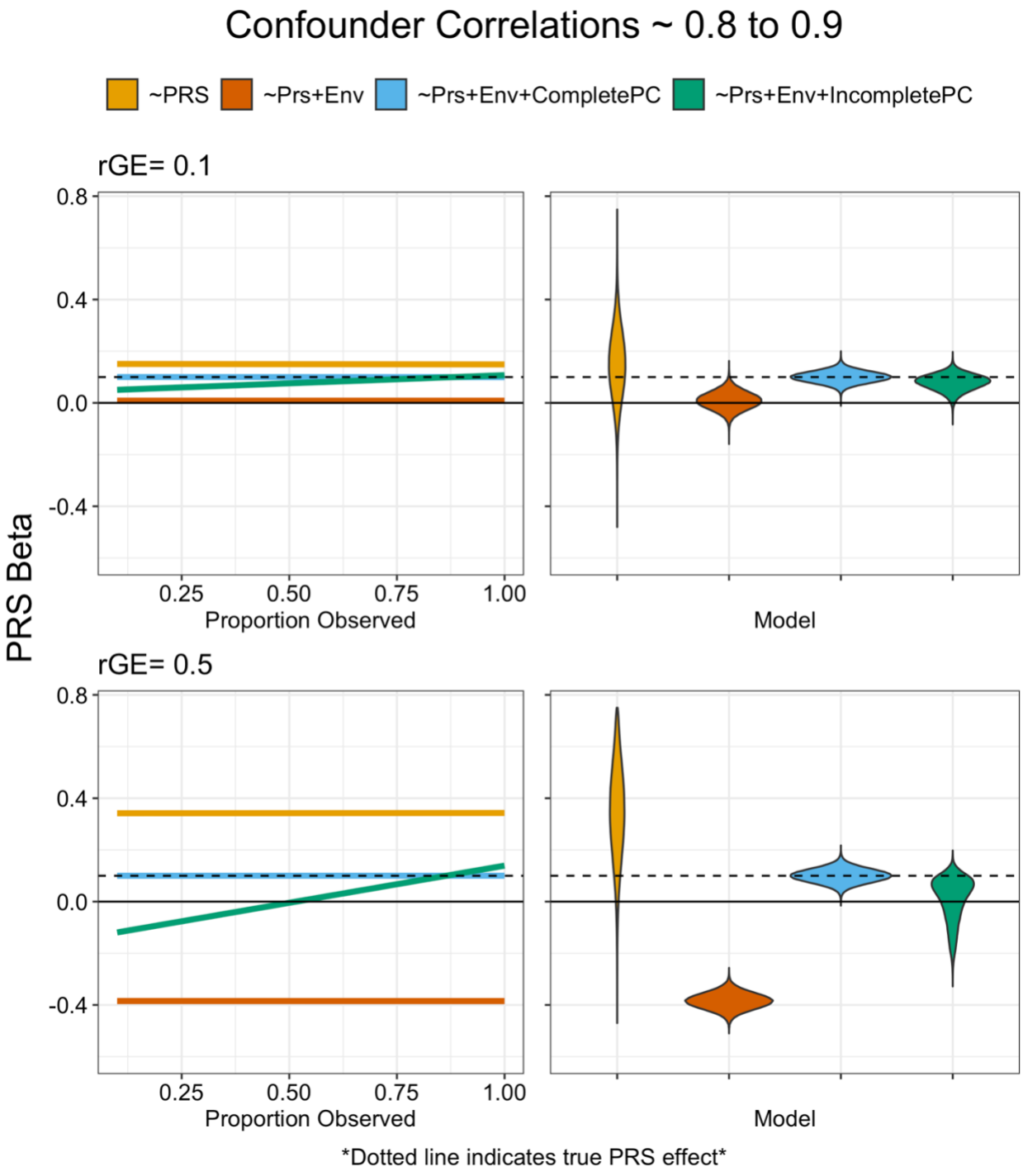

Figure 3 caption

PRS = polygenic risk score; Env = heritable covariate; PCs = principal components; rGE = gene-environment correlation

The magnitude of collider bias is larger when rGE is higher. The PRS beta estimate that is corrected by the incomplete PC approaches the true value of 0.1 as the proportion of confounders included in the PC increases. Relative to Figure 2 where confounder correlations are lower, the corrected beta is closer to the true value of 0.1 with a lower proportion of the confounding data.

