Testing and controlling for horizontal pleiotropy with the probabilistic Mendelian randomization in transcriptome-wide association studies

Zhongshang Yuan[1,2], Huanhuan Zhu[2], Ping Zeng[3], Sheng Yang[2], Shiquan Sun[2], Can Yang[4], Jin Liu[5], Xiang Zhou[2,6,*]

1. Department of Biostatistics, School of Public Health, Shandong University, Jinan, China.
2. Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
3. Department of Epidemiology and Biostatistics, Xuzhou Medical University, Xuzhou, Jiangsu, China
4. Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China
5. Centre for Quantitative Medicine, Program in Health Services and Systems Research, Duke-NUS Medical School, Singapore 169857
6. Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed: xzhousph@umich.edu

Abstract

Integrating association results from both genome-wide association studies (GWASs) and expression quantitative trait locus (eQTL) mapping studies has the potential to shed light on the molecular mechanism underlying disease etiology. Several statistical methods have been recently developed to integrate GWASs with eQTL studies in the form of transcriptome-wide association studies (TWASs). These existing methods can all be viewed as two sample Mendelian randomization (MR) methods, which are also widely used in various GWASs for inferring the causal relationship among complex traits. Unfortunately, most existing TWAS and MR methods make an unrealistic modeling assumption that the instrumental variables do not exhibit horizontal pleiotropic effects. However, horizontal pleiotropic effects have been recently discovered to be wide spread across complex traits, and as we will show here, are also wide spread across gene expression traits. Therefore, allowing for no horizontal pleiotropic effects can be overall restrictive, and, as we will be show here, can lead to a substantial inflation of test statistics and subsequently false discoveries in TWAS applications. Here, we present a probabilistic MR method, which we refer to as PMR-Egger, for testing and controlling of horizontal pleiotropic effects in TWAS applications. PMR-Egger relies on a new MR likelihood framework that unifies many existing TWAS and MR methods, accommodates multiple correlated instruments, tests the causal effect of gene on trait in the presence of horizontal pleiotropy, and, with a newly developed parameter expansion version of the expectation maximization algorithm, is scalable to hundreds of thousands of individuals. With extensive simulations, we show that PMR-Egger provides calibrated type I error control for causal effect testing in the presence of horizontal pleiotropic effects, is reasonably robust for various types of horizontal pleiotropic effect mis-specifications, is more powerful than existing MR approaches,

43 and, as a by-product, can directly test for horizontal pleiotropy. We illustrate the benefits of our
44 method in applications to 39 diseases and complex traits obtained from three GWASs including
45 the UK Biobank and show how PMR-Egger can lead to new biological discoveries through
46 integrative analysis.
47
48

Introduction

Genome-wide association studies (GWASs) have identified many SNPs associated with common diseases or disease related traits. Parallel expression quantitative trait loci (eQTL) mapping studies have also identified many cis-acting SNPs associated with gene expression level of nearby genes. Integrating association results from both GWASs and eQTL mapping studies has the potential to shield light on the molecular mechanism underlying disease etiology. Several methods have been recently proposed to integrate GWASs with eQTL mapping studies. For example prediXcan[1] proposes to perform a weighted SNP set test in GWAS by inferring SNP weights from eQTL studies. TWAS[2] proposes to infer the association between gene expression and disease trait by leveraging the shared common set of cis-SNPs. SMR[3] or GSMR[4] directly tests the causal association between gene expression and disease trait under a Mendelian randomization (MR) framework through selecting a single instrument or multiple independent or near independent instruments. While each of these integrative methods was originally proposed to solve a different problem, all the aforementioned methods can be viewed as a two-sample MR method but with different modeling assumptions, aiming to identify genes causally associated with complex traits or diseases in the context of transcriptome-wide association studies (TWAS).

MR analysis is a form of instrumental variable analysis that was originally developed in the field of causal inference[5]. MR aims to determine the causal relationship between an exposure variable (e.g. gene expression) and an outcome variable (e.g. complex trait) in observational studies. MR treats SNPs as instrumental variables for the exposure variable of interest and uses these SNP instruments to estimate and test the causal effect of the exposure variable on the outcome variable. MR methods have been widely applied to investigate the causal relationship among various complex traits[6-9], and, through a two-sample design, can be easily adapted to

72 settings where the exposure and outcome are measured on two different sets of individuals[10,11].
73 However, MR analysis for TWAS is not straightforward and requires the development of new
74 methods that can accommodate two important TWAS data features.

75 First, both GWASs and eQTL mapping studies collect multiple SNPs that are in high linkage
76 disequilibrium (LD) with each other. Traditional MR methods, such as the random effects
77 version or the fixed effect version of the inverse variance weighted regression[12], MR-Egger[13],
78 median-based regression[14], SMR[3], or GSMR[4], are only applicable to a single SNP instrument or
79 multiple independent/near-independent SNP instruments. Handling only independent SNPs is
80 restrictive, as most exposure variables/molecular traits are polygenic/omni-genic and are
81 influenced by multiple SNPs that are in potential LD with each other. Indeed, incorporating
82 multiple correlated SNPs often can help explain a greater proportion of variance in the exposure
83 variable than using independent SNPs, and thus can help increase power and improve estimation
84 accuracy of MR analysis[5,15-17]. Due to the benefits of multiple correlated instruments, both
85 prediXcan and TWAS, rely on either ElasticNet[18] or BSLMM[19], incorporate all cis-SNPs that are
86 in high LD for TWAS applications, which, as we will show below, leads to a substantial power
87 improvement over standard MR approaches that use independent SNPs. Unfortunately, both
88 prediXcan and TWAS rely on a two-stage MR inference procedure, estimating SNP effect sizes
89 in the exposure study and plugging in these estimates to the outcome study for causal effect
90 inference. Two-stage inference procedure in MR fails to account for the uncertainty in parameter
91 estimates in the exposure study and can often lead to biased causal effect estimates and power
92 loss especially in the presence of weak instruments[5,16]. Indeed, similar to what have been
93 observed in the MR filed, our previous study also suggest that likelihood based inference can

substantially improve power for TWAS[20]. Therefore, it is important to incorporate multiple correlated instruments in a likelihood inference framework for MR analysis in TWAS.

Second, perhaps more importantly, SNP instruments often exhibit pervasive horizontal pleiotropic effects[21]. Horizontal pleiotropy occurs when a genetic variant affects the outcome variable through pathways other than or in addition to the exposure variable[22]. Horizontal pleiotropy is in contrast with vertical pleiotropy, which characterizes instrument effects on the outcome variable through the path of the exposure. Horizontal pleiotropy is widely distributed across the genome, affects a wide spectrum of complex traits, and can be driven by LD and extreme polygenicity of traits[21 23]. Despite its wide prevalence, however, only a limited number of MR methods have been developed that can test and control for horizontal pleiotropy; even fewer are applicable for TWAS applications. For example, some existing methods (e.g. Cohran's Q statistic[24,25] and MR-PRESSO[21]) test for horizontal pleiotropic effects without directly controlling for them. Some methods (e.g. CaMMEL[26]) control for horizontal pleiotropic effects without directly testing them[27,28]. Some methods (e.g. Egger regression[13,29], GLIDE[30], GSMR[4], MR-median method[14], profile score approach[31], MRMix[32] and Bayesian MR[33,34]) test and control for horizontal pleiotropic effects, but can only accommodate independent or near-independent instruments. As far as we are aware, there is only one two-sample MR method currently developed for testing and controlling for pleiotropic effects in the presence of correlated instruments: LDA MR-Egger[35]. Unfortunately, as we will show below, LDA MR-Egger cannot handle realistic LD pattern among cis-SNPs for TWAS applications.

Here, we develop a generative two-sample MR method in a likelihood framework, which we refer to as the probabilistic two-sample Mendelian randomization (PMR), to perform MR analysis using multiple correlated instruments for TWAS applications. We illustrate how the

PMR framework can facilitate the understanding of many existing MR approaches as well as many existing integrative analysis approaches. Within the PMR framework, we focus on a particular horizontal pleiotropy effect modeling assumption based on the burden test assumption commonly used for rare variant test, which, as we will show later, effectively generalizes the Egger regression assumption commonly used for MR analysis to correlated instruments. Our method allows us to test the causal effect in the presence of horizontal pleiotropy, and, with a parameter expansion version of the expectation maximization algorithm (PX-EM), is scalable to hundreds of thousands of individuals. We refer to our method as PMR-Egger. With simulations, we show that PMR-Egger provides calibrated type I error for causal effect testing in the presence of horizontal pleiotropic effects, is more powerful than existing MR approaches, and, as a by-product, can directly test for horizontal pleiotropy. We apply our method to perform TWAS for 39 diseases and complex traits obtained from three GWASs with sample size ranging from 4,686 to 337,199.

Methods

PMR-Egger Overview

We consider a probabilistic Mendelian randomization framework for performing two-sample Mendelian randomization analysis with correlated SNP instruments. Two-sample Mendelian randomization analysis aims to estimate and test for the causal effect of an explanatory variable on an outcome variable in the setting where the exposure and outcome variables are measured in two separate studies with no sample overlap. In the TWAS applications we consider here, the explanatory variable is gene expression level that is measured in a gene expression study, while the outcome variable is a quantitative trait or a dichotomous disease status that is measured in a GWAS. Often times, in such TWAS setting, the gene expression study and GWAS are performed on two separate samples. While we mostly focus on TWAS applications in the present study, we note that two-sample Mendelian randomization is also commonly performed in settings where both the explanatory and outcome variables are complex traits that are measured in two separate GWASs. An illustrative diagram of MR analysis is displayed in Supplementary Fig. 1.

