MR-APSS: a unified approach to Mendelian Randomization accounting for pleiotropy and sample structure using genome-wide summary statistics

Xianghong Hu[1†], Jia Zhao[1†], Zhixiang Lin[2], Yang Wang[1], Heng Peng[3], Hongyu Zhao[4*], Xiang Wan[5*], Can Yang[1*]

[1]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

[2]Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China

[3]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

[4]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[5]Shen Zhen Research Institute of Big Data, Shen zhen, China

Abstract (150 words)

Mendelian Randomization (MR) has proved to be a powerful tool for inferring causal relationships among a wide range of traits using GWAS summary statistics. Great efforts have been made to relax MR assumptions to account for confounding due to pleiotropy. Here we show that sample structure is another major confounding factor, including population stratification, cryptic relatedness, and sample overlap. We propose a unified MR approach, MR-APSS, to account for pleiotropy and sample structure simultaneously by leveraging genome-wide information. By further correcting bias in selecting genetic instruments, MR-APSS allows to include more genetic instruments with moderate effects to improve statistical power without inflating type I errors. We first evaluated MR-APSS using comprehensive simulations and negative controls, and then applied MR-APSS to study the causal relationships among a collection of diverse complex traits. The results suggest that MR-APSS can better identify plausible causal relationships with high reliability, in particular for highly polygenic traits.

†Co-first Author

*Correspondence should be addressed to Hongyu Zhao (hongyu.zhao@yale.edu), Xiang Wan (wanxiang@sribd.cn) and Can Yang (macyang@ust.hk)

Introduction

Inferring the causal relationship between a risk factor (exposure) and a phenotype of interest (outcome) is essential in biomedical research and social science [1]. Although randomized controlled trials (RCTs) are the gold standard for causal inference, RCTs can be very costly and sometimes even infeasible and unethical (e.g., random allocation to prenatal smoking) [2]. Mendelian randomization (MR) was introduced to mimic RCTs for causal inference in observational studies [3, 4]. Recently, MR analysis has drawn increasing attention [5] because it can take summary statistics from Genome-Wide Association Studies (GWASs) as input, including SNP effect size estimates and their standard errors, to investigate causal relationship among human complex traits.

MR is an instrumental variable (IV) method to infer causal relationship from the association between an exposure and an outcome, where genetic variants, e.g., single-nucleotide polymorphisms (SNPs), serve as IVs of the exposure [6, 7]. To eliminate the influence of confounding factors, conventional MR methods rely on strong assumptions of valid IVs, including (i) IVs are associated with the exposure; (ii) IVs are independent of confounding factors; and (iii) IVs only affect the outcome through the exposure. However, IV assumptions (ii) and (iii) are often not satisfied in practice due to confounding factors hidden in GWAS summary statistics, leading to false positive findings [8, 9]. To perform causal inference with genetic data, it is indispensable to distinguish two major confounding factors: pleiotropy [10] and sample structure [11, 12].

First, SNPs exhibit pervasive pleiotropic effects. Pleiotropy occurs when a genetic variant directly affects both exposure and outcome traits or indirectly through an intermediate phenotype [13]. Pleiotropy can induce trait association or genetic correlation in the absence of causal relationship [13]. Due to the polygenicity of complex traits and linkage disequilibrium (LD) in the human genome, the pleiotropic effects can widely spread across the whole genome [14]. Therefore, a substantial proportion of SNPs can carry pleiotropic effects and they fail to satisfy assumptions (ii) and (iii) on IVs in conventional MR methods.

Second, sample structure can lead to bias in SNP effect size estimates and introduce spurious trait associations. Here, sample structure encompasses population stratification, cryptic relatedness, and sample overlap in GWASs of the exposure and outcome traits. In the presence of population stratification and cryptic relatedness, SNPs can affect the outcome through the sample structure and thus they violate assumptions (ii) and (iii) on IVs. Without correcting for sample structure, SNP effect size estimates can be severely biased, which may lead to mis-interpretation on trait association and thus many false positive discoveries in causal inference. Sample overlap can also lead to spurious trait associations [15]. Although principal component analysis (PCA) [16] and linear mixed models (LMM) [17] are widely used to account for sample structure in GWASs, the results from LDSC [18] show that sample structure is often unsatisfactorily corrected in publicly available GWAS summary statistics.

Recently, a number of MR methods have been developed, including Egger [19], MR-PRESSO [20], GSMR [21], RAPS [22], BWMR [23], MRMix [24] and CAUSE [25], for more robust MR analysis. Despite much effort, there are two major limitations in the existing MR methods. First, most methods only use a small subset of strong IVs passing the genome-wide significance (p-value $< 5 \times 10^{-8}$) to account for pleiotropy. With limited information from a small number of strong IVs, it is very challenging to fit a flexible model to account for pleiotropy. Second, existing MR methods presume that PCA or LMM-based approaches have satisfactorily accounted for

sample structure and thus they largely ignore the influence of sample structure in the released GWAS summary statistics. Due to the complexity of human genetics, sample structure driven by socioeconomic status [26] or geographic structure [27] may not be corrected by routine adjustment and it still remains as a major confounding factor hidden in the released GWAS summary statistics.

Here, we develop MR-APSS, a unified approach to Mendelian Randomization Accounting for Pleiotropy and Sample Structure simultaneously. Specifically, we propose to decompose the observed SNP effect sizes into background and foreground signals, where the background signal can be attributed to the confounding factors hidden in the observed SNP effect sizes, including pleiotropy and sample structure, and the foreground signal corresponds to the remaining SNP effect sizes used for causal inference. MR-APSS differs from existing methods in the following aspects. First, under the assumptions of LD score regression (LDSC) [28], a background model is built to analytically account for pleiotropy and sample structure. In contrast to existing MR methods that account for confounding factors using a small number of selected IVs under strong assumptions, we estimate the background model using genome-wide summary statistics. Second, MR-APSS allows to include more SNPs with moderate effects as IVs to improve statistical power without inflating type I errors. With the pre-estimated background model, MR-APSS can inform whether a SNP belongs to the background component or the foreground component. Even in the presence of many invalid IVs, the type I error will not be inflated because only the foreground signals are used for causal inference. As more SNPs are included, the increasing amount of the foreground signal can improve statistical power.

To demonstrate the effectiveness of MR-APSS, we have performed a comprehensive simulation study and analysed 640 pairs of exposure and outcome traits from 26 GWASs. In the simulation study, we showed that MR-APSS had satisfactory performance when the assumptions of IVs were violated. We examined MR-APSS on a wide spectrum of complex traits using GWAS summary statistics, including psychiatric/neurological disorders, social traits, anthropometric traits, cardiovascular traits, metabolic traits and immune-related traits. Real data results indicate that pleiotropy and sample structure are two major confounding factors. By rigorous statistical modelling of these confounding factors, MR-APSS not only avoids many false positive findings but also improves statistical power of MR. When inferring causal relationships among highly polygenic traits, such as psychiatric disorders and social traits, the strengths of IVs tend to be relatively weak and causal inference is vulnerable to confounding effects. Thus, existing MR methods can suffer from either low statistical power or inflated type I errors. The empirical results indicate that MR-APSS is particularly useful in this scenario because it accounts for confounding factors and allows for incorporating many IVs with moderate effects, thus demonstrates its advantage over existing MR methods.

Results

Method Overview

Causality, pleiotropy, and sample structure are three major sources to induce correlation between GWAS estimates of exposure-outcome traits. To distinguish causality from correlation, it is indispensable to eliminate the possibility that correlation is induced by confounding factors, such as pleiotropy and sample structure (population stratification, cryptic relatedness and

sample overlap).

MR-APSS takes GWAS summary statistics of exposure and outcome traits as its input, and performs causal inference based on a proposed background-foreground model (see an overview in Fig. 1 and details in the Method section). Under the assumptions of LDSC, the background model can effectively account for confounding factors by disentangling pleiotropy (Fig. 1b) and sample structure (Fig. 1c). This is because the pleiotropic effects can be tagged by LD and the influence of sample structure is uncorrelated with LD [28]. By further accounting for winner's curse [29] due to selection of IVs (see Method section), the foreground model can use the classical causal diagram to perform causal inference (Fig. 1a). With these advances in statistical modeling, MR-APSS relaxes those strong assumptions in the conventional MR analysis.

The keys to the success of MR-APSS are twofold. First, MR-APSS accounts for confounding factors by exploiting the LD structure of the human genome. More specifically, pleiotropy and sample structure are captured by the first-order term and the zero-order term of LD score (slope and intercept of LDSC), respectively. Our real data results indicate that confounding factors associated with the first-order and zero-order terms of LD score are the major sources of confounding. Although confounding factors associated with higher-order terms of LD score may exist, they are nearly ignorable in real data analysis. Second, MR-APSS accounts for winner's curse [29] due to IV selection. Thus, it allows to include more IVs with moderate effects to improve statistical power without inflated type I errors. Through simulation and real data study, we demonstrate that MR-APSS can be applied to infer causal relationships with high reliability, in particular, for highly polygenic traits. The software of MR-APSS is freely available at `https://github.com/YangLabHKUST/MR-APSS`.

Compared methods

We compared the performance of MR-APSS with five representative MR methods: Inverse Variance Weighted regression (IVW) [6], Egger [19], RAPS [22], MRMix [24], and CAUSE [25]. We roughly grouped these methods into three categories based on their assumptions (Table 1). In the first group, MR methods require that all IVs satisfy the three strong assumptions in MR analysis. IVW is such a method and it tends to report many false positive findings when some assumptions do not hold. We included IVW in simulation study and real data analysis as a baseline method. In the second group, MR methods, such as Egger and RAPS, relax assumption (iii) by allowing for a direct effect between IV and outcome. Egger assumes a non-zero but constant direct effect to model the directional pleiotropy. For this purpose, Egger requires the orientation of genetic variants to avoid ambiguity [19, 30]. Therefore, Egger often suffers from large estimation errors and thus has low power to detect the causal effect [25]. RAPS also allows direct effects from IVs to the outcome trait but requires that the Instrument Strength must be Independent of the Direct Effect, which is known as the InSIDE condition [19]. In the third group, recently developed MR methods, such as MRMix and CAUSE, further relax the three strong assumptions. MRMix is a four-components mixture model, where its first component requires aforementioned MR assumptions (i), (ii) and (iii), and the other three components handle scenarios in which some of these assumptions do not hold. In other words, MRMix requires that there exist a proportion of IVs which exactly satisfy all the three