Figure 4. Change in Alcohol Problems PRS Beta across models with TOB/EDU environment.

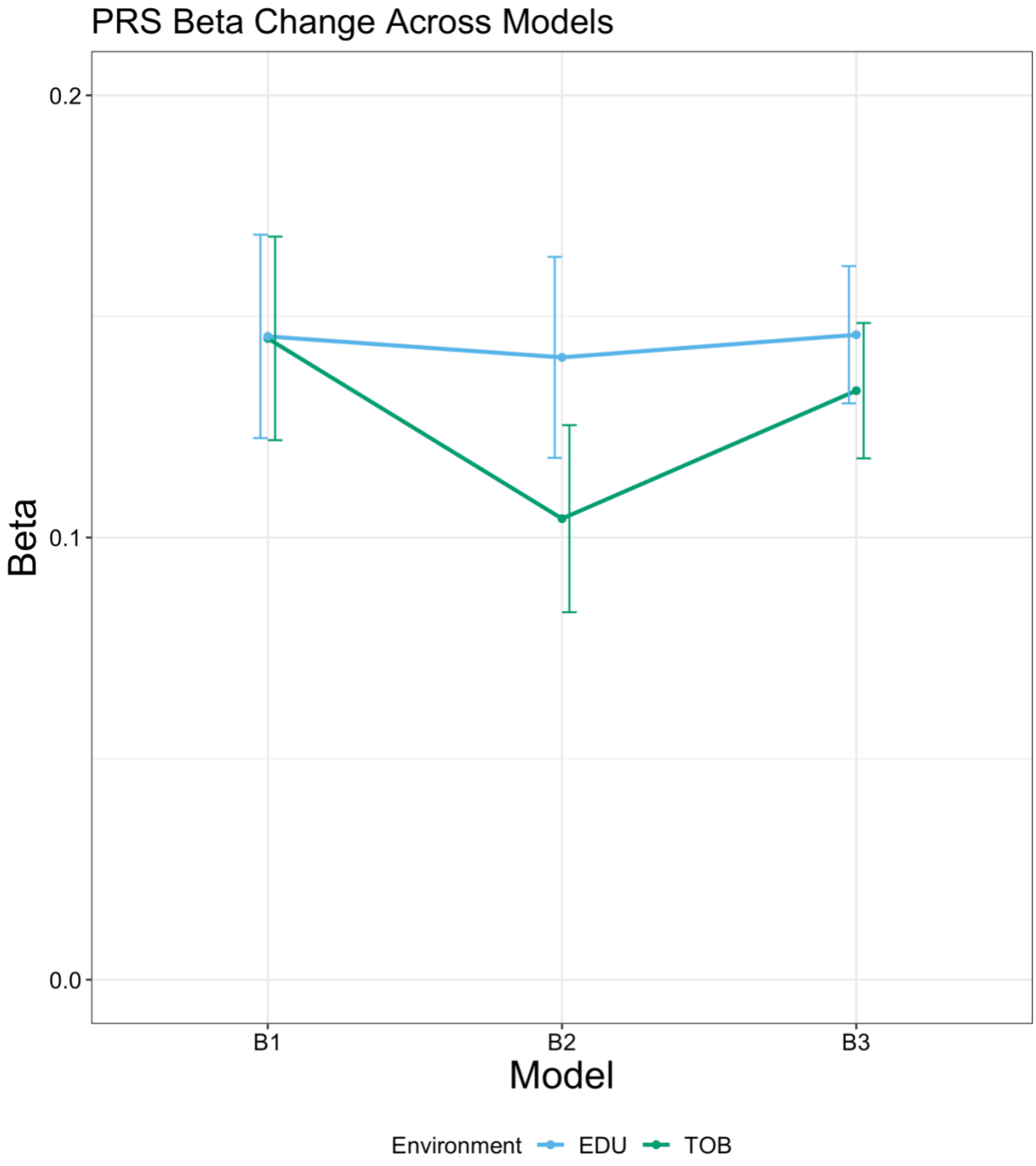

Figure 4 caption

PRS = polygenic risk score; TOB = tobacco use; EDU = educational attainment; PCs = principal components

The PRS beta is lower when TOB is included in the model. The beta increases when phenotypic PCs are added to the model**.** The PRS beta does not decrease substantially in the presence of EDU.

Table I Descriptive statistics for target phenotype and heritable environment in TOB and EDU models.