We denote $\boldsymbol{x}$ as an n_1-vector of exposure variable (i.e. gene expression measurements) that is measured on n_1 individuals in the gene expression study and denote $\mathbf{Z}_\mathrm{x}$ as an n_1 by p matrix of genotypes for p selected instruments (i.e. cis-SNPs) in the same study. We denote $\mathbf{y}$ as an n_2-vector of outcome variable (i.e. trait) that is measured on n_2 individuals in the GWAS and denote $\mathbf{Z}_y$ as an n_2 by p matrix of genotypes for the same p selected instruments there. We consider three linear regressions to model the two studies separately

$$\boldsymbol{x} = \mathbf{1}_{\mathrm{n}_1}\mu_\mathrm{x} + \mathbf{Z}_\mathrm{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathrm{x} \quad (1)$$

$$\tilde{\boldsymbol{x}} = \mathbf{1}_{\mathrm{n}_2}\mu_\mathrm{x} + \mathbf{Z}_\mathrm{y}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\tilde{x}} \quad (2)$$

$$\boldsymbol{y} = \mathbf{1}_{n_2}\mu_y + \tilde{\boldsymbol{x}}\alpha + \mathbf{Z}_y\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad (3)$$

where the equation (1) is for the gene expression data and the equations (2)-(3) are for the GWAS data. Here, μ_x and μ_y are the intercepts; $\tilde{\boldsymbol{x}}$ is an unobserved n_2-vector of exposure variable on the n_2 individuals in the GWAS; $\boldsymbol{\beta}$ is a p-vector of instrumental effect sizes on the explanatory variable; α is a scalar that represents the causal effect of the exposure variable on the outcome variable; $\boldsymbol{\gamma}$ is a p-vector of horizontal pleiotropic effect sizes of p instruments on the outcome variable; $\boldsymbol{\varepsilon}_x$ is an n_1-vector of residual error with each element independently and identically distributed from a normal distribution $N(0, \sigma_x^2)$; $\boldsymbol{\varepsilon}_{\tilde{x}}$ is an n_2-vector of residual error with each element independently and identically distributed from the same normal distribution $N(0, \sigma_x^2)$; and $\boldsymbol{\epsilon}$ is an n_2-vector of residual error with each element independently and identically distributed from a normal distribution $N(0, \sigma_y^2)$. We note that while the above three equations are specified based on two separate studies, they are joined together with the common parameter $\boldsymbol{\beta}$ and the unobserved gene expression measurements $\tilde{\boldsymbol{x}}$. Note that equations (2)-(3) can be combined into

$$\boldsymbol{y} = \mathbf{1}_{n_2}\mu_y + \mathbf{Z}_y\boldsymbol{\beta}\alpha + \mathbf{Z}_y\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_y \quad (4)$$

where $\boldsymbol{\varepsilon}_y = \boldsymbol{\varepsilon}_{\tilde{x}}\alpha + \boldsymbol{\epsilon}$.

Our key parameter of interest in the above joint model is the causal effect α. The causal interpretation of α requires two assumptions of MR analysis to hold: (i) instruments are associated with the exposure; (ii) instruments are not associated with any other confounders that may be associated with both exposure and outcome. Note that our model no longer requires the exclusion restriction condition of traditional MR (i.e. instruments only influence the outcome by the path of exposure), as we directly model horizontal pleiotropy effects $\boldsymbol{\gamma}$. With the above two assumptions of MR analysis, we derive the causal interpretation and identification of α under the

framework of decision-theoretic causal inference[33,36-38] (details in Supplementary Note). Because the causal effect interpretation of α depends on the two MR assumptions, at least one of which is often not testable in practice, MR analysis in observational studies likely provides weaker causality evidence than randomized clinical trials. Therefore, while we follow standard MR analysis and use the term "causal effect" through the text, we only intend to use this term to emphasize the fact that α estimate from an MR analysis is more trustworthy than the effect size estimate in a standard linear regression of $\boldsymbol{y}$ on $\tilde{\boldsymbol{x}}$.

Because p is often larger than n_1, we will need to make additional modeling assumptions on $\boldsymbol{\beta}$ to make the model identifiable. In addition, the two instrumental effect terms defined in equation (4), the vertical pleiotropic effect $\mathbf{Z}_y\boldsymbol{\beta}\alpha$ and the horizontal pleiotropic effect $\mathbf{Z}_y\boldsymbol{\gamma}$, are also not identifiable from each other, unless we make additional modeling assumptions on $\boldsymbol{\gamma}$. Here, we follow standard polygenic model and assume that all elements in $\boldsymbol{\beta}$ are non-zero and that each follows a normal distribution $\mathrm{N}(0, \sigma_\beta^2)$. In addition, we follow the burden test assumption commonly used for rare variant test and assume that $\boldsymbol{\gamma}_j = \gamma$ for $j = 1, \ldots p$. With the burden test assumption on $\boldsymbol{\gamma}$, in the special case where instruments are independent and treated as fixed effects and where a two-stage estimation procedure is used for inference, our model reduces to the commonly used MR-Egger regression model. Therefore, our model effectively generalizes the Egger regression model to a likelihood inference framework with correlated instruments.

We are interested in estimating the causal effect α and testing the null hypothesis $\mathrm{H}_0: \alpha = 0$ in the presence of horizontal pleiotropy $\boldsymbol{\gamma}$. In addition, we are interested in estimating the horizontal pleiotropic effect size γ and testing the null hypothesis $\mathrm{H}_0: \gamma = 0$. We accomplish both tasks through the maximum likelihood inference framework. In particular, we develop an expectation maximization (EM) algorithm for parameter inference by maximizing the joint likelihood

defined based on equations (1) and (4) (details in the Supplementary Note). The EM algorithm allows us to obtain the maximum likelihood of the joint model, together with maximum likelihood estimates for both α and γ. In addition, we apply the EM algorithm to two reduced models, one without α and the other without γ, to obtain the corresponding maximum likelihoods. Afterwards, we perform likelihood ratio tests for either $\mathrm{H}_0: \alpha = 0$ or $\mathrm{H}_0: \gamma = 0$, by contrasting the maximum likelihood obtained from the joint model to that obtained from each of the two reduced models, respectively. We refer to the above inference procedure as probabilistic, as we place estimation and testing into a maximum likelihood framework. Our inference procedure is in contrast to the commonly used two-stage estimation procedure (as used in, for example, Egger regression[13,29], PrediXcan[1] and TWAS[2]), which estimates $\widehat{\boldsymbol{\beta}}$ from equation (1) first and then directly plug in the estimates into equation (4) for inference. The previous two-stage estimation procedure fails to properly account for the estimation uncertainty in $\widehat{\boldsymbol{\beta}}$ and is known to lose power compared to a formal likelihood inference procedure[5,16,20].

We refer to our model and algorithm together as the two-sample probabilistic Mendelian randomization with Egger regression (PMR-Egger). As explained above, we use "probabilistic" to refer to both the data generative model and the maximum likelihood inference procedure. We use "Egger" to refer to the horizontal pleiotropic assumption on $\boldsymbol{\gamma}$ that effectively generalizes the Egger-regression assumption to correlated instruments. We also note that the joint generative Mendelian randomization model defined in equations (1) and (4) is a useful conceptual framework that unifies many existing MR methods. In particular, almost all existing MR methods are built upon on the joint model, but with different modeling assumptions on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and with different inference procedures (Table 1). Compared with these existing MR approaches,

PMR-Egger is capable of modeling multiple correlated instruments, effectively controls for horizontal pleiotropy, and places inference into a likelihood framework.

Simulations

We performed simulations to assess the performance of PMR-Egger and compare it with existing approaches. To do so, we first obtained 556 cis-SNPs for the gene *BACE1* on chromosome 11 from the GEUVADIS data[39] (data processing details in the next section) and simulated gene expression values. We used the gene *BACE1* because the number of cis-SNPs in this gene represents the median of that across all genes. With the scaled genotype data $\boldsymbol{Z}_x$, we simulated SNP effect sizes $\boldsymbol{\beta}$ from a normal distribution $N(0, PVE_{\mathrm{zx}}/556)$ and summed the genetic effects across all cis-SNPs as $\boldsymbol{Z}_x\boldsymbol{\beta}$. We also simulated residual errors $\boldsymbol{\varepsilon}_{\mathrm{x}}$ from a normal distribution $N(0, 1 - PVE_{\mathrm{zx}})$. The variances of the two normal distributions are chosen so that in expectation, the proportion of phenotypic variance explained by genetic effects is PVE_{zx}.

Next, we obtained genotypes for the same 556 SNPs from 2,000 randomly selected control individuals in the Kaiser Permanente/UCSF Genetic Epidemiology Research Study on Adult Health and Aging (GERA)[40,41] and simulated a quantitative trait. Here, we directly used $\boldsymbol{\beta}$ from the gene expression data, which, when paired with causal effect α, yielded the vertical pleiotropic effects $\alpha\boldsymbol{\beta}$. We set $\alpha = \sqrt{PVE_{\mathrm{zy}}/PVE_{\mathrm{zx}}}$, and we simulated residual errors $\boldsymbol{\varepsilon}_{\mathrm{y}}$ from a normal distribution $N(0, 1 - PVE_{\mathrm{zy}})$. These parameters were selected so that in expectation, in the absence of horizontal pleiotropic effects, the proportion of phenotypic variance explained by vertical pleiotropic effects is PVE_{zy}. Afterwards, we simulated horizontal pleiotropic effects $\boldsymbol{\gamma}$ for these SNPs (more details below). We summed the horizontal pleiotropic effects, vertical pleiotropic effects and residual errors to yield the simulated trait.