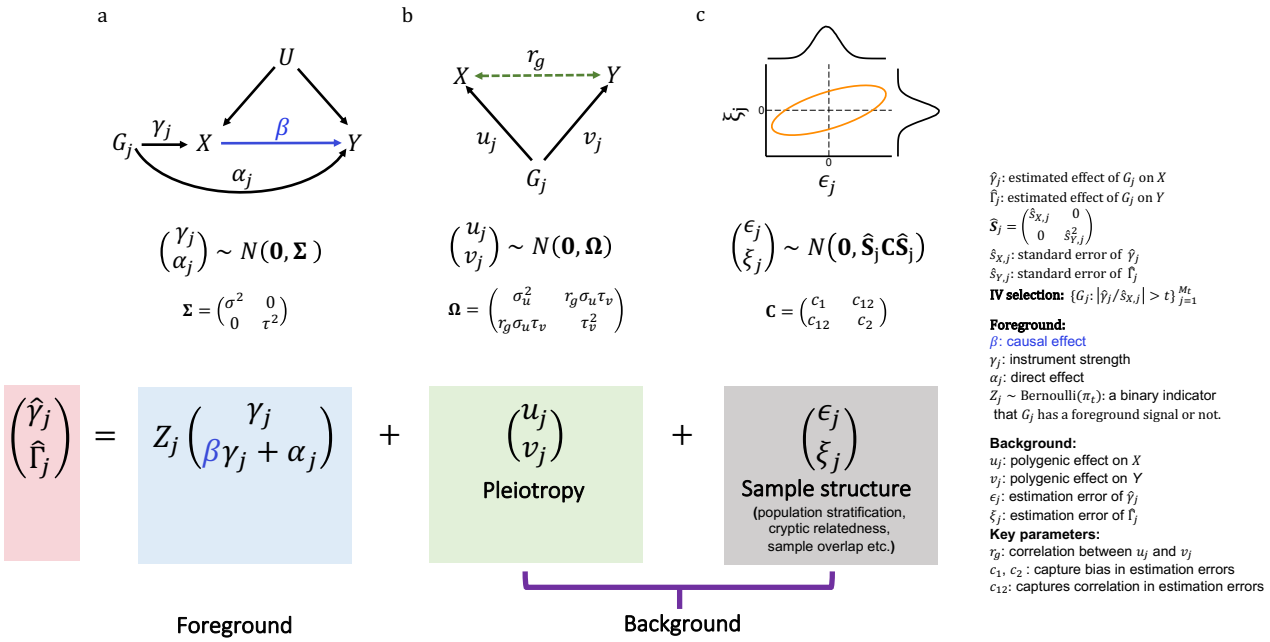

Figure 1: **The MR-APSS model**. To infer the causal effect β between exposure X and outcome Y, MR-APSS uses a foreground-background model to characterize the estimated effects of SNPs G_j on X and Y $(\hat{\gamma}_j, \hat{\Gamma}_j)$, where the background model accounts for pleiotropy (**b**) and sample structure (**c**), and the foreground model (**a**) performs causal inference. Specifically, we assume a variance component to model polygenic effects (u_j, v_j) and their correlation induced by pleiotropy $corr(u_j, v_j) = r_g$. We also explicitly account for the bias of estimation errors (ϵ_j, ξ_j) and their correlation, where c_1 and c_2 capture the biases and c_{12} captures the correlation. In the presence of population stratification and cryptic relatedness, c_1 and c_2 deviate from one, and c_{12} becomes nonzero due to either population stratification or sample overlap. Under the assumptions of LDSC, the background model can be well estimated using genome-wide summary statistics. After accounting for the major confounding factors, the foreground model can use the classical causal diagram to infer the causal effect. Note that IVs are selected by a z-score threshold t with $|\hat{\gamma}_j/\hat{s}_{X,j}| \geq t$ or an equivalent p-value threshold. The number of IVs M_t depends on threshold t: the smaller t, the larger M_t. For a given threshold t, a Bernoulli variable Z_j is introduced to indicate whether SNP G_j carries a foreground signal $(Z_j = 1)$ or not $(Z_j = 0)$. By further accounting for the bias in IV selection process, MR-APSS can accurately estimate the foreground model. With the estimated foreground-background model, it is easy for MR-APSS to exclude IVs without foreground signals $(Z_j = 0)$ and utilize valid IVs $(Z_j = 1)$ for causal inference. By using a loose threshold t, MR-APSS can include more SNP with moderate effects to improve statistical power without inflated type I errors.

strong assumptions. For example, by assumption (iii), IVs belonging to the first component are not allowed to have direct effects on outcome traits. However, SNPs often have small or moderate direct effects on outcome traits due to polygenicity and pleiotropy, which makes the assumptions of MRMix violated (see more discussion about MRMix in Section 1.6.1 of the Supplementary Note). CAUSE aims to account for pleiotropic effects by distinguishing two types of pleiotropy: correlated pleiotropy and uncorrelated pleiotropy. Despite this conceptual advance, it is very challenging for CAUSE to distinguish the underlying true causal effects from correlated pleiotropic effects. A closer examination of CAUSE shows that CAUSE tends to consider the causal effect as correlated pleiotropy during the model fitting, leading to very conservative performance for detecting causal effects (see more discussions in Section 1.6.2 of the Supplementary Note). Compared with existing MR methods, MR-APSS has much weaker assumptions. Specifically, it relaxes the InSIDE condition for all IVs by analytically accounting for pleiotropy and sample structure under the assumptions of LDSC. Moreover, MR-APSS accounts for winner's curse to avoid bias due to the IV selection. Thus, it allows to include more SNPs with moderate effects as IVs to improve statistical power without inflated type I errors.

Table 1: Summary of compared MR methods

Method	IV selection	Three IV assumptions	InSIDE assumption	Account for winner's curse	Account for sample structure
IVW	Strong IVs	(i)(ii)(iii)	Required	No	No
Egger	Strong IVs	(i)(ii)	Required	No	No
RAPS	Strong IVs	(i)(ii)	Required	No	No
MRMix	Strong IVs	(i)(ii)(iii) (for a subset of IVs)	Relaxed (for a subset of IVs)	No	No
CAUSE	Strong and moderate IVs	(i)(ii) (for a subset of IVs)	Relaxed (for a subset of IVs)	No	Partially (only account for sample overlap)
MR-APSS	Strong and moderate IVs	(i)	Relaxed (for all IVs)	Yes	Yes

IV: Instrumental Variable; Three IV assumptions: (i) IVs are associated with the exposure; (ii) IVs are independent of confounders; and (iii) IVs only affect the outcome through the exposure. InSIDE assumption: the Instrument Strength must be Independent of the Direct Effect. Winner's curse is also known as selection bias, i.e., the bias induced by selecting IVs.

Simulation studies

To evaluate MR-APSS in various scenarios and compare it with five MR methods in Table 1, we performed simulations based on real genotype data. For exposure and outcome traits, we used 47,049 SNPs on chromosomes 1 and 2 of 20,000 individuals of white British ancestry randomly drawn from the UK BioBank (UKBB). SNP effect sizes $(\gamma_j, \alpha_j, u_j, v_j)$ were generated from the relationship shown in Fig. 1 and Eq. (1) in the Method section. Based on real genotype data and simulated SNP effect sizes, we generated both traits and obtained summary statistics (as discussed in Section 1.7 of the Supplementary Note). The relationship shown in Fig. 1 is composed of the background signal and the foreground signal. For the background signal, polygenic effects (u_j, v_j) of all SNPs were normally distributed with variance components $(\sigma_u^2 = \tau_v^2 = 0.5/47,049)$, such that the heritabilities of both exposure X and outcome Y were specified at 0.5. The magnitudes of the error terms (ϵ_j, ξ_j) were determined by the fixed sample sizes of $20,000$. For the foreground signal, we randomly assigned 500 out of 47,049 SNPs as IVs. Note that the instrument strength (γ_j) and the magnitude of the direct effect (α_j) are

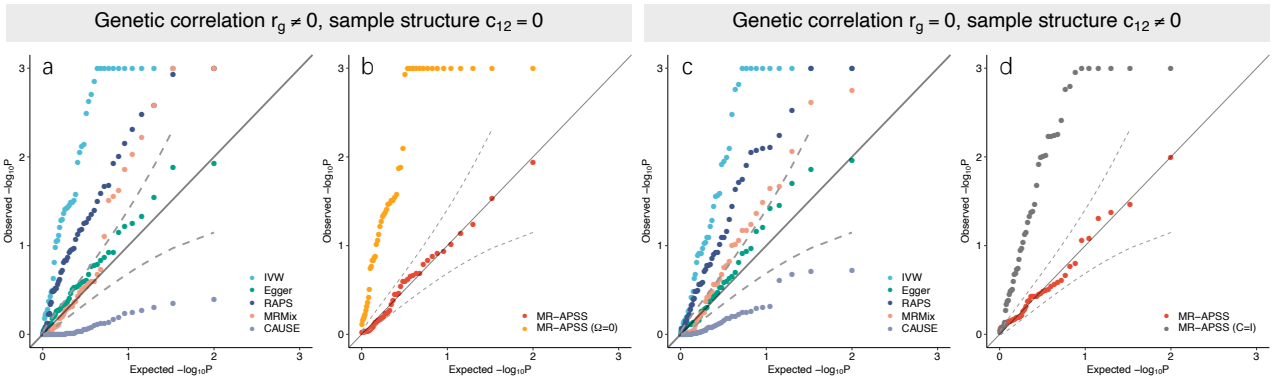

Figure 2: **Type I error control of different methods on inferring causal effects.** Quantile-quantile plots of $-\log_{10}(p)$-values from different methods in the absence of causal effect ($\beta = 0$). Null simulations were performed under different scenarios: (**a-b**) Null simulations with genetic correlation ($r_g = 0.2$) induced by pleiotropy, but without correlation in estimation errors ($c_{12} = 0$). (**c-d**) Null simulations in the presence of correlation in estimation errors ($c_{12} = 0.15$) due to sample structure, but in the absence of non-zero genetic correlation ($r_g = 0$). The results were summarized from 50 replications.

given by variance components σ^2 and τ^2 (Fig. 1). To mimic real data scenarios, we specified $\sigma^2 : \sigma_u^2 = 20$. We set $\tau^2 : \tau_v^2 = 1$ (i.e., magnitude of the direct effects in the foreground model is the same as that of the polygenic effects).

We compared MR-APSS to five MR methods, including IVW, Egger, RAPS, MRMix and CAUSE. Note that the performance of MR methods depends on the selected IVs. Using a stringent criterion, fewer SNPs will be selected as IVs and MR methods tend to have lower power of detecting the causal effect and low false positive rate. When more SNPs are included using a loose criterion, MR methods tend to have higher power but more false positive findings because their model assumptions are more likely to be violated. To evaluate the performance of MR methods under null ($\beta = 0$), we used a stringent criterion (IV threshold $p = 5 \times 10^{-6}$) to select IVs for IVW, Egger, RAPS and MRMix. For CAUSE, we used its default threshold $p = 1 \times 10^{-3}$ to include IVs. For MR-APSS, we used $p = 5 \times 10^{-4}$ as the IV threshold. For all six MR methods, we applied LD pruning ($r^2 = 0.01$) to selected IVs, and ensured that they were nearly independent.

We first examined type I error control of different MR methods under the null ($\beta = 0$) in the presence of genetic correlation induced by pleiotropy. We simulated data with genetic correlation but without correlation in estimation errors. Quantile-quantile plots of different MR methods are shown in Fig. 2 (a,b) for genetic correlation $r_g = 0.2$ (more results for different genetic correlations are given in Supplementary Fig. S2). Clearly, MR-APSS is the only method that produces well calibrated p-values. To better examine how MR-APSS accounted for polygenicity and pleiotropy, we manually set the variance component of MR-APSS to zero, i.e., $\boldsymbol{\Omega} = \mathbf{0}$. We denote this version of MR-APSS as MR-APSS ($\boldsymbol{\Omega} = \mathbf{0}$). As shown in Fig. 2b, MR-APSS produced well calibrated p-values while MR-APSS ($\boldsymbol{\Omega} = \mathbf{0}$) produced overly inflated p-values. This suggests that variance component $\boldsymbol{\Omega}$ plays a critical role in accounting for polygenicity and pleiotropy. We also noticed different performance of alternative MR methods (Fig. 2a). In the presence of non-zero genetic correlation, MR methods, such as IVW, RAPS,

MRMix and MR-APSS ($\mathbf{\Omega} = \mathbf{0}$), tended to produce inflated p-values. Different from other MR methods, CAUSE produced very deflated p-values and thus CAUSE was very conservative in identifying causal effects.

Next, we examined type I error control under null ($\beta = 0$) in the presence of correlation between estimation errors due to sample structure. To focus on correlation in estimation errors, we set genetic correlation $r_g = 0$ and simply generated correlation of estimation errors ($c_{12} = 0.15$) using 10,000 overlapped samples in exposure and outcome studies (more results for different c_{12} are given in Supplementary Fig. S3). We notice that correlation between estimation errors can also be induced by population stratification and cryptic relatedness. To avoid unrealistic simulation of population stratification, we investigated this issue when we performed real data analysis. The quantile-quantile plots of different MR methods are shown in Fig. 2 (c,d). IVW, RAPS, and MRMix produced overly inflated p-values. These results indicate that correlation between estimator errors can be a major confounding factor which leads to false positive findings. Again, CAUSE produced very deflated p-values. To see how MR-APSS accounts for correlation between estimation errors, we set $\mathbf{C} = \mathbf{I}$, i.e., $c_1 = c_2 = 1$ and $c_{12} = 0$. In such a way, MR-APSS was forced to ignore the correlation between estimation errors. We denote this version of MR-APSS as MR-APSS ($\mathbf{C} = \mathbf{I}$). As shown in Fig. 2d, MR-APSS ($\mathbf{C} = \mathbf{I}$) produced inflated p-values. In contrast, MR-APSS also produced well calibrated p-values. These evidence suggests that MR-APSS can satisfactorily account for correlation between estimation error due to sample structure.