	Total n	Mean / n*	SD / Proportion*	Minimum	Maximum
TOB Sample					
Female*	7270	3850	0.53		
Age	7270	37.67	14.44	17	91
AUD Sx	7270	3.35	3.55	0	11
TOB*	7270	3739	0.51		
EDU Sample					
Female*	7286	3854	0.53		
Age	7286	37.69	14.45	17	91
AUD Sx	7286	3.35	3.55	0	11
EDU	7286	13.46	2.23	2	17

TOB = tobacco use; EDU = educational attainment; AUD Sx = DSM-5 Alcohol Use Disorder clinical criterion counts

Table II Base model R^2 and PRS Change R^2 across TOB models.

Base Model	Base Model R^2	Change R^2 with PRS	Percent Change
~ Ancestry PCs	0.013	0.020	
~ Ancestry PCs + TOB	0.181	0.010	50%
~ Ancestry PCs + TOB + Phenotypic PCs	0.562	0.017	35%

PRS = polygenic risk score; TOB = tobacco use; PCs = principal components

Table III Change in betas across models with TOB environment.

Model		B	SE	LowerCI95	UpperCI95
~ PRS (B1) / ~ TOB					
	PRS	0.145	0.012	0.122	0.168
	TOB	0.825	0.021	0.783	0.867
~ PRS + TOB (B2)					
	PRS	0.104	0.011	0.083	0.125
	TOB	0.804	0.021	0.763	0.846
~ PRS + TOB + Phenotypic PCs (B3)					
	PRS	0.133	0.008	0.118	0.149
	TOB	0.237	0.017	0.203	0.271

PRS = polygenic risk score; TOB = tobacco use; PCs = principal components

Table IV Base model R^2 and PRS Change R^2 across EDU models.

Base Model	Base Model R^2	Change R^2 with PRS	Percent Change
~ Ancestry PCs	0.012	0.020	
~ Ancestry PCs + EDU	0.038	0.019	5%
~ Ancestry PCs + EDU + Phenotypic PCs	0.541	0.020	5%

PRS = polygenic risk score; EDU = educational attainment; PCs = principal components

Table V Change in betas across models with EDU environment.

Model		B	SE	LowerCI95	UpperCI95
~ PRS (B1) / ~ EDU					
	PRS	0.145	0.012	0.122	0.168
	EDU	-0.160	0.012	-0.183	-0.137
~ PRS + EDU (B2)					
	PRS	0.141	0.012	0.118	0.163
	EDU	-0.156	0.011	-0.178	-0.133
~ PRS + EDU + Phenotypic PCs (B3)					
	PRS	0.146	0.008	0.130	0.161
	EDU	-0.003	0.008	-0.019	0.013

PRS = polygenic risk score; EDU = educational attainment; PCs = principal components

Acknowledgements:

The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, T. Foroud; Scientific Director, A. Agrawal; Translational Director, D. Dick, includes eleven different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, T. Foroud, Y. Liu, M. Plawecki); University of Iowa Carver College of Medicine (S. Kuperman, J. Kramer); SUNY Downstate Health Sciences University (B. Porjesz, J. Meyers, C. Kamarajan, A. Pandey); Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz, A. Agrawal); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, R. Hart, J. Salvatore); The Children's Hospital of Philadelphia, University of Pennsylvania (L. Almasy); Virginia Commonwealth University (D. Dick); Icahn School of Medicine at Mount Sinai (A. Goate, P. Slesinger); and Howard University (D. Scott). Other COGA collaborators include: L. Bauer (University of Connecticut); J. Nurnberger Jr., L. Wetherill, X., Xuei, D. Lai, S. O'Connor, (Indiana University); G. Chan (University of Iowa; University of Connecticut); D.B. Chorlian, J. Zhang, P. Barr, S. Kinreich, G. Pandey (SUNY Downstate); N. Mullins (Icahn School of Medicine at Mount Sinai); A. Anokhin, S. Hartz, E. Johnson, V. McCutcheon, S. Saccone (Washington University); J. Moore, Z. Pang, S. Kuo (Rutgers University); A. Merikangas (The Children's Hospital of Philadelphia and University of Pennsylvania); F. Aliev (Virginia Commonwealth University); H. Chin and A. Parsian are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting- Kai Li, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions. This national collaborative study is supported by

NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA).

This work was also supported by the National Institutes of Health (NIH) Grants R01AA028064 (PI: Salvatore) and K01AA024152 (PI: Salvatore) from the National Institute on Alcohol Abuse and Alcoholism (NIAAA).