In the simulations, we first examined a baseline simulation setting where we set $PVE_{\mathrm{zx}} = 10\%$, $PVE_{\mathrm{zy}} = 0$, with all $\gamma_{\mathrm{j}} = 0$. On top of the baseline setting, we varied one parameter at a time to examine the influence of various parameters. For PVE_{zx}, we set it to be either 1%, 5% or 10%, close to the median gene expression heritability estimates across genes[42,43]. For $\boldsymbol{\beta}$, we examined alternative SNP effect size distributions that deviate from the polygenic assumption in the baseline setting. Specifically, we randomly selected either 1 SNP, 1%, 10% or 100% of the SNPs to have non-zero effect, while simulated their effects from a normal distribution to explain a fixed PVE_{zx}. For PVE_{zy}, we varied its value to be either 0% (for null simulations), 0.2%, 0.4% or 0.6% (for power simulations). For $\boldsymbol{\gamma}$, we randomly assigned a fixed proportion of γ_{j} to be non-zero (proportion equals 10%, 30%, 50%, or 100%). Afterwards, we set the absolute value of non-zero γ_{j} to be the same value of γ, but randomly assign some of their signs to be positive and some of their signs to be negative, with the ratio of positive effects to negative effects being either 1:9, 3:7, or 5:5. Here, we set γ to be 1×10^{-4}, 5×10^{-4}, 1×10^{-3} or 2×10^{-3}, which corresponds to the 50%, 70%, 90%, 95% quantiles of horizontal pleiotropic effect estimates across all genes and all traits in the WTCCC data (more details below), respectively.

For null simulations and type I error control examination, we performed 10,000 simulation replicates for each simulation scenario described above. For power calculation, for each scenario, we performed 1,000 alternative simulations together with 9,000 null simulations and calculated power based on false discovery rate (FDR).

Real Data Applications

We applied our method to perform TWAS by integrating gene expression data with several GWASs. Specifically, we obtained GEUVADIS data[39] as the gene expression data and examined 39 phenotypes from three GWASs. The three GWASs include the Wellcome trust case control

264 study (WTCCC)[44], the Kaiser Permanente/UCSF Genetic Epidemiology Research Study on
265 Adult Health and Aging (GERA)[40,41], and the UK Biobank[45].

266 The GEUVADIS data[39] contains gene expression measurements for 465 individuals collected
267 from five different populations that include CEPH (CEU), Finns (FIN), British (GBR), Toscani
268 (TSI) and Yoruba (YRI). In the expression data, we only focused on protein coding genes and
269 lincRNAs that are annotated in GENCODE [46,4746; 47] (release 12). Among these genes, we
270 removed lowly expressed genes that have zero counts in at least half of the individuals to obtain
271 a final set of 15,810 genes. We performed PEER normalization to remove confounding effects
272 and unwanted variations following previous studies[48,49]. Afterwards, following[49], to remove
273 remaining population stratification, we quantile normalized the gene expression measurements
274 across individuals in each population to a standard normal distribution, and then further quantile
275 normalized the gene expression measurements to a standard normal distribution across
276 individuals from all five populations. Besides expression data, all individuals in GEUVADIS
277 also have their genotypes sequenced in the 1,000 Genomes Project. We obtained genotype data
278 from the 1,000 Genomes Project phase 3. We filtered out SNPs that have a Hardy-Weinberg
279 equilibrium (HWE) p-value $< 10^{-4}$, a genotype call rate <95%, or a minor allele frequency (MAF)
280 <0.01. We retained a total of 7,072,917 SNPs for analysis.

281 The WTCCC data consists of about 14,000 cases from seven common diseases and 2,938
282 shared controls[44]. The diseases include type 1 diabetes (T1D; *n*=1,963), Crohn's disease (CD;
283 *n*=1,748), rheumatoid arthritis (RA; *n*=1,861), bipolar disorder (BD; *n*=1,868), type 2 diabetes
284 (T2D; *n*=1,924), coronary artery disease (CAD; *n*=1,926), and hypertension (HT; *n*=1,952). We
285 obtained quality controlled genotypes from WTCCC and initially imputed missing genotypes
286 using BIMBAM[50] to arrive at a total of 458,868 SNPs shared across all individuals. Afterwards,

287 we further imputed SNPs using the 1,000 Genomes as the reference panel with SHAPEIT and
288 IMPUTE2[51]. We filtered out SNPs that have an HWE p-value < 10^{-4}, a genotype call rate <95%,
289 or an MAF<0.01 to obtain a total of 2,793,818 imputed SNPs.

290 The GERA study consists of 61,953 individuals and 675,367genotyped SNPs. We filtered out
291 SNPs that had a genotype calling rate below 0.95, MAF<0.01, or HWE p value<10^{-4} to yield a
292 total of 487,609 SNPs. We then phased genotypes using SHAPEIT[52] and imputed SNPs based on
293 the Haplotype Reference Consortium (HRC version r1.1) reference panel[53] on the Michigan
294 Imputation Server using Minimac3[54]. Afterwards, we further filtered out SNPs that have a HWE
295 p-value < 10^{-4}, a genotype call rate <95%, an MAF<0.01, or an imputation score<0.30 to arrive
296 at a total of 8,385,867 SNPs shared across 61,953 individuals. We examined 22 diseases in
297 GERA that include Asthma (number of cases n=10,101), Allergic Rhinitis (n=15,193),
298 Cardiovascular Disease (CARD, n=16,431), Cancers (n=18,714), Depressive Disorder (n=7,900),
299 Dermatophytosis (n=8,443), Type 2 Diabetes (T2D, n=7,638), Dyslipidemia (n=33,071),
300 Hypertension (HT, n=31,044), Hemorrhoids (n=9,922), Abdominal Hernia (n=6,876), Insomnia
301 (n=4,357), Iron Deficiency (n=2,706), Irritable Bowel Syndrome (n=3,367), Macular
302 Degeneration (n=4,031), Osteoarthritis (n=22,062), Osteoporosis (n=5,909), Peripheral Vascular
303 Disease (PVD, n=4,718), Peptic Ulcer (n=1,007), Psychiatric disorders (n=9408), Stress
304 Disorders (n=4,706), and Varicose Veins (n=2,714).

305 The UK Biobank data consists of 487,409 individuals and 92,693,895 imputed SNPs[45]. We
306 followed the same sample QC procedure in Neale lab
307 (https://github.com/Nealelab/UK_Biobank_GWAS/tree/master/imputed-v2-gwas) to retain a
308 total of 337,199 individuals of European ancestry. In addition, we filtered out SNPs with an
309 HWE p-value < 10^{-7}, a genotype call rate <95%, or an MAF<0.001 to obtain a total of

13,876,958 SNPs. We selected 10 UK Biobank quantitative traits that have a phenotyping rate > 80%, a SNP-heritability > 0.2 and a low correlation among them following a previous study[55]. The 10 traits include Height ($h^2 = 0.579$;), Platelet count ($h^2 = 0.404$), Bone mineral density ($h^2 = 0.401$), Red blood cell count ($h^2 = 0.324$), FEV1-FVC ratio ($h^2 = 0.313$), Body mass index (BMI, $h^2 = 0.308$), RBC distribution width ($h^2 = 0.288$), Eosinophils count ($h^2 = 0.277$), Forced vital capacity ($h^2 = 0.277$), White blood cell count ($h^2 = 0.272$). For each trait in turn, we standardized the phenotype to have a mean of zero and a standard deviation of one, regressed the resulting standardized phenotypes on sex and top 10 genotype principal components (PCs) to obtain the residuals, and finally used the scaled residuals to conduct TWAS analysis.

In the GEUVADIS data, for each gene in turn, we extracted cis-SNPs that are within either 100 kb upstream of the transcription start site (TSS) or 100 kb downstream of the transcription end site (TES). We overlapped these SNPs in GEUVADIS with the SNPs obtained from each of the three GWASs to obtain common sets of SNPs. The median number of the overlapped cis-SNPs between GEUVADIS and WTCCC, GERA or UK Biobank are 200, 556 or 500, respectively. Afterwards, for each pair of gene (from GEUVADIS) and trait (from GWAS) in turn, we examined the causal relationship between gene expression and trait of interest while testing and controlling for potential horizontal pleiotropic effects.

Compared Methods

For testing the causal effect, we compared the performance of PMR-Egger with five existing methods that include: (1) SMR that uses a single instrument but does not control for horizontal pleiotropy; here, we first performed a linear regression to choose the top associated cis-SNP to be the instrumental variable; (2) PrediXcan that uses multiple correlated instruments but does not

333 control for horizontal pleiotropy; we rely on the R package glmnet to obtain the coefficient
334 estimates of the cis-SNPs; (3) TWAS that uses multiple correlated instruments but does not
335 control for horizontal pleiotropy; we rely on the BSLMM[19] implemented in the GEMMA
336 software[56] to obtain coefficient estimates of the cis-SNPs; (4) CoMM that uses multiple
337 correlated instruments but does not control for horizontal pleiotropy; CoMM is implemented in
338 the R package CoMM; (5) LDA MR-Egger that uses multiple correlated instruments and
339 controls for horizontal pleiotropy; we contacted the authors to obtain the method source code.
340 All these methods are suitable for two-sample design and yield p values for testing the causal
341 effect α. Note that PrediXcan, TWAS and CoMM are not originally described as an MR method
342 but conceptually rely on the same joint MR model based on equations (1) and (4). These three
343 methods differ in their prior assumptions on $\boldsymbol{\beta}$: PrediXcan relies on Elastic Net assumption;
344 TWAS relies on BSLMM[19] assumption; while CoMM relies on the normal prior assumption. In
345 addition, PrediXcan and TWAS rely on a two-stage regression procedure while CoMM is based
346 on maximum likelihood. In addition, SMR is only able to use a single instrument while all other
347 methods are capable of using multiple correlated instruments. We did not compare with GSMR,
348 which requires multiple near independent instruments that are generally not feasible to obtain in
349 TWAS applications. Finally, we note that we are unable to compare our method directly with the
350 standard Egger regression, because the later requires multiple independent SNP instruments that
351 are not readily available for TWAS.