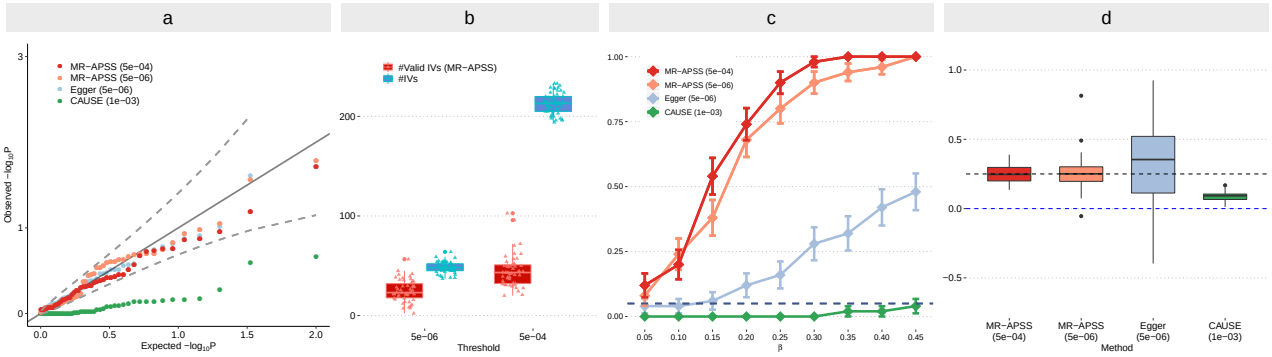

Figure 3: **Comparison of MR-APSS, Egger and CAUSE under both the null and the alternative simulations.** (**a**) Quantile-quantile plots of $-\log_{10}(p)$-values under null simulations with both genetic correlation ($r_g = 0.1$) and correlation of estimation error ($c_{12} = 0.1$). (**b**) Boxplots comparing the estimated number of valid IVs obtained from MR-APSS with the number of selected IVs. (**c**) The power under causal effect size β varied from 0.0 to 0.45. (**d**) Estimate of causal effect under the alternative simulations ($\beta = 0.25$). The results are summarized from 50 replications.

Finally, we examined power of MR methods. As shown above, IVW, RAPS and MRMix often produced overly inflated type I errors in the presence of either pleiotropy or sample structure. Hence, we only compared MR-APSS with Egger and CAUSE. Recall that the power of MR methods depends on the number of IVs. As a common practice, we used a stringent threshold $p = 5 \times 10^{-6}$ for Egger and denote it as Egger (5×10^{-6}). We used the default threshold $p = 1 \times 10^{-3}$ for CAUSE. To have a fair comparison, we considered two IV thresholds

$p = 5 \times 10^{-6}$ and $p = 5 \times 10^{-4}$ for MR-APSS, and denote them as MR-APSS (5×10^{-6}) and MR-APSS (5×10^{-4}), respectively. We simulated data with both genetic correlation ($r_g = 0.1$) and correlation between estimation error ($c_{12} = 0.1$). We varied the causal effect size β from 0.0 to 0.35. Under the null ($\beta = 0$), MR-APSS (5×10^{-4}), MR-APSS (5×10^{-6}), and Egger (5×10^{-6}) produced well calibrated p-values (Fig. 3a). Again, CAUSE produced overly deflated p-values (Fig. 3a). Under the alternative ($\beta \neq 0$), MR-APSS (5×10^{-4}) was the overall winner in terms of the power (Fig. 3c). To see why MR-APSS (5×10^{-4}) could further improve MR-APSS (5×10^{-6}), we recorded the estimated number of valid IVs and the number of selected IVs corresponding to the two thresholds (Fig. 3b). When the p-value threshold varied from $p = 5 \times 10^{-6}$ to $p = 5 \times 10^{-4}$, the number of SNPs included as IVs increased a lot. However, MR-APSS did not use all of them as valid IVs. Recall that a binary variable Z_j was introduced in MR-APSS to SNP j to indicate whether it had a foreground signal ($Z_j = 1$) or not ($Z_j = 0$). MR-APSS only used SNPs with foreground signals for causal inference, avoiding inflated type I errors. When the p-values varied from $p = 5 \times 10^{-6}$ to $p = 5 \times 10^{-4}$, the number of valid IVs increased steadily. Therefore, statistical power of MR-APSS was improved by including more SNPs with moderate effects. We further compared the estimation accuracy of the causal effects using MR-APSS, Egger and CAUSE (Fig. 3d). Consistent with the literature [30], we observed that Egger had a very large estimation error. As discussed in Section 1.6.2 of the Supplementary Note, CAUSE often mis-interprets the causal effect as correlated pleiotropy, leading to underestimation of the true causal effect. Consistently, we observed that the estimate of CAUSE was biased to the null ($\beta = 0$). In the above simulations, the foreground-background variance ratio was fixed at $\sigma : \sigma_u = 20 : 1$. We provide more results with different foreground-background variance ratios ($\sigma : \sigma_u \in \{40, 10\}$) in Supplementary Figs. S4 and S5.

Besides the above simulation study with our own model, we also conducted simulation with the CAUSE model. The main patterns of the performance of the six MR methods largely remained the same. We provide details in Section 1.7 and Figs. S6, S7, S8 of the Supplementary Note.

Real data analysis: negative control outcomes

In this section, we evaluate type I error control of MR methods by applying them to infer the causal effect of exposures on negative control outcomes, where we do not expect any causal effects. The traits that can serve as ideal negative control outcomes should satisfy two conditions. First, they should not be causally affected by any of the exposure traits considered here. Second, the exposure and outcome traits could be affected by some unmeasured confounders, e.g., population stratification. Following the same way of [11] to choose negative control outcomes, we considered natural hair color before greying (Hair colour: black, Hair colour: blonde, Hair color: light brown, Hair color: dark brown) and skin tanning ability (Tanning) from UKBB because they are largely determined at birth and they could be affected by sample structure.

We considered 26 exposure traits from UKBB and Genomics Consortiums (Details for the GWAS sources are given in Supplementary Table S1). These traits can be roughly divided into five categories, including psychiatric/neurological disorders, social traits, anthropometric traits, cardiometabolic traits, immune-related traits. The sample sizes of those GWASs range from

114,244 to 385,603, with a minimum of 15,954 for ASD and a maximum of 898,130 for T2D. Given the large sample size of GWASs, we used genome-wide significance threshold 5×10^{-8} as the IV threshold for IVW, RAPS, Egger, and MRMix in real data analysis. This stringent criterion helps to exclude invalid IVs for these methods and thus reduce their false positive rates. Due to the stringent IV selection, we were not able to find enough SNPs (> 4) as IVs for four exposure traits, i.e., major depressive disorder (MDD), autism spectrum disorder (ASD), subject well-being (SWB) and number of children Ever born (NEB). For CAUSE [25], we used its default p-value threshold $p = 1 \times 10^{-3}$ to select IVs. For MR-APSS, we used 5×10^{-5} as the default IV threshold.

First, we applied MR-APSS and five other MR methods to infer the causal effects between these 26 exposure traits and 5 negative control outcomes. To make fair comparison, we focus on the results for 110 pairs where each method had sufficient IVs for MR analysis. Ideally, these p-values should be uniformly distributed between 0 and 1 under the null ($\beta = 0$). Fig. 4a shows the QQ-plots of $-\log_{10}(p)$ values of the six methods (red dots). Clearly, MR-APSS produced well calibrated p-values. IVW, RAPS and MRMix produced overly inflated p-values, and Egger produced slightly inflated p-values. CAUSE produced deflated p-values in the beginning but inflated p-values later. We investigated the reason why the five MR methods performed unsatisfactorily. As shown in Fig. 4b, we examined the estimates of two key parameters, r_g and c_{12}, of our background model, where r_g captures the overall correlated pleiotropic effects and c_{12} captures the correlation of estimation errors due to sample structure. Among the 110 pairs of exposure and outcome traits, 81 trait pairs had nearly zero genetic correlation and 29 trait pairs had nonzero genetic correlation at the nominal level 0.05 (marked by *). We also examined the correlation of estimation errors due to sample structure (e.g., population stratification, cryptic relatedness, and sample overlap). Among the 110 trait pairs, 63 trait pairs had significant nonzero $\hat{c}_{12}$ at the nominal level 0.05 (marked by *). To identify the major reason for the inflated p-values produced by the five MR methods, we restricted ourselves to the 81 trait pairs whose genetic correlation was nearly zero. For these 81 trait pairs, we generated the QQ-plots of $-\log_{10}(p)$ values of the six MR methods (blue triangles in Fig. 4a). Clearly, IVW, RAPS and MRMix still produced overly inflated p-values, Egger produced slightly better calibrated p-values, and CAUSE produced deflated p-values in the beginning but inflated p-values later. We further restricted ourselves to trait pairs whose genetic correlation and correlation of estimation error were both nearly zero. For these trait pairs (green diamond), MR-APSS, RAPS, MRMix and Egger produced well calibrated p-values. IVW still produced inflated p-values because it did not allow direct effects. Again, CAUSE produced very conservative p values. These results suggest that sample structure is another major confounding factor in addition to pleiotropy.

It is worthwhile to mention that nonzero c_{12} can be induced by either population stratification or sample overlap. To see this, let us consider the relationship between Height (GIANT) [31] and Tanning from UKBB. Recall that parameters c_1 and c_2 capture the bias in estimation errors (ϵ_j, ξ_j) and parameter c_{12} captures their correlation (Fig. 1). By applying LDSC to estimate our background model, we obtained the estimates of c_1 and c_2: $\hat{c}_1 = 1.34$ (s.e. $= 0.022$) for Height (GIANT), $\hat{c}_2 = 1.81$ (s.e. $= 0.023$) for Tanning, respectively. These results indicate that the publicly released GWAS summary statistics are affected by confounding factors, such as population stratification. By applying bivariate LDSC, we obtained $\hat{c}_{12} = -0.17$ (s.e. $= 0.011$).

As we know, the samples from GIANT do not overlap with UKBB [32]. Therefore, the nonzero $\hat{c}_{12}$ value should be mainly attributed to population stratification. As a comparison, we also considered Height (UKBB) [33] and Tanning from UKBB. By applying LDSC, we obtained $\hat{c}_1 = 1.97$ (s.e. $= 0.040$) for Height (UKBB), suggesting that the released GWAS summary statistics of Height (UKBB) might potentially suffer from population stratification. By applying bivariate LDSC, we obtained $\hat{c}_{12} = -0.36$ (s.e. $= 0.014$) for Height (UKBB) and Tanning (UKBB). Such a nonzero value could be attributed to both population stratification and sample overlap.

To better examine the role of MR-APSS in accounting for pleiotropy or sample structure, we applied MR-APSS but fixed $\mathbf{\Omega} = \mathbf{0}$ and $\mathbf{C} = \mathbf{I}$, respectively. We denote the two variations as MR-APSS ($\mathbf{\Omega} = \mathbf{0}$) and MR-APSS ($\mathbf{C} = \mathbf{I}$), where MR-APSS ($\mathbf{\Omega} = \mathbf{0}$) does not account for pleiotropy and MR-APSS ($\mathbf{C} = \mathbf{I}$) does not account for sample structure. As shown in 4c, both MR-APSS ($\mathbf{\Omega} = \mathbf{0}$) and MR-APSS ($\mathbf{C} = \mathbf{I}$) reported inflated p-values. For example, Based on Bonferroni correction, several trait pairs (marked with black circles in Fig. 4c) were falsely detected as causal by MR-APSS ($\mathbf{\Omega} = \mathbf{0}$) and MR-APSS ($\mathbf{C} = \mathbf{I}$). As shown in Fig. 4b (marked by squares), their corresponding $\hat{r}_g$ and $\hat{c}_{12}$ values were significantly different from zero. By using negative control outcomes, we show that MR-APSS can produce well calibrated p-values by accounting for pleiotropy and sample structure.