Declarations

Funding: This work was supported by the National Institutes of Health (NIH) Grants R01AA028064 (PI: Salvatore) and K01AA024152 (PI: Salvatore) from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). The Collaborative Study on the Genetics of Alcoholism (COGA) is supported by NIH Grant U10AA008401 (PI: Porjesz).

Conflicts of interest/Competing interests: Nathaniel S. Thomas, Peter Barr, Fazil Aliev, Mallory Stephenson, Sally I-Chun Kuo, Grace Chan, Danielle M. Dick, Howard J. Edenberg, Victor Hesselbrock, Chella Kamarajan, and Jessica E. Salvatore declare that they have no conflicts of interest.

Ethics approval: The Institutional Review Board at all data collection sites approved the study.

Consent to participate: Written consent was obtained from all participants.

Consent for publication: NA

Availability of data and material: Data from the Collaborative Study on the Genetics of Alcoholism (COGA) are available via dbGaP (phs000763.v1.p1, phs000125.v1.p1) or through the National Institute on Alcohol Abuse and Alcoholism.

Code availability: The R scripts used in this work are available on GitHub at https://github.com/thomasns0/PCA_Collider.git

Authors' contributions: Nathaniel S. Thomas: conceived of the study, conducted statistical analyses, and wrote the manuscript. Peter Barr, Fazil Aliev, Mallory Stephenson, and Sally I-Chun Kuo assisted with the design and implementation of the study and provided editorial feedback on the whole manuscript. Grace Chan, Danielle M. Dick, Howard J. Edenberg, Victor Hesselbrock, and Chella Kamarajan provided editorial feedback on the whole manuscript. Jessica E. Salvatore supervised the design and implementation of the study and provided

editorial feedback on the whole manuscript. All authors contributed to and have approved the final manuscript.

References

Akimova, E. T., Breen, R., Brazel, D. M., & Mills, M. C. (2021). Gene-environment dependencies lead to collider bias in models with polygenic scores. *Scientific Reports*, *11*(1), 9457. https://doi.org/10.1038/s41598-021-89020-x

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5.* (5th edition). American Psychiatric Association.

Barr, P. B., Ksinan, A., Su, J., Johnson, E. C., Meyers, J. L., Wetherill, L., Latvala, A., Aliev, F., Chan, G., Kuperman, S., Nurnberger, J., Kamarajan, C., Anokhin, A., Agrawal, A., Rose, R. J., Edenberg, H. J., Schuckit, M., Kaprio, J., & Dick, D. M. (2020). Using polygenic scores for identifying individuals at increased risk of substance use disorders in clinical and population samples. *Translational Psychiatry*, *10*(1), 1–9. https://doi.org/10.1038/s41398-020-00865-8

Begleiter, H. (1995). The Collaborative Study on the Genetics of Alcoholism. *Alcohol Health and Research World*, *19*(3), 228–236.

Bucholz, K. K., Cadoret, R., Cloninger, C. R., Dinwiddie, S. H., Hesselbrock, V. M., Nurnberger, J. I., Reich, T., Schmidt, I., & Schuckit, M. A. (1994). A new, semi-structured psychiatric interview for use in genetic linkage studies: A report on the reliability of the SSAGA. *Journal of Studies on Alcohol*, *55*(2), 149–158. https://doi.org/10.15288/jsa.1994.55.149

Bucholz, K. K., McCutcheon, V. V., Agrawal, A., Dick, D. M., Hesselbrock, V. M., Kramer, J. R., Kuperman, S., Nurnberger, J. I., Salvatore, J. E., Schuckit, M. A., Bierut, L. J., Foroud, T. M., Chan, G., Hesselbrock, M., Meyers, J. L., Edenberg, H. J., & Porjesz, B. (2017). Comparison of parent, peer, psychiatric, and cannabis use influences across

stages of offspring alcohol involvement: Evidence from the COGA Prospective Study. *Alcoholism, Clinical and Experimental Research*, *41*(2), 359–368. https://doi.org/10.1111/acer.13293

Cheng, H., & Furnham, A. (2021). Personality, educational and social class predictors of adult tobacco usage. *Personality and Individual Differences*, *182*, 111085. https://doi.org/10.1016/j.paid.2021.111085