352 For testing horizontal pleiotropic effect, we compared the performance of PMR-Egger with
353 two existing methods that include (1) LDA MR-Egger; and (2) the global test in MR-PRESSO,
354 which is implemented as an R package. Both these methods are capable of producing p values to
355 test horizontal pleiotropic effect for one gene at a time.

Software Availability

PMR-Egger is implemented in the R package PMR, freely available on GitHub (https://www.xzlab.org./software.html). The code to reproduce all the analyses presented in the paper are available on GitHub (https://github.com/yuanzhongshang/PMRreproduce).

Results

Our method is described in the Methods, with technical details provided in the Supplementary Note. For TWAS applications, our method examines one gene at a time and estimates and tests its causal effect on a trait of interest. Our method models multiple correlated instruments, performs MR inference in a maximum likelihood inference framework, and is capable of testing and controlling for horizontal pleiotropic effects commonly encountered in TWAS. We refer to our method as the probabilistic Mendelian randomization with Egger regression (PMR-Egger), which is implemented in the R package PMR. Our method is computationally efficient and can analyze one gene with a few hundred thousand individuals in minutes (Table 2).

Simulations: Testing and estimating causal effect

We performed simulations to examine the effectiveness of our method and compared it with existing MR approaches. Simulation details are provided in the Methods. Briefly, we simulated gene expression values based on genotypes from 456 individuals in GEUVADIS and simulated phenotypes based on genotypes from 2,000 randomly selected individuals in GERA. In the simulations, we varied the genetic architecture underlying gene expression from sparse (one SNP or 1% of SNPs are causal) to polygenic (10% or 100% of SNPs are causal). We varied the proportion of SNPs exhibiting horizontal pleiotropic effects in a wide range (from 0%, 10%, 30%, 50% to 100%). We examined directional pleiotropy setting (the ratio of SNPs with negative vs positive horizontal pleiotropic effects is 0:10), approximately directional pleiotropy setting (1:9 or 3:7) and balanced pleiotropy settings (5:5). We varied the magnitude of horizontal pleiotropic effects γ to be either $1\text{x}10^{-4}$, $5\text{x}10^{-4}$, $1\text{x}10^{-3}$, or $2\text{x}10^{-3}$, which corresponds to the 50%, 70%, 90%, 95% percentiles of the horizontal pleiotropic effect estimate in real data. We also varied the magnitude of causal effect α to be either 0, 0.14, 0.2 or 0.245, which corresponds to a

385 proportion of phenotypic variance explained by vertical pleiotropic effects (PVE_{zy}) as 0, 0.2%,
386 0.4% and 0.6% respectively.

387 Our first set of simulations is focused on causal effect testing. Here, we compared PMR-Egger
388 with five different methods that include SMR, PrediXcan, TWAS, CoMM, and LDA MR-Egger.
389 We first examined type I error control of different methods under the null ($\alpha = 0$). In the
390 absence of horizontal pleiotropic effects, PMR-Egger, together with PrediXcan, TWAS, and
391 CoMM, all provides calibrated type I error (Fig. 1a). SMR produces overly-conservative/deflated
392 p-values, presumably because SMR requires the selected instrument being a true causal SNP
393 with large effect size, which is not always guaranteed in practice. LDA MR-Egger produces
394 inflated p-values, presumably because LDA MR-Egger makes a fixed effect assumption on $\boldsymbol{\beta}$,
395 which often does not work well in the TWAS setting where the number of SNPs are in the same
396 order as the number of samples in the gene expression study. Subsequently, the p-values from
397 LDA MR-Egger are sensitive to the linkage disequilibrium (LD) structure among SNPs and are
398 not well behaved in realistic LD settings (Supplementary Fig. 2).

399 In the presence of horizontal pleiotropic effects, PMR-Egger becomes the only method that
400 produces calibrated (or slightly conservative) p-values (Fig. 1b, c, d). In contrast, the p-values
401 from all other methods become inflated, and more so with increasingly large horizontal
402 pleiotropic effect. For example, when γ is $5\mathrm{x}10^{-4}$, the genomic control factors from PMR-Egger,
403 SMR, PrediXcan, TWAS, CoMM, and LDA MR-Egger are 0.93, 1.30, 1.33, 1.33, 1.49 and 2.61
404 respectively. When γ is increased to $1\mathrm{x}10^{-3}$, the genomic control factors from PMR-Egger, SMR,
405 PrediXcan, TWAS, CoMM, and LDA MR-Egger become 0.93, 2.39, 2.27, 2.46, 4.03 and 2.57
406 respectively. The null p-value distributions from different methods remain largely similar
407 regardless of the genetic architecture underlying gene expression being sparse or polygenic

408 (Supplementary Fig. 3). Note that, here, the p-values from SMR become less deflated when there
409 is a sparse set of SNPs affecting gene expression; however, such deflation is not completely
410 abolished even when one SNP has non-zero effect on gene expression, presumably because we
411 cannot always identify the true non-zero effect SNP through eQTL mapping and may supply a
412 tagged SNP for SMR analysis. The p-value distribution pattern for different methods under the
413 null does not change much with reduced the gene expression heritability value PVE_{zx}. When
414 PVE_{zx} is either 5% or 1%, PMR-Egger still produces well-calibrated p values (Supplementary
415 Fig. 4). Finally, like standard MR-Egger regression, our PMR-Egger makes a relatively strong
416 assumption on the horizontal pleiotropic effect and assumes that all SNPs have the same
417 horizontal pleiotropic effect. To examine the robustness of such assumption, besides the above
418 settings where either 0% or 100% SNPs have horizontal pleiotropic effects, we varied the
419 proportion of horizontal pleiotropic SNPs to be either 10%, 30%, 50%. We found that PMR-
420 Egger remain calibrated regardless of the sparsity of horizontal pleiotropic SNPs (Supplementary
421 Fig. 5). In addition, besides the above directional pleiotropy setting where the ratio of SNPs with
422 negative vs positive effects is set to be 0:10, we also examined two approximately directional
423 pleiotropy settings (1:9 or 3:7) and one balanced setting (5:5). We found that p-values of PMR-
424 Egger remains calibrated in either the approximately directional pleiotropy settings or the
425 balanced setting when horizontal pleiotropic effect is small or moderate ($\gamma = 1\text{x}10^{-4}$, $5\text{x}10^{-4}$, or
426 $1\text{x}10^{-3}$; Supplementary Fig. 6a, b, c). However, when horizontal pleiotropic effect is large
427 ($\gamma = 2\text{x}10^{-3}$), as one would expect, the p-values from PMR-Egger becomes inflated, with
428 genomic control factor being 1.08, 1.31 and 1.37, for settings where the ratio is 1:9, 3:7 and 5:5,
429 respectively (Supplementary Fig. 6d).

430 Next, we examined the power of different methods to identify the causal effect for a range of
431 possible causal effect sizes α. Because the same p-value from different methods may correspond
432 to different type I errors, we computed power based on FDR of 0.1 instead of a nominal p-value
433 threshold to allow for fair comparison across methods. In the absence of horizontal pleiotropic
434 effects or in the presence of small horizontal pleiotropic effects, PMR-Egger, TWAS and CoMM
435 have similarly power, all outperforming the other three methods (Fig. 2a, b). The power of PMR-
436 Egger is slightly lower than the other two, presumably because PMR-Egger uses extra
437 parameters to model horizontal pleiotropy and thus lose degree of freedom and subsequently
438 power in the absence of horizontal pleiotropy. The power of all methods increases with α,
439 though their relative performance rank does not change. In the presence of horizontal pleiotropy,
440 the power of all methods reduces (Fig. 2c, d). However, the power reduction from PMR-Egger is
441 substantially smaller than all other methods. For example, when $PVE_{\mathrm{zy}} = 0.006$ and $\gamma = 0.0005$,
442 PMR-Egger reaches a power of 41%; the power of SMR, PrediXcan, TWAS, CoMM, and LDA
443 MR-Egger are 7%, 24%, 31%, 33% and 1%, respectively. When $PVE_{\mathrm{zy}} = 0.006$ but $\gamma = 0.001$,
444 the power of PMR-Egger remains similar and is 40%; the power of SMR, PrediXcan, TWAS,
445 CoMM, and LDAMR-Egger reduces to 3%, 13%, 16%, 16% and 0.9%, respectively. Besides
446 horizontal pleiotropic effects $\boldsymbol{\gamma}$, we examined how power is influenced by the genetic
447 architecture underlying gene expression, $\boldsymbol{\beta}$ (Supplementary Fig. 7). We found that the power of
448 different methods in the setting where 10% of SNPs have non-zero effects on gene expression
449 are similar to the baseline setting where all SNPs have non-zero effects, both in the absence
450 (Supplementary Fig. 7e vs Fig. 2a) or in the presence of horizontal pleiotropic effects
451 (Supplementary Fig. 7F vs Fig. 2d). However, the relative performance of different methods
452 changes when there is only one SNP or 1% SNPs having non-zero effect on gene expression.