Inferring causal relationships among complex traits

To perform causal inference, we considered 26 complex traits from five categories including psychiatric/neurological disorders, social trait, anthropometric traits, cardiometabolic traits, and immune-related traits. Before applying MR methods, we examined the estimates of r_g and c_{12} in the background model of MR-APSS for all 325 pairwise combination of the 26 traits. We found that genetic correlation of 198 pairs significantly differed from zero at the nominal level 0.05 (marked by $*$ in Fig. 5a). Among them, genetic correlation of 130 pairs remained to be significant after Bonferroni correction with $p \leq 0.05/325$ (marked by $**$ in Fig. 5a). For the estimates of c_{12}, 126 pairs had significant nonzero $\hat{c}_{12}$ at the nominal level 0.05 (marked by $*$ in Fig. 5b), and 76 pairs of them remained to be significantly different from zero after Bonferroni correction (marked by $**$ in Fig. 5b). Of note, 56 pairs of traits had significantly nonzero estimates of both $\hat{r}_g$ and $\hat{c}_{12}$ after Bonferroni correction. The above results suggest that both pleiotropy and sample structure are presented as major confounding factors for causal inference.

We considered inferring the causal relationship between traits X and Y in both directions, i.e., $X \to Y$ (X as exposure and Y as outcome) and $Y \to X$ (Y as exposure and X as outcome). To avoid causal inference between two very similar phenotypes (e.g., Angina and CAD), we excluded several trait pairs which were marked in grey color as non-diagonal cells in Fig. 5c. Therefore, 640 trait pairs remained for MR tests in total. We applied MR-APSS to these trait pairs using IV threshold $p = 5 \times 10^{-5}$ and identified 34 significant causal relationships after Bonferroni correction (Fig. 5c, marked by triangles). As shown in Fig. 5a, many traits in social or neurological/psychiatric categories were observed to be genetically correlated with a wide range of complex traits from different categories. After accounting for pleiotropy and sample structure, the results from MR-APSS indicate that genetic correlation of many trait

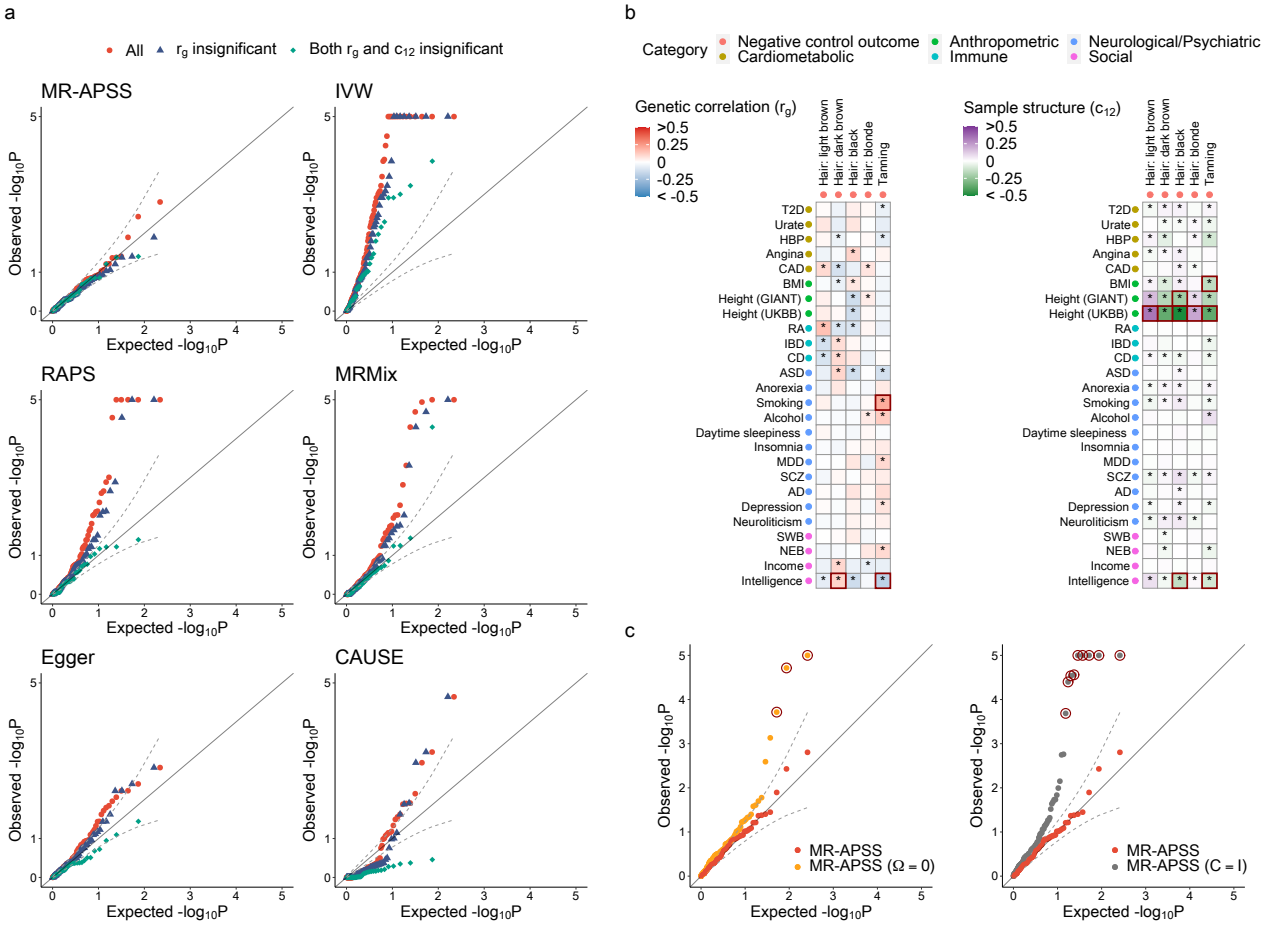

Figure 4: **Evaluation of type I error control of MR methods using negative control outcomes.** (**a**) Quantile-quantile plots of $-\log_{10}(p)$-values from six MR methods for causal inference between complex traits and negative control outcome. Red dots present all 110 trait pairs tested by each method. Blue triangles present the 81 trait pairs with insignificant genetic correlation at the nominal level 0.05. Green diamonds present the 29 trait pairs whose genetic correlation r_g and c_{12} are both insignificant at the nominal level 0.05. (**b**) Estimates of r_g and c_{12} for trait pairs between 26 complex traits and five negative control outcomes. (**c**) Quantile-quantile plots of $-\log_{10}(p)$-values from MR-APSS, MR-APSS ($\mathbf{\Omega = 0}$) and MR-APSS ($\mathbf{C = I}$) for trait pairs between 26 complex traits and five negative control outcomes. The circled p-values correspond to the trait pairs squared in (b), which are largely confounded by pleiotropy and sample structure.

pairs should be not be attributed to the causal effects. An example is Intelligence which was observed to be genetically correlated with 18 complex traits, such as type 2 diabetes (T2D) ($\hat{r}_g = -0.132, \text{s.e.} = 0.023$) and coronary artery disease (CAD) ($\hat{r}_g = -0.183, \text{s.e.} = 0.030$) from the Cardiometabolic category, body mass index (BMI) ($\hat{r}_g = -0.129, \text{s.e.} = 0.021$), Height (GIANT) ($\hat{r}_g = 0.090, \text{s.e.} = 0.019$) and Height (UKBB) ($\hat{r}_g = 0.156, \text{s.e.} = 0.018$) from the Anthropometric category, schizophrenia (SCZ) ($\hat{r}_g = -0.264, \text{s.e.} = 0.024$) and major depressive disorder (MDD) ($\hat{r}_g = -0.206, \text{s.e.} = 0.048$) from the neurological/psychiatric category, household income (Income) ($\hat{r}_g = 0.577, \text{s.e.} = 0.048$) and number of children ever born (NEB) ($\hat{r}_g = 0.172, \text{s.e.} = 0.039$). Among the 18 traits, MR-APSS only confirmed one significant causal effect of Intelligence on household income (Income) ($\hat{\beta} = 0.355$, p-value $= 2.23 \times 10^{-9}$). Another example is Depression which was also genetically correlated with 18 complex traits from different categories, such as T2D ($\hat{r}_g = 0.194, \text{s.e.} = 0.027$) and CAD ($\hat{r}_g = 0.157, \text{s.e.} = 0.035$) from Cardiometabolic category, BMI ($\hat{r}_g = 0.220, \text{s.e.} = 0.024$) from Anthropometric category, Insomnia ($\hat{r}_g = 0.454, \text{s.e.} = 0.025$) and SCZ ($\hat{r}_g = 0.321, \text{s.e.} = 0.027$) from neurological/psychiatric category, Income ($\hat{r}_g = -0.450, \text{s.e.} = 0.029$) and NEB ($\hat{r}_g = 0.214, \text{s.e.} = 0.048$). Again, MR-APSS only confirmed the causal effect of Depression on Insomnia ($\hat{\beta} = 0.570$, p-value $= 4.38 \times 10^{-5}$). Clearly, MR-APSS can serve as an effective tool to distinguish causality from genetic correlation.

As a comparison, we also applied the five other MR methods to infer the causal relationships for the 640 trait pairs. We used $p = 5 \times 10^{-8}$ as the IV selection threshold for IVW, RAPS, MRMix and Egger and used $p = 1 \times 10^{-3}$ as the IV selection threshold for CAUSE. For MR methods including IVW, RAPS, Egger and MRMix, only 541 trait pairs were tested because 99 trait pairs had less than four SNPs as IVs. For CAUSE, all 640 trait pairs were included. A summary of the causal relationship detected by the five compared methods are given in Supplementary Figs. S13-S17. RAPS reported 58 trait pairs with significant causal effects after Bonferroni correction. Among them, 24 trait pairs were considered insignificant by MR-APSS after Bonferroni correction. Notably, RAPS made a similar assumption with the foreground model of MR-APSS, however, it has no background model to account for pleiotropy and sample structure. To better understand the difference between RAPS and MR-APSS, we applied MR-APSS ($\mathbf{\Omega} = \mathbf{0}$) or MR-APSS ($\mathbf{C} = \mathbf{I}$) to those trait pairs. The testing p-values of 18 trait pairs became significant based on Bonferroni correction. An example was BMI and Insomnia (see Fig. 6a) with $\hat{r}_g = 0.184$ (s.e. $= 0.025$) and $\hat{c}_{12} = 0.058$ (s.e. $= 0.010$). RAPS produced $\hat{\beta} = 0.07$ with p-value 3.04×10^{-9}. Without accounting for pleiotropy or sample structure, MR-APSS ($\mathbf{\Omega} = \mathbf{0}$) and MR-APSS ($\mathbf{C} = \mathbf{I}$) reported $\hat{\beta} = 0.070$ with p-value $= 1.70 \times 10^{-7}$ and $\hat{\beta} = 0.063$ with p-value $= 1.01 \times 10^{-4}$, respectively. After accounting for both pleiotropy and sample structure, MR-APSS estimated causal effect between BMI and Insomnia as $\hat{\beta} = 0.0337$ with p-value $= 0.128$. The result indicated that RAPS was likely affected by pleiotropy and sample structure.