Dinno, A. (2018). *paran: Horn's Test of Principal Components/Factors* (R package version 1.5.2) [Computer software]. https://CRAN.R-project.org/package=paran

Domingue, B. W., Trejo, S., Armstrong-Carter, E., & Tucker-Drob, E. M. (2020). Interactions between Polygenic Scores and Environments: Methodological and Conceptual Challenges. *Sociological Science*, *7*, 465–486. https://doi.org/10.15195/v7.a19

Duncan, L. E., Ostacher, M., & Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*, *44*(9), 1518–1523. https://doi.org/10.1038/s41386-019-0389-5

Esch, P., Bocquet, V., Pull, C., Couffignal, S., Lehnert, T., Graas, M., Fond-Harmant, L., & Ansseau, M. (2014). The downward spiral of mental disorders and educational attainment: A systematic review on early school leaving. *BMC Psychiatry*, *14*(1), 237. https://doi.org/10.1186/s12888-014-0237-4

Fox, J. (2019). *polycor: Polychoric and Polyserial Correlations.* (R package version 0.7-10) [Computer software]. https://CRAN.R-project.org/package=polycor

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The structure of

haplotype blocks in the human genome. *Science (New York, N.Y.)*, *296*(5576), 2225–2229. https://doi.org/10.1126/science.1069424

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-09718-5

Green, V. R., Conway, K. P., Silveira, M. L., Kasza, K. A., Cohn, A., Cummings, K. M., Stanton, C. A., Callahan-Lyon, P., Slavit, W., Sargent, J. D., Hilmi, N., Niaura, R. S., Reissig, C. J., Lambert, E., Zandberg, I., Brunette, M. F., Tanski, S. E., Borek, N., Hyland, A. J., & Compton, W. M. (2018). Mental Health Problems and Onset of Tobacco Use Among 12- to 24-Year-Olds in the PATH Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*(12), 944-954.e4. https://doi.org/10.1016/j.jaac.2018.06.029

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerström, K. O. (1991). The Fagerström Test for Nicotine Dependence: A revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction*, *86*(9), 1119–1127. https://doi.org/10.1111/j.1360-0443.1991.tb01879.x

Hesselbrock, M., Easton, C., Bucholz, K. K., Schuckit, M., & Hesselbrock, V. (1999). A validity study of the SSAGA--a comparison with the SCAN. *Addiction (Abingdon, England)*, *94*(9), 1361–1370. https://doi.org/10.1046/j.1360-0443.1999.94913618.x

Keller, M. C. (2014). Gene × environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. *Biological Psychiatry*, *75*(1), 18–24. https://doi.org/10.1016/j.biopsych.2013.09.006

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(1), 1–16. https://doi.org/10.18637/jss.v074.i07

Kranzler, H. R., Zhou, H., Kember, R. L., Smith, R. V., Justice, A. C., Damrauer, S., Tsao, P. S., Klarin, D., Baras, A., Reid, J., Overton, J., Rader, D. J., Cheng, Z., Tate, J. P., Becker, W. C., Concato, J., Xu, K., Polimanti, R., Zhao, H., & Gelernter, J. (2019). Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-09480-8

Krapohl, E., Rimfeld, K., Shakeshaft, N. G., Trzaskowski, M., McMillan, A., Pingault, J.-B., Asbury, K., Harlaar, N., Kovas, Y., Dale, P. S., & Plomin, R. (2014). The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proceedings of the National Academy of Sciences*, *111*(42), 15273–15278. https://doi.org/10.1073/pnas.1408777111

Kuperman, S., Chan, G., Kramer, J. R., Wetherill, L., Bucholz, K. K., Dick, D., Hesselbrock, V., Porjesz, B., Rangaswamy, M., & Schuckit, M. (2013). A Model to Determine the Likely Age of an Adolescent's First Drink of Alcohol. *Pediatrics*, *131*(2), 242–248. https://doi.org/10.1542/peds.2012-0880

Lai, D., Wetherill, L., Bertelsen, S., Carey, C. E., Kamarajan, C., Kapoor, M., Meyers, J. L., Anokhin, A. P., Bennett, D. A., Bucholz, K. K., Chang, K. K., De Jager, P. L., Dick, D. M., Hesselbrock, V., Kramer, J., Kuperman, S., Nurnberger, J. I., Raj, T., Schuckit, M., … Foroud, T. (2019). Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes, Brain, and Behavior*, *18*(6), e12579. https://doi.org/10.1111/gbb.12579

Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E., & Neale, B. M. (2019). Predicting Polygenic Risk of Psychiatric Disorders. *Biological Psychiatry*, *86*(2), 97–109. https://doi.org/10.1016/j.biopsych.2018.12.015

Mõttus, R., Realo, A., Vainik, U., Allik, J., & Esko, T. (2017). Educational Attainment and Personality Are Genetically Intertwined. *Psychological Science*, *28*(11), 1631–1639. https://doi.org/10.1177/0956797617719083

Pasman, J. A., Verweij, K. J. H., & Vink, J. M. (2019). Systematic Review of Polygenic Gene–Environment Interaction in Tobacco, Alcohol, and Cannabis Use. *Behavior Genetics*, *49*(4), 349–365. https://doi.org/10.1007/s10519-019-09958-7

R Core Team. (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reich, T., Edenberg, H. J., Goate, A., Williams, J. T., Rice, J. P., Van Eerdewegh, P., Foroud, T., Hesselbrock, V., Schuckit, M. A., Bucholz, K., Porjesz, B., Li, T. K., Conneally, P. M., Nurnberger, J. I., Tischfield, J. A., Crowe, R. R., Cloninger, C. R., Wu, W., Shears, S., … Begleiter, H. (1998). Genome-wide search for genes affecting the risk for alcohol dependence. *American Journal of Medical Genetics*, *81*(3), 207–215.

Sanchez-Roige, S., Palmer, A. A., Fontanillas, P., Elson, S. L., Adams, M. J., Howard, D. M., Edenberg, H. J., Davies, G., Crist, R. C., Deary, I. J., McIntosh, A. M., & Clarke, T.-K. (2019). Genome-wide association study meta-analysis of the Alcohol Use Disorder Identification Test (AUDIT) in two population-based cohorts. *The American Journal of Psychiatry*, *176*(2), 107–118. https://doi.org/10.1176/appi.ajp.2018.18040369

Thomas, N. S., Kuo, S. I.-C., Aliev, F., McCutcheon, V. V., Jacquelyn, M. M., Chan, G., Hesselbrock, V., Kamarajan, C., Kinreich, S., Kramer, J. R., Kuperman, S., Lai, D.,

Plawecki, M. H., Porjesz, B., Schuckit, M. A., Dick, D. M., Bucholz, K. K., & Salvatore, J. E. (2021). *Alcohol Use Disorder, Psychiatric Comorbidities, Marriage and Divorce in a High-risk Sample* [Manuscript submitted for publication].

Uher, R., & Zwicker, A. (2017). Etiology in psychiatry: Embracing the reality of poly-gene-environmental causation of mental illness. *World Psychiatry*, *16*(2), 121–129. https://doi.org/10.1002/wps.20436

Veldman, K., Bültmann, U., Stewart, R. E., Ormel, J., Verhulst, F. C., & Reijneveld, S. A. (2014). Mental Health Problems and Educational Attainment in Adolescence: 9-Year Follow-Up of the TRAILS Study. *PLOS ONE*, *9*(7), e101751. https://doi.org/10.1371/journal.pone.0101751

Walters, R. K., Polimanti, R., Johnson, E. C., McClintick, J. N., Adams, M. J., Adkins, A. E., Aliev, F., Bacanu, S.-A., Batzler, A., Bertelsen, S., Biernacka, J. M., Bigdeli, T. B., Chen, L.-S., Clarke, T.-K., Chou, Y.-L., Degenhardt, F., Docherty, A. R., Edwards, A. C., Fontanillas, P., … Agrawal, A. (2018). Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nature Neuroscience*, *21*(12), 1656–1669. https://doi.org/10.1038/s41593-018-0275-1

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://www.springer.com/us/book/9780387981413

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *55*(10), 1068–1087. https://doi.org/10.1111/jcpp.12295

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou H., Sealock J. M., Sanchez-Roige S., Clarke T-K, Levey D. F., Cheng Z., Li B., Polimanti R., Kember R. L., Smith R. V., Thygesen J. H., Morgan M. Y., Atkinson S. R., Thursz M. R., Nyegaard M., Mattheisen M., Børglum A. D., Johnson E. C., Justice A. C., Palmer A. A., McQuillin A., Davis L. K., Edenberg H. J., Agrawal A., Kranzler H. R., Gelernter J. (2020). Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nature Neuroscience, 23*,809–818. https://doi.org/10.1038/s41593-020-0643-5