Specifically, in the absence of horizontal pleiotropic effects, the power of both PrediXcan and SMR become slightly higher than PMR-Egger, TWAS and CoMM, all of which have substantially higher power than LDA MR-Egger (Supplementary Fig. 7a, c). The higher power of PrediXcan and SMR in the sparse setting presumably is because the ElasticNet assumption employed in PrediXcan favors sparse eQTLs while SMR explicitly makes a single eQTL assumption. In the presence of horizontal pleiotropic effects, however, PMR-Egger remains the most powerful, even in the setting where only one SNP has non-zero effect on gene expression (Supplementary Fig. 7b, d). Finally, we found that PMR-Egger produces accurate estimate of the causal effect α, both under the null and under various alternatives, in the presence or absence of horizontal pleiotropic effects (Supplementary Fig. 8).

Simulations: Testing and estimating horizontal pleiotropic effect

Our second set of simulations is focused on horizontal pleiotropic effect testing. Here, we compared PMR-Egger with two different methods: LDA MR-Egger and MR-PRESSO. All three methods examine one gene at a time and test whether cis-SNPs within the gene exhibit non-zero horizontal pleiotropic effects. Note that, unlike PMR-Egger and LDA MR-Egger, MR-PRESSO requires independent instruments and uses permutation to obtain the empirical p-values. Due to the heavy computational burden resulting from permutations, we are unable to apply MR-PRESSO to all simulation scenarios and also have to restrict the number of permutations to 10,000 for each explored simulation scenario. Subsequently, the lowest possible p value from MR-PRESSO is 10^{-4}. We first examined type I error control of different methods under the null, where there is no horizontal pleiotropic effect. We found that the p-values from PMR-Egger provide calibrated type I error control under a range of causal effect sizes α (Fig. 3). However, p-values from both LDA MR-Egger and MR-PRESSO are inflated, and more so with increasingly

476 large causal effect α. For example, when $\mathrm{PVE}_{\mathrm{zy}} = 0$, the genomic control factor from PMR-
477 Egger and LDA MR-Egger are 0.96 and 2.31, respectively. When $\mathrm{PVE}_{\mathrm{zy}}$ is increased to 0.6%,
478 the genomic control factor from PMR-Egger remains 0.96, while the genomic control factors
479 from LDA MR-Egger become 3.04. The overly inflated p-values from LDA MR-Egger is
480 presumably due to its fixed effect modeling assumption on $\boldsymbol{\beta}$ and the subsequent failure to
481 control for realistic LD pattern. The inflation of MR-PRESSO p values is presumably because
482 MR-PRESSO can only handle independent instruments and thus does not fare well in TWAS
483 settings. Inflation of p-value on testing horizontal pleiotropy would incorrectly identify genes
484 with no pleiotropic effects, thus likely reducing the power to detect true causal effect α.
485 Importantly, the p-values from PMR-Egger remain calibrated regardless of the genetic
486 architecture underlying gene expression (Supplementary Fig. 9).

487 Next, we examined the power of different methods in detecting non-zero horizontal
488 pleiotropic effect. Again, we computed power based on a FDR of 0.1 instead of the nominal p-
489 value to allow for fair comparison across methods. We dropped MR-PRESSO for comparison
490 here due to its heavy computational burden. We found that PMR-Egger outperforms LDA MR-
491 Egger in a range of possible horizontal pleiotropic effect sizes, and that the power of both
492 methods increases with increasing horizontal pleiotropy (Supplementary Fig. 2e, f). For example,
493 when $\mathrm{PVE}_{\mathrm{zy}} = 0.6\%$ and $\gamma = 0.0005$, PMR-Egger achieves a power of 1.6% while LDA MR-
494 Egger achieves a power of 1% (note that both power are relatively small due to the small sample
495 size used in the simulations). When $\mathrm{PVE}_{\mathrm{zy}} = 0.6\%$ but $\gamma = 0.001$, the power of PMR-Egger
496 increases to 58.9% while the power of LDA MR-Egger increases to 32%. In addition, the power
497 to detect horizontal pleiotropic effects is not influenced by the sparsity level of the genetic
498 architecture underlying gene expression (Supplementary Fig. 10). The power to detect horizontal

pleiotropic effects does, however, depend on the sparsity level of $\boldsymbol{\gamma}$ (Supplementary Fig. 11a). Specifically, power of both PMR-Egger and LDA MR-Egger reduces with increasing sparsity of $\boldsymbol{\gamma}$, though the power of PMR-Egger remains higher than LDA MR-Egger across a range of sparsity values. Similar trend can be found under various directional horizontal pleiotropic effect assumptions (Supplementary Fig. 11b). Finally, we note that, in the presence of directional pleiotropic effect, PMR-Egger can also estimate the horizontal pleiotropic effect size accurately (Supplementary Fig. 12).

Real data applications

We performed TWAS to detect genes causally associated with any of the 39 phenotypes collected from three GWASs (details in Methods). The examined gene expression data is obtained from the GEUVADIS study and contains 15,810 genes. The examined phenotypes include 7 common diseases from WTCCC, 22 diseases from GERA, and 10 quantitative traits from UK Biobank. The GWAS sample size ranges from 4,686 (for Crohn's disease in WTCCC) to 337,199 (for UK Biobank). We applied PMR-Egger together with five other approaches (SMR, PrediXcan, TWAS, CoMM, and LDA MR-Egger) to examine pairs of gene and phenotype one at a time. The p-values for testing the causal effect of each gene on the phenotype are shown for WTCCC traits (Fig. 4a, b and Supplementary Fig. 13), GERA traits (Fig. 5a, b and Supplementary Fig. 14), and UK Biobank traits (Fig. 6a, b and Supplementary Fig. 15); with genomic control factors listed in Supplementary Table 1 and visualized in Fig. 4c, Fig. 5c and Fig. 6c. We also display qq-plots for two selected traits in each data, one with a relatively low number of gene associations and the other with a relatively high number gene associations, in Fig. 4a, b, Fig. 5a, b and Fig. 6a, b, respectively. Among the selected six traits, the one with zero number of associated genes (BD in WTCCC; Fig. 4a) and the one with one associated gene

522 (Irritable Bowel Syndrome in GERA; Fig. 5a), represent approximately null traits with no
523 apparently associated genes. For the six selected traits, consistent with the simulations, we found
524 that the p-values from PMR-Egger are well calibrated, at least more so than the other methods. In
525 contrast, the p-values from CoMM, TWAS, PrediXcan and LDA MR-Egger are inflated and
526 deviated upward from the diagonal line, while the p-values from SMR are overly conservative
527 and lie below the diagonal line. The results observed in these exemplary traits generalize to all
528 other examined traits. For example, the genomic control factor from PMR-Egger is the lowest
529 among all methods in 25 out of the 39 traits, and ranges from 0.92 to 1.04 in WTCCC (Fig. 4c),
530 from 0.92 to 1.18 in GERA (Fig. 5c), and from 1.12 to 1.34 in UK Biobank (Fig. 6c). Note that
531 the high genomic control factor in the large UK Biobank is expected under polygenic
532 architecture[57]. In contrast, the genomic control factors from CoMM, TWAS, PrediXcan are often
533 higher than that from PMR-Egger for most traits examined. For example, the genomic control
534 factor from CoMM is often the highest among all other methods (except for LDA MR-Egger) in
535 22 out of the 39 traits, and ranges from 1.12 to 1.26 in WTCCC, 0.96 to 1.92 in GERA, and 1.45
536 to 1.90 in UK Biobank. The genomic control factor from TWAS is the highest among all other
537 methods (except for LDA MR-Egger) in 14 out of the 39 traits, ranges from 1.17 to 1.33 in
538 WTCCC, 1.01 to 1.45 in GERA, and 1.30 to 2.17 in UK Biobank. The genomic control factor
539 from PrediXcan is the highest among all other methods (except for LDA MR-Egger) in 5 out of
540 the 39 traits, and ranges from 1.21 to 1.32 in WTCCC, 1.02 to 1.38 in GERA, and 1.09 to 1.46 in
541 UK Biobank. In addition, consistent with simulations, we observed a substantial inflation of
542 LDA MR-Egger p-values: its genomic control factor ranges from 7.10 to 10.48 in WTCCC,
543 32.15 to 35.24 in GERA, and 10.48 to 16.65 in UK Biobank. Also consistent with simulations,

544 the p-values from SMR often lies underneath the expected null, even though its genomic control
545 factors are often well behaved (Fig. 4a, b, Fig. 5a, b, Fig. 6a, b and Supplementary Figs. 13-15).
546 We examined the number of associated genes detected by different methods based on a
547 Bonferroni corrected genome-wide threshold (Fig. 4d, Fig. 5d and Fig. 6d; Supplementary Table
548 2). For this analysis, we excluded LDA MR-Egger for comparison, as its p-values are overly
549 inflated. Comparing across the remaining methods, we found that SMR can barely detect any
550 genes significantly associated with traits across all three data sets, much less than that detected
551 by the other four methods. The much lower number of genes detected by SMR than the other
552 four methods are consistent with the relatively low power of SMR observed in simulations. For
553 the other four methods, we found that the number of gene-trait pairs detected by CoMM and
554 PMR-Egger is higher than that detected by TWAS and PrediXcan in all three GWASs (Fig. 4d,
555 Fig. 5d and Fig. 6d; Supplementary Table 2). The higher number of discoveries by both CoMM
556 and PMR-Egger in the three GWASs is consistent with our simulations as well as previous
557 observations that likelihood-based inference often achieves higher power than two-stage
558 inference for MR analysis. However, we do notice that PMR-Egger detects slightly lower
559 number of gene-trait pairs than CoMM based on the same genome-wide p-value threshold,
560 consistent with the inflated genomic inflation factors observed for CoMM. Indeed, we found that
561 the estimated $|\frac{\alpha}{\gamma}|$ for the common set of genes detected by both CoMM and PMR-Egger is
562 higher than the set of genes only detected by CoMM across traits (Supplementary Fig. 16a, b).
563 Therefore, the genes detected by CoMM but not PMR-Egger tend to have large $|\gamma|$ and small $|\alpha|$,
564 likely reflecting false associations due to horizontal pleiotropic confounding.
565 Overall, by controlling for horizontal pleiotropic effects, PMR-Egger detected many likely
566 causal genes that the other methods failed to detect. As the first example, the *C2* gene