Since IVW, RAPS, MRMix tended to have higher type I error than the nominal level, we mainly compared statistical power of MR-APSS with Egger and CAUSE (Fig. 5d). A complete list of causal relationship among these traits detected by MR-APSS, Egger and CAUSE are summarized in Supplementary Table S2. Based on Bonferroni correction, MR-APSS detected 18 significant causal effects which were not reported by CAUSE and Egger, showing higher statistical power of MR-APSS than both CAUSE and Egger. For example, MR-APSS detected

significant causal effects of BMI on eight traits. Five of them were supported with evidence of causality from previous literature, including T2D [34], serum urate (Urate) [35] and three cardiovascular diseases (high blood pressure (HBP), Angina and CAD) [36]. For these five supported trait pairs, Egger only detected two significant causal relationships (BMI and CAD; T2D and HBP), and CAUSE only detected two significant causal relationships (BMI and Urate; HBP and T2D). In addition to the confirmed findings, MR-APSS detected significant causal effects of BMI on Depression ($\hat{\beta} = 0.07$, p-value $= 2.09 \times 10^{-5}$), ever smoked regularly (Smoking) ($\hat{\beta} = 0.11$, p-value $= 1.36 \times 10^{-6}$) and Income ($\hat{\beta} = -0.17$, p-value $= 1.83 \times 10^{-11}$). Those findings are consistent with results from previous MR studies [37, 38, 39], suggesting that being overweight not only increases the risk of depression and tobacco dependence but also suffers from reduced income. For these three traits, Egger only detected a significant causal effect of BMI on Income. CAUSE detected significant causal effects of BMI on both Income and Smoking. Neither CAUSE nor Egger detected significant causal effects between BMI and Depression (Fig. 6b). Our results also revealed Neuroticism as an important health indicator especially for human psychiatric health. Neuroticism is one of the big five personality traits, characterized by negative emotional states including sadness, moodiness, and emotional instability. Higher neuroticism is associated with premature mortality and a wide range of mental illnesses or psychiatric disorders [32, 40]. There is growing evidence that neuroticism plays a causal role in psychiatric disorders, such as SCZ [41] and MDD [42]. Evidence from MR-APSS also supported the significant causal effect of Neuroticism on SCZ ($\hat{\beta} = 0.57$, p-value $= 7.02 \times 10^{-7}$) and MDD ($\hat{\beta} = 0.18$, p-value $= 2.06 \times 10^{-5}$). Neither CAUSE nor Egger detected significant causal effects of Neuroticism on MDD (Fig. 6c) or SCZ. MR-APSS also revealed that Neuroticism could be causally linked to Insomnia ($\hat{\beta} = 0.29$, p-value $= 2.7 \times 10^{-10}$) and Anorexia ($\hat{\beta} = 0.4$, p-value $= 6.90 \times 10^{-7}$). Egger did not report these two cases, and CAUSE only detected a significant causal effect between Neuroticism and Insomnia ($\hat{\beta} = 0.14$, p-value $= 3.89 \times 10^{-6}$).

It is worthwhile to mention the two trait pairs reported by Egger but not reported by MR-APSS. One trait pair is HBP and BMI. It is well known that HBP and BMI are positively correlated [36]. Among all methods, only Egger reported a causal effect with negative direction (Fig. 6d). As suggested by strong confounding effect from pleiotropy ($\hat{r}_g = 0.346$, s.e. $= 0.019$) and sample structure ($\hat{c}_{12} = 0.289$, s.e. $= 0.0102$), the InSIDE condition has been severely violated and Egger tends to report a biased result. By accounting for pleiotropy and sample structure between HBP and BMI, the result given MR-APSS implies no clear evidence for the causal effect of HBP on BMI. Another trait pair detected by Egger but not MR-APSS is HBP and Urate. Again, this result is likely a false positive finding because of the unaccounted confounding effects (Supplementary Fig. S18). Consistently with literature [30], the real data results here indicate that Egger can produce unreliable estimates when the InSIDE condition does not hold.

Improving statistical power of MR-APSS by including more IVs with moderate effects

We have presented the real data results of MR-APSS using a relative looser IV threshold 5×10^{-5} where more IVs below the genome-wide significance (5×10^{-8}) are involved. As more

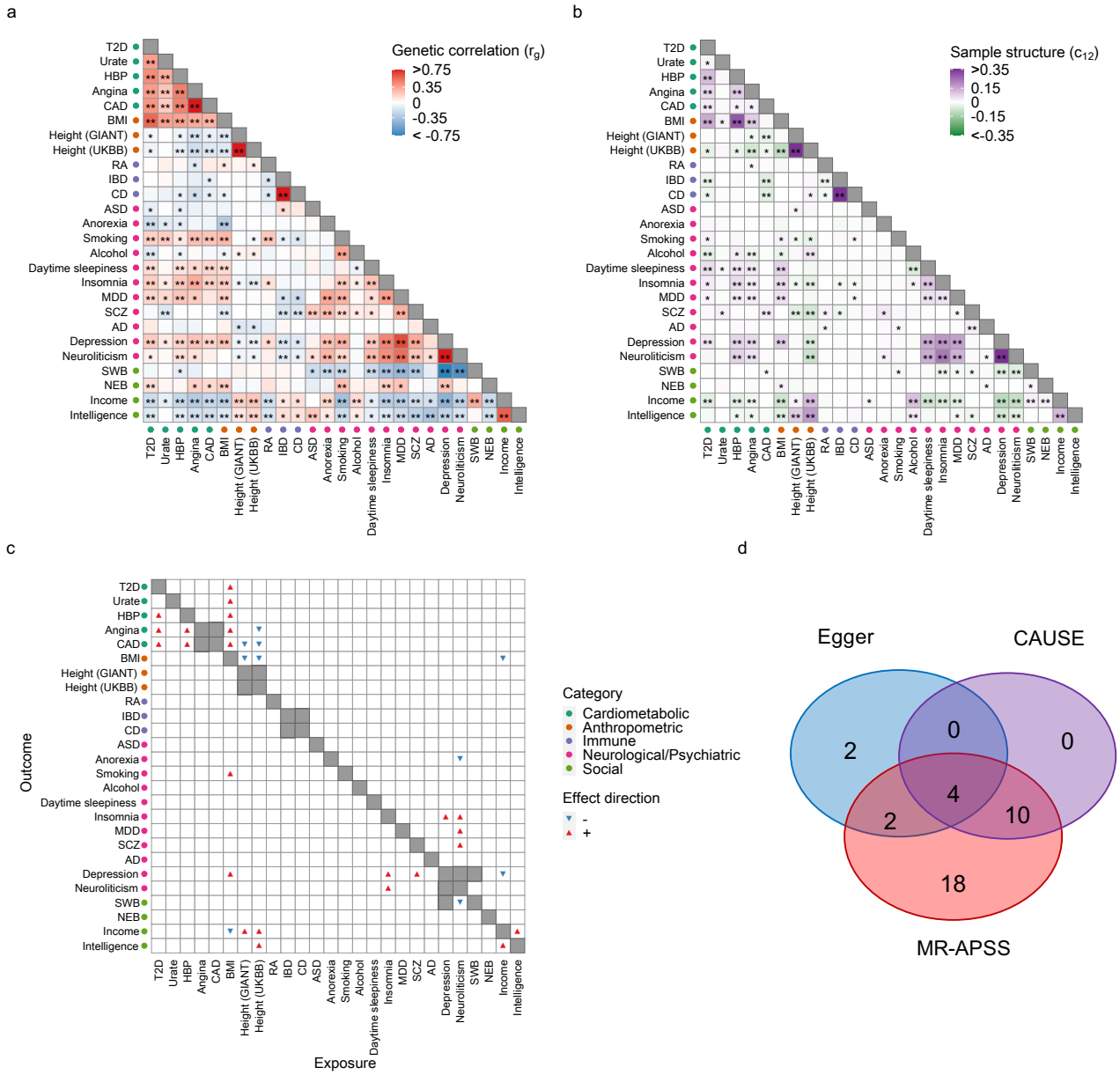

Figure 5: Application of MR-APSS to infer causal relationships between 26 complex traits. (**a**) Estimates of genetic correlation between 26 complex traits. Positive and negative estimates of genetic correlation $\hat{r}_g$ are indicated in red and blue, respectively. Trait pairs with significant $\hat{r}_g$ at the nominal level 0.05 are marked by $*$. Trait pairs that remain to be significant after Bonferroni correction with $p \leq 0.05/325$ are marked by $**$. (**b**) Estimates of c_{12} between 26 complex traits. Positive and negative estimates of c_{12} are shown in purple and blue, respectively. Trait pairs with significant $\hat{c}_{12}$ at the nominal level 0.05 are marked by $*$. Trait pairs remain to be significant after Bonferroni correction with $p \leq 0.05/325$ are marked by $**$. (**c**) Causal relationships detected by MR-APSS. The positive and negative estimates of causal effects of the exposure on the outcome are indicated by red up-pointing triangles and blue down-pointing triangles, respectively. (**d**) The Venn diagram plot shows the causal effects detected by MR-APSS, CAUSE and Egger after Bonferroni correction.

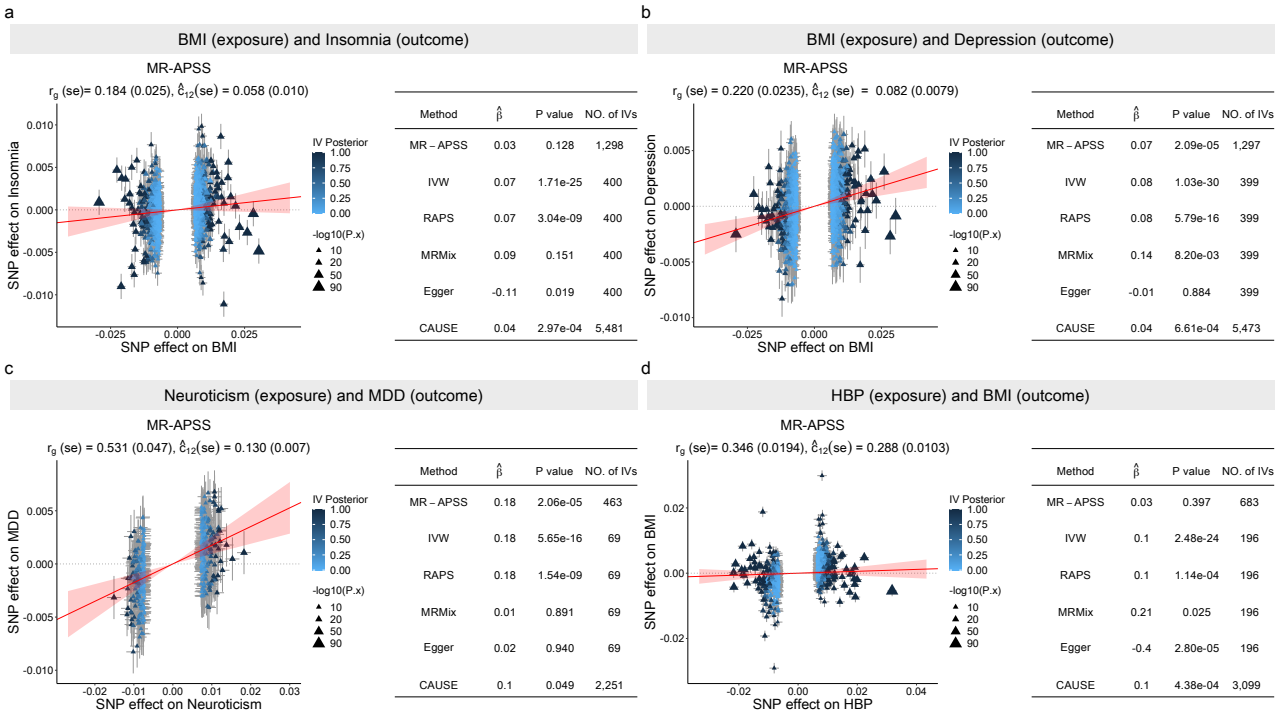

Figure 6: **Four examples of MR-APSS analysis results**. The left panel of each subplot shows the result of MR-APSS with its default IV threshold $p = 5 \times 10^{-5}$. The estimated causal effect is indicated by a red line with its 95% confidence interval indicated by the shaded area in transparent red color. Triangles indicate the observed SNP effect sizes ($\hat{\gamma}_j$ and $\hat{\Gamma}_j$). The color of triangles indicates the posterior of a valid IV, i.e., the posterior of an IV carrying the foreground signal ($Z_j = 1$, dark blue) or not ($Z_j = 0$, light blue). The precise definition of the posterior is given in the Method section. The right panel in each subplot shows the estimated causal effect $\hat{\beta}$, p-value as well as the number of IVs used by the six MR methods. (**a**) BMI and Insomnia. (**b**) BMI and Depression. (**c**) Neuroticism and MDD. (**d**) HBP (high blood pressure) and BMI.

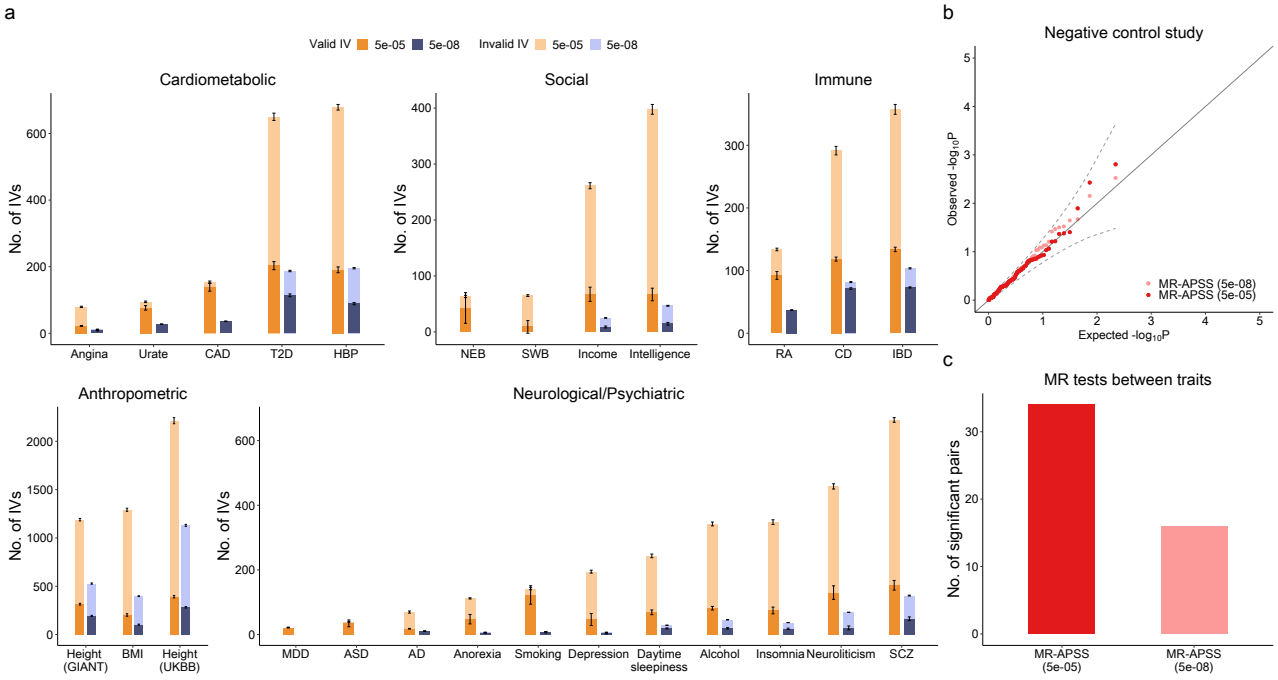

Figure 7: **Comparison of MR-APSS using IV thresholds $p = 5 \times 10^{-5}$ and $p = 5 \times 10^{-8}$.** (**a**) The average estimated number of valid IVs (dark color) and invalid IVs (light color) for traits from each category using IV thresholds $p = 5 \times 10^{-5}$ and $p = 5 \times 10^{-8}$. The number of valid IVs for one trait pair is the number of IVs that carry the foreground signal, as defined in the Method section. Clearly, the number of valid IVs increases as the IV threshold becomes looser. (**b**) Quantile-quantile plots of $-\log_{10}(p)$-values from MR-APSS for MR tests between 26 complex traits and negative control outcomes, using IV thresholds $p = 5 \times 10^{-5}$ and $p = 5 \times 10^{-8}$. MR-APSS produces well calibrated p-values for both loose and stringent IV thresholds. (**c**) The number of significant pairs detected by MR-APSS using different IV thresholds when applied to infer causal relationship among 26 complex traits. The statistical power of MR-APSS increases by including more IVs with a looser threshold.

IVs involved with a looser IV threshold, the number of invalid IVs increases because they are prone to violation of MR assumptions. Moreover, the number of IVs with moderate effects increases, bias due to confounding becomes increasingly influential. Without accounting for pleiotropy and sample structure, the type I error rate of MR methods can be more inflated when using a looser IV threshold. To have a better understanding of how MR-APSS works with the selected IVs, we examined the results of MR-APSS using a loose IV threshold ($p = 5 \times 10^{-5}$) as well as the results using a stringent IV threshold ($p = 5 \times 10^{-8}$). As we show later, the proposed background-foreground model can include more IVs with moderate effects to improve statistical power without inflated type I errors. On one hand, the number of selected IVs increased dramatically as the IV threshold become looser. However, most of IVs were detected by MR-APSS as invalid IVs (Fig. 7a). Since MR-APSS only uses the valid instrument strength in the foreground model for causal inference, the type I error will not be inflated when more invalid IVs are included. As shown in Fig. 7b, the p-values from MR-APSS for trait pairs between 26 complex traits and five negative control outcomes remain to be well calibrated using IV thresholds at 5×10^{-5} and 5×10^{-8}. These results confirm that the type I error of MR-APSS

is insensitive to the IV threshold. On the other hand, MR-APSS allows to use a looser IV threshold to include more IVs with moderate effects, and thus increase its statistical power (Fig. 7c). When investigating the causal relationship among 26 complex traits, the number of valid IVs increased a lot by changing the IV threshold from 5×10^{-8} to 5×10^{-5}. Clearly, the social and neurological/psychiatric traits can benefit a lot from this property. Despite the large sample sizes for these traits, the number of IVs is too small to perform powerful MR analysis when using the IV threshold $p = 5 \times 10^{-8}$ (Fig. 7a). For example, Depression only had seven IVs using a stringent IV threshold $p = 5 \times 10^{-8}$. Due to limited number of IVs, MR-APSS could not detect a significant causal effect of Depression on Insominia ($\hat{\beta} = 0.197$, s.e. $= 0.214$, p-value $= 0.358$). With a looser IV threshold $p = 5 \times 10^{-5}$, the number of IVs increased to 197 and the number of valid IVs was estimated around 71. As a result, MR-APSS detected a significant causal relationship between Depression and Insomnia ($\hat{\beta} = 0.569$, s.e. $= 0.139$, p-value $= 4.38 \times 10^{-5}$).

It is important to note that the correction of bias due to IV selection allows MR-APSS to include more SNPs with moderate effects. In the context of GWAS, the bias induced by selecting SNPs with different p-value thresholds is known as winner's curse [29, 43]. In the presence of winner's curse, the magnitude of the true effect of an IV is typically overestimated and the bias for SNPs with moderate effects tends to be much more severe than SNPs with large effects [44]. Without accounting for winner's curse in MR-APSS, IV strength would be falsely amplified. As a result, more SNPs from the background component would be falsely detected as valid IVs carrying foreground signals. When they are used for causal inference as valid IVs, the type I error rate increases dramatically. To verify this, we modified MR-APSS to ignore the winner's curse and applied this modified version to infer the causal relationships between 26 complex traits and the five negative control outcomes. We changed the IV thresholds from 5×10^{-5} to 5×10^{-8}. Without accounting for winner's curse, the proportions of valid IVs estimated by MR-APSS were close to 1 (Supplementary Fig S19(a)), indicating that most of IVs were falsely detected as valid IVs. As a result, the p-values produced by MR-APSS became inflated (Supplementary Fig S19 (b-c)), and the inflation was more severe at a looser IV threshold $p = 5 \times 10^{-5}$ where more SNPs with moderate effects were falsely detected as valid IVs.

Discussion

Through comprehensive simulations and real data analysis, we have demonstrated that MR-APSS can be applied to infer the causal relationship between a wide spectrum of exposure and outcome traits. The major advance of MR-APSS for causal inference can be attributed to the background-foreground model, where the background model accounts for confounding due to pleiotropy and sample structure, and the foreground model safely makes use of the valid signal for causal inference under the classical causal diagram. With the background-foreground model, MR-APSS allows to include more IVs with moderate effects to effectively improve statistical power without inflated type I errors.

A salient feature of MR-APSS is that it explicitly accounts for sample structure, including population stratification, cryptic relatedness and sample overlap. In particular, we find that population stratification is a major confounding factor in MR analysis. Recently, accumulating evidence suggests that population stratification in the UK Biobank can be driven by socioe-

conomic status [26] or geographic structure [27], which cannot be accounted for by routine adjustment (e.g., using principal components). As a result, the publicly available GWAS summary statistics still suffer from confounding effects [18]. In our real data analysis, we have also shown that causal relationship among psychiatric/neurological disorders, social traits, and anthropometric traits are often confounded by sample structure. Examples include Height (UKB), Height (GIANT), BMI and intelligence (see Fig. 4). However, this issue has been largely ignored by existing MR methods, resulting in many false positive findings.

We have also largely relaxed MR assumptions by leveraging genome-wide information. Many existing MR methods perform causal inference by only using information from the selected IVs. Due to limited information, they cannot easily distinguish causality from correlation without relying on strong assumptions, such as the InSIDE condition. To address this challenge, MR-APSS uses genome-wide information to estimate the background model. The benefits of pre-estimated background model are twofold. First, the pre-estimated background model can account for correlation between polygenic effects across the whole genome (u_j and v_j) and correlation between estimation errors (ϵ_j and ξ_j). This means that MR-APSS allows all SNPs to carry direct effects correlated with their instrument strength. Thus, the InSIDE condition is relaxed for all IVs. Second, it is helpful to ensure the model identifiability. With the pre-estimated background model, it is much easier to estimate the foreground model by using information from selected IVs. In contrast, CAUSE requires an additional condition to ensure its model identifiability: the portion of IVs with correlated pleiotropy to be small. This condition may not hold due to polygenic or omnigenic genetic architecture [45], where many genes jointly affect phenotypes through shared genetic pathways. From this perspective, relaxation of the InSIDE condition to all IVs is essential to address the complexity of human genetics.

In addition to pleiotropy and sample structure, one may wonder whether there exist some other major confounding factors in MR analysis. We believe that the influence of other confounding factors should be nearly ignorable. Under the LDSC assumptions, we exploit the LD structure of human genome to account for confounding factors, where pleiotropy and sample structure are captured by the first-order and zero-order terms of LD score, respectively. In this sense, what we ignored in MR-APSS is the confounding factors associated with higher-order terms of LD score. As shown in our real data analysis (Figs. 4-5), statistical inference based on MR-APSS is well calibrated, indicating the influence of other confounding factors should be very minor. Regarding the winner's curse, we have noticed that LD clumping may also potentially introduce some bias because SNPs with smaller p-values will be selected as IVs. Such a bias is very small when IV threshold $p \leq 5 \times 10^{-5}$. We can empirically correct the bias via adjusting the IV threshold by the ratio of the median after the LD clumping to the median before LD clumping (see the details in the Supplementary Note 1.8 and and Supplementary Figs S9- S10).

Despite the improvement of MR-APPSS over many existing MR methods, more research is needed for causal inference with genetic data. First, the background model is proposed to account for pleiotropy and sample structure hidden in GWAS of complex traits. The direct application of this model in some other contexts may not be suitable. For example, transcriptome-wide association studies (TWAS) have identified genetically regulated genes that are associated with complex diseases [46, 47, 48]. Recent studies have shown that a significant

proportion of phenotypic variation can be attributed to the genetically regulated gene expression [49, 50]. It is of great interest to infer the causal relationship between gene expression and complex diseases based on transcriptome-wide Mendelian randomization. However, it remains unclear what kind of signals should be represented as the background signals. The development of new statistical methods for transcriptome-wide Mendelian randomization is highly desirable, where cell-type shared or specific information may be helpful to define the background signal. Second, multivariate Mendelian randomization (MVMR) is drawing more and more attention [51, 52]. As some risk factors are known to a certain type of disease, it is more interesting to ask what other risk factors can be inferred conditioning on the known ones. For example, obesity is known to be a causal risk factor for CAD and T2D. The identification of some other causal risk factors for CAD and T2D through MVMR gains more biological and etiological understanding of these complex diseases.