567 (31,765,562-32,013,449 on chr 6) is only identified by PMR-Egger to be associated with T1D
568 ($p = 3.65 \times 10^{-176}$), but not by the other methods (CoMM p=0.97; TWAS $p = 8.74 \times 10^{-4}$;
569 PrediXcan $p = 6.95 \times 10^{-3}$; SMR p=0.082). *C2* gene belongs to the classical pathway of the
570 complement immune system, is located within the class III MHC region, and has been confirmed
571 by several studies to be associated with autoimmune diseases[58,59]. As the second example, the
572 *HLA-F* gene (29,590,552-29,806,305 on chr 6), which is a HLA class I heavy chain paralogue, is
573 only identified by PMR-Egger to be associated with RA ($p = 2.57 \times 10^{-6}$), but not by the other
574 methods (CoMM $p = 4.34 \times 10^{-5}$; TWAS p=0.03; PrediXcan p=0.02; SMR p=0.08). As the
575 third example, the *HERPUD1* gene (56,865,960-57,077,798 on chr 16), which is a well-known
576 lipid-related gene[60,61], is only detected by PMR-Egger to be associated with dyslipidemia
577 ($p = 2.48 \times 10^{-6}$), but not by the other methods (CoMM $p = 2.32 \times 10^{-5}$; TWAS $p = 6.92 \times$
578 10^{-3}; PrediXcan $p = 5.96 \times 10^{-3}$; SMR p=0.10). As the fourth example, the *ADAM15* gene
579 (154,923,042-155,135,252 on chr 1), which has been confirmed to be associated with height in
580 the UK Biobank Pheweb, is only detected by PMR-Egger to be associated with height ($p =$
581 3.60×10^{-8}), but not by the other methods (CoMM $p = 1.52 \times 10^{-5}$; TWAS p=0.17;
582 PrediXcan p=0.44; SMR p=0.14). Finally, the *DNAJC27-AS1* gene (25,094,259-25,362,563 on
583 chr 2) is only identified by PMR-Egger to be associated with height ($p = 3.16 \times 10^{-24}$), but not
584 by the other methods (CoMM p=0.99; TWAS $p = 1.10 \times 10^{-4}$; PrediXcan p=0.02; SMR
585 p=0.05). *DNAJC27-AS1* is a *DNAJC27* antisense RNA gene and can inhibit the expression of
586 *DNAJC27*, which has been confirmed to be associated with height in the UK Biobank Pheweb
587 and other large association studies[62].

588 Next, we shift our focus to testing horizontal pleiotropic effects. The p-values for testing the
589 causal effect of gene on phenotype are shown for WTCCC traits (Fig. 4e, f and Supplementary

590 Fig. 13), GERA traits (Fig. 5e, f and Supplementary Fig. 17), and UK Biobank traits (Fig. 6e, f
591 and Supplementary Fig. 18); with genomic control factors visualized in Fig. 4g, Fig. 5g and Fig.
592 6g. We also display qq-plots for the previously selected exemplary traits in Fig. 4e, f, Fig. 5e, f,
593 and Fig. 6e, F. Overall, consistent with simulations, the p-values from PMR-Egger are well
594 behaved while the p-value from LDA MR-Egger display substantial inflation. For example, the
595 genomic control factor from PMR-Egger ranges from 0.94 to 1.03 in WTCCC (Fig. 4g), from
596 0.92 to 1.26 in GERA (Fig. 5g), and from 1.13 to 1.71 in UK Biobank (Fig. 6g). In contrast, the
597 genomic control factor from LDA MR-Egger ranges from 10.16 to 15.64 in WTCCC, from 70.43
598 to 72.19 in GERA and from 17.75 to 29.85 in UK Biobank (Supplementary Table 1). With the
599 same Bonferroni adjusted genome-wide p-value threshold, PMR-Egger detected 33 gene-trait
600 pairs in WTCCC in which the cis-SNPs exhibit significant horizontal pleiotropy, 37 gene-trait
601 pairs in GERA, and 626 gene-trait pairs in the UK Biobank.

602 Horizontal pleiotropic effect tests can help us explain some of the discrepancy in terms of the
603 causal associations detected by PMR-Egger and the other methods. For example, for T1D in
604 WTCCC, the *ZKSCAN4* gene on chromosome 6 shows a significant pleiotropy effect ($p =$
605 2.93×10^{-8}) but displays no significant causal effect (p=0.99) by PMR-Egger. In contrast,
606 *ZKSCAN4* is detected to be significantly associated with T1D by PrediXcan ($p = 2.80 \times 10^{-6}$),
607 and, to a lesser extent, by TWAS ($p = 4.12 \times 10^{-5}$). However, *ZKSCAN4* has not been reported
608 to be associated with T1D by previous GWASs. A careful examination of the *ZKSCAN4* locus
609 (28,112,401-28,327,011) revealed a neighboring MHC region (28,477,797-33,448,354), a region
610 previously well recognized to affect T1D and other autoimmune diseases. Indeed, by controlling
611 for the predicted expression level of the *APOM* gene (31,520,193-31,725,987) in the MHC
612 region in the PrediXcan framework, the association between the predicted *ZKSCAN4* expression

613 level and T1D is no longer significant (p = 0.38). Therefore, the causal association between
614 *ZKSCAN4* and T1D detected by PrediXcan and TWAS likely reflect the horizontal pleiotropic
615 effect of *ZKSCAN4* cis-SNPs on T1D through genes in the MHC region. As another example, for
616 the trait of red blood cell count in UK Biobank, the *MAPT* gene on chromosome 17 shows a
617 significant pleiotropy effect ($p = 2.35 \times 10^{-9}$) but displays no significant causal effect (p=0.98)
618 by PMR-Egger. In contrast, *MAPT* is detected to be significantly associated with red blood cell
619 count by PrediXcan ($p = 8.11 \times 10^{-10}$), and, to a lesser extent, by TWAS ($p = 1.72 \times 10^{-3}$).
620 However, no previous evidence suggests that *MAPT* is associated with red blood cell count.
621 Indeed, we found that the genomic location of *MAPT* (43,871,748-44,205,700) is close to and
622 partially overlapped with *KANSL1* (44,007,282-44,402,733), which has been previously
623 identified to be associated with red blood cell traits[61,63]. Such association between *KANSL1* and
624 red blood cell count is also detected by PMR-Egger ($p = 1.02 \times 10^{-7}$), by CoMM ($p = 2.72 \times$
625 10^{-8}), and, to a lesser extent, by TWAS ($p = 1.66 \times 10^{-3}$) in the present study. By controlling
626 for the predicted expression level of the *KANSL1* gene in the PrediXcan framework, the
627 association between the predicted *MAPT* expression level and red blood cell count is no longer
628 significant (p = 0.10). Therefore, the causal association between *MAPT* and red blood cell count
629 detected by PrediXcan likely reflects the horizontal pleiotropic effect of *MAPT* cis-SNPs on red
630 blood cell count through *KANSL1*. As a final example, for height in the UK Biobank, the
631 pseudogene *RP11-9E13.2* (70,137,755-70,340,521) on chromosome 10 has a significant
632 pleiotropy effect ($p = 1.08 \times 10^{-13}$) but displays no significant causal effect (p=0.93) by PMR-
633 Egger. In contrast, *RP11-9E13.2* is detected to be significantly associated with height by
634 PrediXcan ($p = 4.34 \times 10^{-10}$), and, to a lesser extent, by TWAS ($p = 9.05 \times 10^{-6}$). The
635 pseudogene *RP11-9E13.2* is in the neighborhood of *MYPN* (69,765,912-70,071,774), which has

been previously identified to be associated with height[64]. The association between *MYPN* and height is also detected by PMR-Egger ($p = 1.82 \times 10^{-7}$), CoMM ($p = 2.13 \times 10^{-14}$), and to a lesser extent, PrediXcan ($p = 3.94 \times 10^{-4}$) and TWAS ($p = 1.55 \times 10^{-3}$), in the present study. By controlling for the predicted expression level of *MYPN* gene in the PrediXcan framework, the association between the predicted *RP11-9E13.2* expression level and height is no longer significant at the genome-wide threshold ($p = 3.37 \times 10^{-4}$). Therefore, the causal association between the pseudogene *RP11-9E13.2* and height as detected by PrediXcan and TWAS likely reflects the horizontal pleiotropic effect of *RP11-9E13.2* cis-SNPs on height at least partially through *MYPN*. The results suggest the practical importance of testing and controlling for pleiotropic effects in TWAS applications.