We have mainly focused on comparing MR-APSS with methods using IVs. We note that causal inference can be performed without using IVs. A recently developed method is the latent causal variable (LCV) model [53]. We have also compared MR-APSS with LCV and found that LCV is often more conservative than MR-APSS (see details in the Supplementary Note 1.9 and Supplementary Figs S20-S21). In summary, we believe that MR-APSS will be widely applicable for causal inference, and we hope that MR-APSS can motivate more researchers to uncover more reliable causal relationships using rich genetic data resources.

Methods

The MR-APSS model. MR-APSS takes GWAS summary statistics $\{\hat{\gamma}_j, \hat{\Gamma}_j, \hat{s}_{X,j}, \hat{s}_{Y,j} \big| |\hat{\gamma}_j/\hat{s}_{X,j}| \geq t\}_{j=1,\ldots,M_t}$ as input to perform causal inference, where $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ are the estimated j-th SNP's effects on exposure X and outcome Y, respectively, and $\hat{s}_{X,j}$ and $\hat{s}_{Y,j}$ are their standard errors, $|\hat{\gamma}_j/\hat{s}_{X,j}| \geq t$ is the selection criterion to ensure that SNP j is associated with X, and M_t is the number of SNPs selected as IVs using a threshold t of z-values. To infer the causal effect β of exposure X on outcome Y, we propose to decompose the observed SNP effect sizes into background and foreground signals (Fig. 1):

$$\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} = Z_j \underbrace{\begin{pmatrix}\gamma_j \\ \beta\gamma_j + \alpha_j\end{pmatrix}}_{\text{Foreground}} + \underbrace{\begin{pmatrix}u_j \\ v_j\end{pmatrix}}_{\text{Pleiotropy}} + \underbrace{\begin{pmatrix}\epsilon_j \\ \xi_j\end{pmatrix}}_{\substack{\text{Sample structure} \\ \text{(Population stratification,} \\ \text{cryptic relatedness,} \\ \text{sample overlap etc.)}}} \quad , \quad j = 1, \cdots M_t, \tag{1}$$

where Pleiotropy + Sample structure (Population stratification, cryptic relatedness, sample overlap etc.) = Background

where u_j and v_j are the polygenic effects of SNP j on X and Y, ϵ_j and ξ_j are the estimation errors of SNP effect sizes, γ_j is the remaining SNP effect on exposure X as the instrument strength, α_j is the direct SNP effect on outcome Y, and Z_j is a Bernoulli variable indicating whether SNP j has a foreground component ($Z_j = 1$) or not ($Z_j = 0$).

The background model of MR-APSS. To model polygenic effects and their correlation induced by pleiotropy (Fig. 1b), we assume a variance component model

$$p\left(u_j, v_j | \mathbf{\Omega}\right) = \mathcal{N}\left(\begin{pmatrix}u_j \\ v_j\end{pmatrix} \Big| \mathbf{0}, \mathbf{\Omega}\right), \text{ with } \mathbf{\Omega} = \begin{pmatrix}\sigma_u^2 & r_g\sigma_u\tau_v \\ r_g\sigma_u\tau_v & \tau_v^2\end{pmatrix}, \tag{2}$$

where r_g is the genetic correlation induced by pleiotropic effects between X and Y, σ_u^2 and τ_v^2 are the variance of polygenic effects on X and Y, respectively. To account for bias and correlation in estimation errors due to sample structure (Fig. 1c), we consider the following model:

$$p\left(\epsilon_j, \xi_j | \mathbf{C}, \hat{\mathbf{S}}_j\right) = \mathcal{N}\left(\begin{pmatrix}\epsilon_j \\ \xi_j\end{pmatrix} \middle| \mathbf{0}, \hat{\mathbf{S}}_j \mathbf{C} \hat{\mathbf{S}}_j\right), \text{ with } \hat{\mathbf{S}}_j = \begin{pmatrix}\hat{s}_{X,j} & 0 \\ 0 & \hat{s}_{Y,j}\end{pmatrix}, \mathbf{C} = \begin{pmatrix}c_1 & c_{12} \\ c_{12} & c_2\end{pmatrix}, \quad (3)$$

where parameters c_1 and c_2 are used to adjust the bias in estimator errors, c_{12} accounts for the correlation between the estimation errors. In the presence of population stratification and cryptic relatedness, c_1 and c_2 will deviate from one (typically larger than one). Moreover, either population stratification or sample overlap can induce covariance between the estimation errors, resulting in nonzero c_{12}. Under the assumptions of LDSC [28], we can exploit the LD structure of human genome to account for confounding factors in the background model. Let $\ell_j = \sum_k r_{jk}^2$ be the LD score of SNP j, where r_{jk} is the correlation between SNP j and SNP k. The key idea to adjust LD effects is based on the following fact: the underlying true genetic effects are tagged by LD while the influence of sample structure is uncorrelated with LD. Then we show that our background model ($Z_j = 0$) can be written as (see details in Section 1.1 of Supplementary Note)

$$p(\hat{\gamma}_j, \hat{\Gamma}_j | \mathbf{\Omega}, \mathbf{C}, \hat{\mathbf{S}}_j, \ell_j) = \mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \mathbf{\Omega} + \hat{\mathbf{S}}_j \mathbf{C} \hat{\mathbf{S}}_j\right), \quad (4)$$

where pleiotropy and sample structure are captured by the first-order and zero-order terms of LD score, respectively. Therefore, both $\mathbf{\Omega}$ and $\mathbf{C}$ in the background model are pre-estimated by LDSC using genome-wide summary statistics (see details in Section 1.4 of Supplementary Note). As we see in real data analysis, genetic correlation of polygenic effects and correlation of estimation errors are two major confounding factors for causal inference.

The foreground model of MR-APSS. By accounting for confounding factors using the background model, we only need three mild assumptions on instrument strength γ_j and direct effect α_j to infer causal effect β, as shown in Fig. 1(a). First, there exist some nonzero values in $\{\gamma_j\}_{j=1,\ldots,M_t}$. Second, the strengths of instruments $\{\gamma_j\}_{j=1,\ldots,M_t}$ are independent of confounding factors. Third, the instrument strengths are independent of the direct effects, i.e., $(\gamma_1, \ldots, \gamma_{M_t}) \perp (\alpha_1, \ldots, \alpha_{M_t})$. Although our assumptions seem similar to those of existing methods, they are only imposed to the foreground signal and thus they are much weaker than existing MR methods. Specifically, we assume that γ_j and α_j are normally distributed and independent of each other:

$$p\left(\gamma_j, \alpha_j | \mathbf{\Sigma}\right) = \mathcal{N}\left(\begin{pmatrix}\gamma_j \\ \alpha_j\end{pmatrix} \middle| \mathbf{0}, \mathbf{\Sigma}\right), \text{ where } \mathbf{\Sigma} = \begin{pmatrix}\sigma^2 & 0 \\ 0 & \tau^2\end{pmatrix}. \quad (5)$$

The background-foreground model of MR-APSS. Now we combine the background model and the foreground model to characterize the observed SNP effect sizes $(\hat{\gamma}_j, \hat{\Gamma}_j)$. Let $\pi = p(Z_j = 1)$ be the probability that SNP j carries the foreground signal. Combining Eqs. (1,2,3,5) and integrating out γ_j, α_j, u_j, v_j, ϵ_j, ξ_j and Z_j, we have the following probabilistic

model:

$$\begin{aligned} &p(\hat{\gamma}_j, \hat{\Gamma}_j | \pi, \beta, \mathbf{\Sigma}, \mathbf{\Omega}, \mathbf{C}, \hat{\mathbf{S}}_j, \ell_j) \\ =&\pi \mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \mathbf{A}(\beta)\mathbf{\Sigma}\mathbf{A}(\beta)^T + \ell_j \mathbf{\Omega} + \hat{\mathbf{S}}_j \mathbf{C} \hat{\mathbf{S}}_j\right) + (1-\pi)\mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \mathbf{\Omega} + \hat{\mathbf{S}}_j \mathbf{C} \hat{\mathbf{S}}_j\right), \end{aligned} \tag{6}$$

where $\mathbf{A}(\beta) = \begin{pmatrix} 1 & 0 \\ \beta & 1 \end{pmatrix}$. A detailed derivation for Eq. (6) is given in Section 1.2 of the Supplementary Note.

Accounting for winner's curse. Recall that SNPs are selected based on a p-value threshold or equivalently a threshold t of z-score, i.e., $|\hat{\gamma}_j / \hat{s}_{X_j}| \geq t$. This selection process introduces non-ignorable bias, i.e., $E(\hat{\gamma}_j \Big| |\hat{\gamma}_j / \hat{s}_{X_j}| \geq t) \neq \gamma_j$, which has been known as winner's curse in GWAS [43, 29]. To correct the selection bias in MR, we further take into account the selection condition $|\hat{\gamma}_j / \hat{s}_{X_j}| \geq t$. After some derivations (Section 1.3 of the Supplementary Note), model (6) becomes mixture of truncated normal distributions:

$$\begin{aligned} &p\left(\hat{\gamma}_j, \hat{\Gamma}_j \middle| |\hat{\gamma}_j / \hat{s}_{X_j}| \geq t, \pi_t, \beta, \mathbf{\Sigma}, \mathbf{\Omega}, \mathbf{C}, \hat{\mathbf{S}}_j, \ell_j\right) \\ =&\pi_t \frac{\mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \mathbf{A}(\beta)\mathbf{\Sigma}\mathbf{A}(\beta)^T + \ell_j \mathbf{\Omega} + \hat{\mathbf{S}}_j \mathbf{C} \hat{\mathbf{S}}_j\right)}{2\Phi\left(\frac{-t\hat{s}_{X_j}}{\sqrt{\ell_j \sigma^2 + \ell_j \sigma_u^2 + \hat{s}_{X_j}^2}}\right)} + (1-\pi_t)\frac{\mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \mathbf{\Omega} + \hat{\mathbf{S}}_j \mathbf{C} \hat{\mathbf{S}}_j\right)}{2\Phi\left(\frac{-t\hat{s}_{X,j}}{\sqrt{\ell_j \sigma_u^2 + \hat{s}_{X,j}^2}}\right)}, \end{aligned} \tag{7}$$

where $\pi_t = p(Z_j = 1 \Big| |\hat{\gamma}_j / \hat{s}_{X_j}| \geq t)$ is the probability that the j-th SNP carries the foreground signal after selection.