Finally, we note that an important feature of PMR-Egger is its ability to test both causal effect and horizontal pleiotropy effect simultaneously. We contrast the p-values obtained from these two different tests across genes for those traits in which at least one gene is detected as significant from either of the two tests (Supplementary Figs. 19-21). We found that different traits exhibit different gene association patterns. For example, some traits may only contain genes with a significant causal effect but without a significant horizontal pleiotropic effect (e.g. CD and CAD in WTCCC; Allergic Rhinitis, Irritable Bowel Syndrome and Psychiatric disorders in GERA). Some traits may only contain genes with a significant horizontal pleiotropic effect but without a significant causal effect (e.g. Dermatophytosis in GERA). Some traits may contain genes with a significant causal effect as well as genes with a significant horizontal pleiotropic effect, but with the two sets of genes being non-overlapped (e.g. Asthma, Dyslipidemia, HT, Abdominal Hernia and Macular Degeneration in GERA; Fored Vitral Capacity in UK Biobank). While the majority of traits contain genes with both a significant causal effect and a significant

horizontal pleiotropic effect. The top gene which is most significant for both causal effect test and pleiotropy test is highlighted in the plots. Being capable of testing both causal effect and horizontal pleiotropy effect facilitates our understanding of the gene association pattern with various different complex traits.

Discussion

We have presented a data generative model and a likelihood framework for MR analysis that unifies many existing transcriptome wide association analysis methods and many existing MR methods. Under the framework, we have presented PMR-Egger, a new method that conducts MR analysis using multiple correlated instruments while properly controlling for horizontal pleiotropic effects. By properly controlling for horizontal pleiotropic effects and making inference under a likelihood framework, PMR-Egger yields calibrated p-values across a wide range of scenarios and improves power of MR analysis over existing approaches. We have illustrated the benefits of PMR-Egger through extensive simulations and multiple real data applications of TWAS.

One important modeling assumption we made in PMR-Egger is that the horizontal pleiotropic effects of all SNPs equal to each other. The equal effect size assumption directly follows the commonly used Egger regression modeling assumption for MR analysis and is analogous to the burden effect size assumption commonly used for rare variant tests. Consistent with existing literature on applications of the Egger regression and burden test, we also found that equal effect size assumption employed in PMR-Egger works reasonably robust with respect to a range of model mis-specifications and appears to be effective in several real data applications examined here. However, we do acknowledge that our equal effect size assumption in PMR-Egger can be overly restrictive in many settings. We have attempted to alleviate this restrictive modeling assumption by imposing an alternative modeling assumption on the horizontal effect sizes based on variance component assumption. In particular, we have attempted to assume that the horizontal pleiotropic effect of each SNP follows a normal distribution with mean zero and a certain variance component parameter, i.e. analogous to the SKAT test assumption[65]. Such

688 variance component assumption is a more flexible modeling assumption than the equal effect
689 size assumption, potentially alleviating much of the concern with respect to the sensitivity and
690 robustness of equal effect size assumption. Unfortunately, under the variance component
691 assumption, inference for the resulting PMR model becomes overly complicated. In particular,
692 due to the estimation uncertainty in the hyper-parameter estimates, the p-values from the PMR
693 variance component model becomes severely deflated even under simple null simulations
694 (Figures S22). Such deflation of p-values has been previously observed in variance component
695 tests for microbiome applications[66]. Only few methods exist to address such p-value in-
696 calibration issue resulting from hyper-parameter estimation uncertainty[67], and it is not
697 straightforward to adapt any of these methods to our PMR variance component model. Besides
698 the equal effect size modeling restriction, we also note that neither PMR-Egger nor the PMR
699 variance component model is capable of accounting for correlation between horizontal
700 pleiotropic effects γ and the SNP effects on gene expression β. Therefore, while we view PMR-
701 Egger as in important first step towards effective control of horizontal pleiotropic effects in
702 TWAS applications, we emphasize that imposing more realistic modeling assumptions on the
703 horizontal pleiotropic effects in the PMR framework will likely yield more fruitful results in the
704 future.

705 We have primarily focused on modeling continue traits with PMR-Egger. For case control
706 studies, we have followed previous approaches and directly treated binary phenotypes as
707 continuous outcomes[19,49,68,69], which appears to work well in both WTCCC and GERA data
708 applications we examined. Treating binary phenotypes as continuous outcomes can be justified
709 by recognizing the linear model as a first order Taylor approximation to a generalized linear
710 model[19]. However, it would be desirable to extend PMR-Egger to accommodate case control

711 data or other discrete data types in a principled way, by, for example, extending PMR-Egger into
712 the generalized linear model framework. In particular, we could use a probit or a logistic link to
713 extend PMR-Egger to directly model case control data. Extending PMR-Egger to model discrete
714 data types using the generalized linear model framework would likely lead to wider applications
715 of PMR-Egger and is thus an important avenue for future research.

716 We have primarily focused on modeling individual-level data with PMR-Egger. However, like
717 many other linear model-based methods in statistical genetics, PMR-Egger can also be easily
718 extended to make sure of summary statistics. The summary statistics version of PMR-Egger is
719 described in detail in the Supplementary Text. Briefly, the summary statistics version of PMR-
720 Egger requires marginal SNP effect size estimates and their standard errors, both on the gene
721 expression and on the trait of interest. In addition, it requires a SNP by SNP correlation matrix
722 that can be constructed based on a reference panel. Being able to make use of summary statistics
723 extends the applicability of PMR-Egger to data sets where individual-level genotype or
724 phenotype are not available.

725 Finally, in addition to what we have already mentioned in the Materials and Methods, we
726 emphasize here again, that, while we have followed the previous MR literature and use "causal
727 effect" through the text, the effect is causal only when certain MR modeling assumptions hold.
728 These MR assumptions are often not straightforward to prove. For example, without measuring
729 all potential confounders, it is not straightforward to argue that the SNP instruments are not
730 associated with any other confounders that may be associated with both exposure and outcome.
731 Therefore, we caution against the over-interpretation of causal inference in observation studies
732 such as TWAS applications. However, we do believe MR is an important step that allows us to

move beyond standard linear regressions and is an important analysis that can provide potentially more trustworthy evidence with regard to causality compared to simpler approaches.

References

1 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091-1098 (2015).

2 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245-252 (2016).

3 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**, 481-487 (2016).

4 Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature communications* **9**, 224 (2018).

5 Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *Statistical methods in medical research* **26**, 2333-2355 (2017).

6 Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and risk of cardiovascular disease and diabetes. *New England Journal of Medicine* **375**, 2144-2153 (2016).

7 Helgadottir, A. *et al.* Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nature genetics* **48**, 634-639 (2016).

8 Pingault, J.-B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19**, 566-580 (2018).

9 Zheng, J. *et al.* Recent developments in Mendelian randomization studies. *Current epidemiology reports* **4**, 330-345 (2017).

10 Haycock, P. C. *et al.* Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American journal of clinical nutrition* **103**, 965-978 (2016).

11 Lawlor, D. A. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International journal of epidemiology* **45**, 908-915 (2016).

12 Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine* **36**, 1783-1802 (2017).

13 Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology* **44**, 512-525 (2015).

14 Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology* **40**, 304-314 (2016).

15 Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic epidemiology* **37**, 658-665 (2013).

16 Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine* **35**, 1880-1906 (2016).

17 Burgess, S. & Thompson, S. G. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in medicine* **30**, 1312-1323 (2011).

18 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320 (2005).

19 Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *Plos Genetics* **9**, : e1003264. (2013).

20 Yang, C. *et al.* CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, doi:10.1093/bioinformatics/bty865 (2018).

21 Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics* **50**, 693-698 (2018).

22 Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics* **27**, R195-R208 (2018).

23 Jordan, D. M., Verbanck, M. & Do, R. The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *bioRxiv*, 311332 (2018).

24 Bowden, J. *et al.* Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. *International journal of epidemiology*, 1-15 (2018).

25 Bowden, J., Hemani, G. & Davey Smith, G. Detecting individual and global horizontal pleiotropy in Mendelian randomization: a job for the humble heterogeneity statistic? *American journal of epidemiology* **187**, 2681-2685 (2018).

26 Park, Y. *et al.* A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. *bioRxiv*, 219428 (2017).

27 Kang, H., Zhang, A., Cai, T. T. & Small, D. S. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association* **111**, 132-144 (2016).

28 Guo, Z., Kang, H., Tony Cai, T. & Small, D. S. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 793-815 (2018).

29 Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *European journal of epidemiology* **32**, 391-392 (2017).

30 Dai, J. Y. *et al.* Diagnostics of Pleiotropy in Mendelian Randomization Studies: Global and Individual Tests for Direct Effects. *American journal of epidemiology* **187**, 2672-2680 (2018).

31 Zhao, Q., Wang, J., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *arXiv:1801.09652* (2018).

809 32 Qi, G. & Chatterjee, N. Mendelian Randomization Analysis Using Mixture Models (MRMix) for
810 Genetic Effect-Size-Distribution Leads to Robust Estimation of Causal Effects. *bioRxiv*, 367821
811 (2018).
812 33 Berzuini, C., Guo, H., Burgess, S. & Bernardinelli, L. A Bayesian approach to Mendelian
813 randomization with multiple pleiotropic variants. *Biostatistics*, 1-16 (2018).
814 34 Li, S. Mendelian randomization when many instruments are invalid: hierarchical empirical Bayes
815 estimation. *arXiv:1706.01389* (2017).
816 35 Barfield, R. *et al.* Transcriptome-wide association studies accounting for colocalization using
817 Egger regression. *Genetic epidemiology* **42**, 418-433 (2018).
818 36 Dawid, A. P. Causal inference without counterfactuals. *Journal of the American statistical*
819 *Association* **95**, 407-424 (2000).
820 37 Dawid, A. P. Statistical causality from a decision-theoretic perspective. *Annual Review of*
821 *Statistics and Its Application* **2**, 273-303 (2015).
822 38 Berzuini, C., Dawid, P. & Bernardinell, L. *Causality: Statistical perspectives and applications*.
823 (John Wiley & Sons, 2012).
824 39 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in
825 humans. *Nature* **501**, 506-511 (2013).
826 40 Banda, Y. *et al.* Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the
827 Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* **200**, 1285-
828 1295 (2015).
829 41 Kvale, M. N. *et al.* Genotyping informatics and quality control for 100,000 subjects in the Genetic
830 Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* **200**, 1051-1060
831 (2015).
832 42 Price, A. L. *et al.* Effects of cis and trans genetic ancestry on gene expression in African
833 Americans. *Plos Genetics* **4**, e1000294 (2008).
834 43 Price, A. L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-
835 descent in related or unrelated individuals. *Plos Genetics* **7**, e1001317 (2011).
836 44 Consortium, W. T. C. C. Genome-wide association study of 14,000 cases of seven common
837 diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
838 45 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**,
839 203-209 (2018).
840 46 Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and
841 functional annotation. *Plos Genetics* **11**, e1005176 (2015).
842 47 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project.
843 *Genome research* **22**, 1760-1774 (2012).