Parameter estimation and statistical inference. In MR-APSS, the parameters of $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{C}}$ in the background model are estimated by LDSC using genome-wide summary statistics. Given $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{C}}$, the log-likelihood function of the observed data $\mathcal{D}_c = \{\hat{\gamma}_j, \hat{\Gamma}_j, \hat{s}_{X,j}, \hat{s}_{Y,j} \Big| |\hat{\gamma}_j / \hat{s}_{X,j}| \geq t\}_{j=1,\ldots,M_t}$ can be written as:

$$\begin{aligned} &L(\boldsymbol{\theta} | \mathcal{D}_c) \\ =&\sum_{j=1}^{M_t} \log \left[\pi_t \frac{\mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \mathbf{A}(\beta)\mathbf{\Sigma}\mathbf{A}(\beta)^T + \ell_j \hat{\mathbf{\Omega}} + \hat{\mathbf{S}}_j \hat{\mathbf{C}} \hat{\mathbf{S}}_j\right)}{2\Phi\left(\frac{-t\hat{s}_{X,j}}{\sqrt{\ell_j \sigma_u^2 + \ell_j \sigma^2 + \hat{c}_1 \hat{s}_{X,j}^2}}\right)} + (1-\pi_t)\frac{\mathcal{N}\left(\begin{pmatrix}\hat{\gamma}_j \\ \hat{\Gamma}_j\end{pmatrix} \middle| \mathbf{0}, \ell_j \hat{\mathbf{\Omega}} + \hat{\mathbf{S}}_j \hat{\mathbf{C}} \hat{\mathbf{S}}_j\right)}{2\Phi\left(\frac{-t\hat{s}_{X,j}}{\sqrt{\ell_j \sigma_u^2 + \hat{c}_1 \hat{s}_{X,j}^2}}\right)} \right], \end{aligned} \tag{8}$$

To obtain the maximum likelihood estimate of model parameters $\theta = \{\beta, \pi_t, \mathbf{\Sigma}\}$, we then derive an efficient expectation-maximization (EM) algorithm with details given in Section 1.5 of the Supplementary Note. As a byproduct, we can estimate the numbers of valid IVs and invalid IVs as $\hat{\pi}_t M_t$ and $(1-\hat{\pi}_t)M_t$, respectively. Real data results of the estimated numbers of valid and invalid IVs are shown in Fig.7a. The posterior of SNP j serving as a valid IV can be estimated as $p(\hat{Z}_j = 1 | \mathcal{D}_c)$, as shown in dark blue in Fig. 6. The likelihood ratio test can be conducted to examine the existence of the causal effect. Consider the following hypothesis test:

$$H_0 : \beta = 0 \quad \text{v.s.} \quad H_1 : \beta \neq 0, \tag{9}$$

the likelihood-ratio test statistic (LRT) is given by

$$T = 2\left(L(\hat{\boldsymbol{\theta}}|\mathcal{D}_c) - L(\hat{\boldsymbol{\theta}}_0|\mathcal{D}_c)\right), \tag{10}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_0$ are the parameter estimates obtained under hypotheses H_1 and H_0, respectively. Under the null hypothesis H_0, the test statistic T is asymptotically distributed as $\chi^2_{df=1}$ and its p-value can be obtained accordingly.

GWAS summary statistics and pre-processing. For all GWAS summary datasets, we used SNPs in the set of HapMap 3 list with minor allele frequency >0.05. We further excluded SNPs in the complex Major Histocompatibility Region (Chromosome 6, 26Mb−34Mb). Following the process in LDSC [54], we checked the χ^2 statistic of each SNP and excluded SNPs with $\chi^2 > \max\{80, N/1000\}$ to prevent the outliers that may unduly affect the results. For a pair of exposure and outcome traits, we took the overlapped SNPs from their GWAS summary statistics and aligned the sign of effect sizes for those SNPs to the same allele. Then we applied bivariate LDSC to estimate the $\boldsymbol{\Omega}$ and $\mathbf{C}$ using genome-wide summary statistics. After this step, we selected SNPs as IVs using an IV threshold (with the default p-value 5×10^{-5}), and applied PLINK clumping ($r^2 < 0.001$, window size 1Mb) to obtain nearly independent IVs. Finally, we fitted the proposed background-foreground model to infer the causal effect.

Acknowledgements

We thank Prof. Lin S. Chen, Prof. Lan Wang and Prof. Baolin Wu for their helpful comments and insightful discussions. This work is supported in part by Hong Kong Research Grant Council [16307818, 16301419, 16308120, 12303618], Hong Kong Innovation and Technology Fund [PRP/029/19FX], Hong Kong University of Science and Technology [startup grant R9405, Z0428 from the Big Data Institute], Key-Area Research and Development Program of Guangdong Province [2020B0101350001], and the Open Research Fund from Shenzhen Research Institute of Big Data [2019ORF01004]. The computational task for this work was partially performed using the X-GPU cluster supported by the RGC Collaborative Research Fund: C6021-19EF.

URLs

PLINK2 software: https://www.cog-genomics.org/plink2
MRMix R package: https://github.com/gqi/MRMix
CAUSE R package: https://github.com/jean997/cause
TwoSampleMR R package: https://github.com/MRCIEU/TwoSampleMR
Methods including IVW, Egger and RAPS are performed using TwoSampleMR R package.

Data availability

All the GWAS summary statistics used in this paper are public available. The URLs for downloading the datasets are summarized in Supplementary Table S1.

Code availability

The MR-APSS software is available at https://github.com/YangLabHKUST/MR-APSS.

References

[1] Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American journal of public health*, 95(S1):S144–S150, 2005.

[2] Lars Bondemark and Sabine Ruf. Randomized controlled trial: the gold standard or an unobtainable fallacy? *European Journal of Orthodontics*, 37(5):457–461, 2015.

[3] George Davey Smith and Shah Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

[4] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23, 2014.

[5] Jean-Baptiste Pingault, Paul F O'reilly, Tabea Schoeler, George B Ploubidis, Frühling Rijsdijk, and Frank Dudbridge. Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, 19(9):566–580, 2018.

[6] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665, 2013.

[7] Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

[8] Jean-Baptiste Pingault, Paul F. O'Reilly, Tabea Schoeler, George B. Ploubidis, Frühling Rijsdijk, and Frank Dudbridge. Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, 19(9):566–580, 2018.

[9] Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, 27(R2):R195–R208, 05 2018.

[10] Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human Molecular Genetics*, 27, 2018.

[11] Eleanor Sanderson, Tom G Richardson, Gibran Hemani, and George Davey Smith. The use of negative control outcomes in mendelian randomisation to detect potential population stratification or selection bias. *bioRxiv*, 2020.

[12] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[13] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 2013.

[14] Daniel M. Jordan, Marie Verbanck, and Ron Do. The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *SSRN Electronic Journal*, 2018.

[15] Stephen Burgess, Neil M. Davies, and Simon G. Thompson. Bias due to participant overlap in two-sample mendelian randomization. *Genetic Epidemiology*, 40(7), 2016.

[16] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[17] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015.

[18] Brendan Buliksullivan, Poru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

[19] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

[20] Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics*, 50(5):693, 2018.

[21] Zhihong Zhu, Zhili Zheng, Futao Zhang, Yang Wu, Maciej Trzaskowski, Robert Maier, Matthew R Robinson, John J McGrath, Peter M Visscher, Naomi R Wray, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature communications*, 9(1):224, 2018.

[22] Qingyuan Zhao, Jingshu Wang, Jack Bowden, and Dylan S Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652*, 2018.

[23] Jia Zhao, Jingsi Ming, Xianghong Hu, Jin Liu, and Can Yang. Bayesian weighted mendelian randomization for causal inference based on summary statistics. *arXiv preprint arXiv:1811.10223*, 2018.

[24] Guanghao Qi and Nilanjan Chatterjee. Mendelian randomization analysis using mixture models (mrmix) for genetic effect-size-distribution leads to robust estimation of causal effects. *bioRxiv*, page 367821, 2018.

[25] Jean Morrison, Nicholas Knoblauch, Joseph H. Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 2020.

[26] Abdel Abdellaoui, David Hugh-Jones, Loic Yengo, Kathryn E. Kemper, and Peter M. Visscher. Genetic correlates of social stratification in great britain. *Nature Human Behaviour*, 3(12), 2019.

[27] Simon Haworth, Ruth Mitchell, Laura Corbin, Kaitlin H. Wade, Tom Dudding, Ashley Budu-Aggrey, David Carslake, Gibran Hemani, Lavinia Paternoster, and George Davey and Smith. Apparent latent structure within the uk biobank sample has implications for epidemiological analysis. *Nature Communications*, 10(1), 2019.

[28] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[29] Sebastian Zollner and Jonathan K Pritchard. Overcoming the winner's curse : Estimating penetrance parameters from case-control data. *American Journal of Human Genetics*, 80(4):605–615, 2007.

[30] Stephen Burgess and Simon G. Thompson. Interpreting findings from mendelian randomization using the mr-egger method. *European Journal of Epidemiology*, 2017.

[31] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Zoltán Kutalik, Najaf Amin, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[32] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, and the GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in $\tilde{7}00000$ individuals of European ancestry. *Human Molecular Genetics*, 27(20):3641–3649, 08 2018.

[33] Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša UmićevićMirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348, 2019.

[34] Michael L Ganz, Neil Wintfeld, Qian Li, Veronica Alas, Jakob Langer, and Mette Hammer. The association of body mass index with the risk of type 2 diabetes: a case–control study nested in an electronic health records system in the united states. *Diabetology and Metabolic Syndrome*, 6(1):50, 2014.

[35] Kentaro Tanaka, Soshiro Ogata, Haruka Tanaka, Kayoko Omura, Chika Honda, Osaka Twin Research Group, and Kazuo Hayakawa. The relationship between body mass index and uric acid: a study on japanese adult twins. *Environmental health and preventive medicine*, 20(5):347–353, 09 2015.

[36] Sadiya S. Khan, Hongyan Ning, John T. Wilkins, Norrina Allen, Mercedes Carnethon, Jarett D. Berry, Ranya N. Sweis, and Donald M. Lloyd-Jones. Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. *JAMA Cardiology*, 2018.

[37] Amy E Taylor, Rebecca C Richmond, Teemu Palviainen, Anu Loukola, Robyn E Wootton, Jaakko Kaprio, Caroline L Relton, George Davey Smith, and Marcus R Munafò. The effect of body mass index on smoking behaviour and nicotine metabolism: a mendelian randomization study. *Human molecular genetics*, 28(8):1322–1330, 2019.

[38] Jessica Tyrrell, Anwar Mulugeta, Andrew R Wood, Ang Zhou, Robin N Beaumont, Marcus A Tuke, Samuel E Jones, Katherine S Ruth, Hanieh Yaghootkar, Seth Sharp, et al. Using genetics to understand the causal influence of higher bmi on depression. *International journal of epidemiology*, 48(3):834–848, 2019.

[39] Jessica Tyrrell, Samuel E Jones, Robin Beaumont, Christina M Astley, Rebecca Lovell, Hanieh Yaghootkar, Marcus Tuke, Katherine S Ruth, Rachel M Freathy, and Joel N Hirschhorn. Height, body mass index, and socioeconomic status: mendelian randomisation study in uk biobank. *Bmj British Medical Journal*, page i582, 2016.

[40] Lahey and B. Benjamin. Public health significance of neuroticism. *American Psychologist*, 64(4):241, 2009.

[41] Jim Van Os and Peter B. Jones. Neuroticism as a risk factor for schizophrenia. *Psychological Medicine*, 31(6):1129–34, 2001.

[42] ANNE, FARMER, KATE, REDMAN, TANYA, HARRIS, ARSHAD, MAHMOOD, STEPHANIE, and SADLER and. Neuroticism, extraversion, life events and depression. *British Journal of Psychiatry*, 2002.

[43] John Ferguson, Judy H Cho, Can Yang, and Hongyu Zhao. Empirical bayes correction for the winner's curse in genetic association studies. *Genetic Epidemiology*, 37(1):60–68, 2013.

[44] Hua Zhong and Ross L Prentice. Correcting "winner's curse" in odds ratios from genome wide association findings for major complex human diseases. *Genetic Epidemiology*, 34(1):78–91, 2009.

[45] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[46] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[47] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[48] Can Yang, Xiang Wan, Xinyi Lin, Mengjie Chen, Xiang Zhou, and Jin Liu. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, 35(10):1644–1652, 2019.

[49] Mingxuan Cai, Lin S Chen, Jin Liu, and Can Yang. IGREX for quantifying the impact of genetically regulated expression on phenotypes. *NAR genomics and bioinformatics*, 2(1):lqaa010, 2020.

[50] Douglas W Yao, Luke J O'Connor, Alkes L Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, pages 1–8, 2020.

[51] Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International journal of epidemiology*, 48(3):713–727, 2019.

[52] Verena Zuber, Johanna Maria Colijn, Caroline Klaver, and Stephen Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nature communications*, 11(1):1–11, 2020.

[53] Luke, J, O'Connor, Alkes, L, and Price. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics*, 2018.

[54] Brendan Buliksullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Poru Loh, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015.