844 48 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression
845 residuals (PEER) to obtain increased power and interpretability of gene expression analyses.
846 *Nature protocols* **7**, 500-507 (2012).
847 49 Zeng, P. & Zhou, X. Non-parametric genetic prediction of complex traits with latent Dirichlet
848 process regression models. *Nature communications* **8**, 456 (2017).
849 50 Guan, Y. & Stephens, M. Practical issues in imputation-based association mapping. *Plos Genetics*
850 **4**, e1000279 (2008).
851 51 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for
852 the next generation of genome-wide association studies. *Plos Genetics* **5**, e1000529 (2009).
853 52 Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and
854 population genetic studies. *Nature methods* **10**, 5-6 (2012).
855 53 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature*
856 *genetics* **48**, 1279-1283 (2016).
857 54 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature genetics* **48**,
858 1284-1287 (2016).
859 55 Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for
860 biobank-scale datasets. *Nature genetics* **50**, 906-908 (2018).
861 56 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies.
862 *Nature genetics* **44**, 821-824 (2012).
863 57 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass
864 index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649,
865 doi:10.1093/hmg/ddy271 (2018).
866 58 Yu, C. Y. Molecular genetics of the human MHC complement gene cluster. *Experimental and*
867 *clinical immunogenetics* **15**, 213-230 (1998).
868 59 Zhu, Z. B., Hsieh, S.-L., Bentley, D., Campbell, R. & Volanakis, J. A variable number of tandem
869 repeats locus within the human complement C2 gene is associated with a retroposon derived
870 from a human endogenous retrovirus. *Journal of Experimental Medicine* **175**, 1783-1787 (1992).
871 60 Smith, E. N. *et al.* Longitudinal genome-wide association of cardiovascular disease risk factors in
872 the Bogalusa heart study. *Plos Genetics* **6**, e1001094 (2010).
873 61 Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types
874 to complex human diseases. *Nature genetics* **50**, 390-400 (2018).
875 62 Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect
876 human height. *Nature* **467**, 832-838 (2010).
877 63 Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common
878 complex disease. *Cell* **167**, 1415-1429 (2016).
879 64 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological
880 architecture of adult human height. *Nature genetics* **46**, 1173-1186 (2014).

65 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82-93 (2011).

66 Zhao, N. *et al.* Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *The American Journal of Human Genetics* **96**, 797-807 (2015).

67 Chen, J., Chen, W., Zhao, N., Wu, M. C. & Schaid, D. J. Small sample kernel association tests for human genetic and microbiome association studies. *Genetic epidemiology* **40**, 5-19 (2016).

68 Yang, J., Fritsche, L. G., Zhou, X., Abecasis, G. & Consortium, I. A.-R. M. D. G. A scalable Bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics* **101**, 404-416 (2017).

69 Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *Plos Genetics* **13**, e1006869 (2017).

Acknowledgements

This study was supported by the National Institutes of Health (NIH) grants R01HG009124 and R01HL142023, National Science Foundation (NSF) grant DMS1712933, and the National Natural Science Foundation of China (81872712 and 81673272). We thank the Wellcome Trust Centre for Human Genetics for making the heterogenous stock mouse data available online. This study makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from http://www. wtccc.org.uk/. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475. The GERA Data (dbGaP accession number phs000788) came from a grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Northern California Community Benefit Programs. The RPGEH and the Resource for Genetic Epidemiology Research in Adult Health and Aging are described in the following publication, Schaefer C, et al., The Kaiser Permanente Research Program on Genes, Environment and Health: Development of a Research Resource in a Multi-Ethnic Health Plan with Electronic Medical Records, in preparation, 2013. This study has been conducted using UK Biobank resource under Application Number 30686. UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It has also had funding from the Welsh Assembly Government, British Heart Foundation and Diabetes UK.

Author contributions

XZ conceived the study. XZ and ZY developed the methods. ZY developed the software tool with assistance from JL and CY. ZY performed simulations and real data analysis with assistance from HZ, PZ, SY, and SS. XZ and ZY wrote the manuscript with input from all other authors. All authors reviewed and approved the final manuscript.

Competing interests: The authors declare no competing interests.

Table 1 Summary of some existing MR methods

	Design	Instrumental variable	β effect assumption	γ effect assumption	Estimation procedure	Individual or summary data
PrediXcan[1]	Two-sample	Correlated	Elastic net	N/A	Two-stage	Both
TWAS[2]	Two-sample	Correlated	BSLMM	N/A	Two-stage	Both
SMR[3]	Two-sample	Univariate	Fixed effect	N/A	Two-stage	Summary
GSMR[4]	Two-sample	Near-independent	Fixed effect	N/A	Two-stage	Summary
MR-Egger[13]	Two-sample	Independent	Fixed effect	Equal effect size	Two-stage	Summary
CoMM[20]	Two-sample	Correlated	Normal	N/A	MLE	Individual
CaMMEL[26]	Two-sample	Correlated	Fixed effect	Normal	Variational Bayes	Summary
Kang et al.[27]	One-sample	Correlated	Fixed effect	Lasso	Two-stage	Individual
MRMix[32]	Two-sample	Independent	Normal Mixture	Normal Mixture	Estimating equation	Summary
Berzuini et al.[33]	One-sample	Correlated	Fixed effect	Horseshoe	MCMC	Individual
LDA MR-Egger[35]	Two-sample	Correlated	Fixed effect	Equal effect size	Two-stage	Summary
PMR-Egger	**Two-sample**	**Correlated**	**Normal**	**Equal effect size**	**MLE**	**Both**

Methods are categorized based on the experimental design (two-sample vs one-sample vs both), the characterizes of selected instrumental variables (univariate vs multiple independent vs multiple correlated), β effect size assumption, γ effect size assumption, estimation/inference procedure (ratio-based vs two-stage estimation vs maximum likelihood vs Bayesian), and input data type (individual-level vs summary). The categorization of inference procedure generally follows ref [5]. In the inference procedure, the two-stage estimation procedure comprises two regression stages: the first-stage regression of the exposure on the instrumental variables, and the second-stage regression of the outcome on the fitted values of the exposure from the first stage. Some inference procedures, such as the inverse variance weighted (IVW) procedure (e.g. MR-Egger[13]) or the ratio method (e.g. for SMR[3]) are categorized as two-stage procedure here, as both are asymptotically equivalent to a two-stage estimation procedure in the case of independent instruments. We only list MR methods that directly take input instruments into the model; many MR methods that performs various selection procedures on the instruments (e.g. Guo et al[28]) are not included in the table. Some recently developed methods that only test for horizontal pleiotropy, such as GLIDE[30], MR-PRESSO[21] and modified second-order weighting of Cohran's Q statistic[24,25], are not included in the table.

Table 2. Mean computational time (in second) of various MR methods

Trait	#SNP in the exemplary gene	CoMM	PMR-Egger	TWAS	LDA MR-Egger	SMR	PrediXcan	MR-PRESSO
T1D from WTCCC (n=4901)	300	0.51(0.19)	0.80(0.57)	96.09(72.39)	0.08(0.02)	0.0003(0.0005)	26.74(2.81)	408.27(74.76)
	500	1.21(0.41)	1.42(0.77)	3.48(1.16)	0.14(0.03)	0.0004(0.0005)	11.77(0.64)	829.04(135.79)
	983	5.85(1.50)	9.79(1.56)	4.69(1.73)	0.60(0.09)	0.0004(0.0005)	9.96(0.78)	2023.77(260.43)
	2106	111.00(12.87)	97.33(7.63)	5.87(2.26)	4.18(0.59)	0.0005(0.0005)	22.90(2.63)	4913.22(554.47)
Asthma from GERA (n=61,953)	300	1.47(0.29)	2.06(0.22)	2.61(1.48)	0.05(0.02)	0.0002(0.0004)	33.39(3.09)	464.64(62.18)
	500	1.21(0.33)	4.21(0.81)	2.54(0.87)	0.09(0.03)	0.0002(0.0004)	11.71(0.70)	919.66(102.83)
	1000	24.37(5.13)	21.68(1.66)	3.07(2.55)	0.46(0.13)	0.0002(0.0004)	14.29(1.30)	2275.42(263.95)
	2008	59.01(4.98)	52.52(4.47)	4.51(1.48)	2.33(0.71)	0.0004(0.0005)	20.18(3.28)	5213.73(601.46)
Platelet Count from UK Biobank (n=337,199)	300	2.56(0.53)	5.57(4.54)	5.04(4.19)	0.09(0.02)	0.0008(0.0004)	10.93(1.96)	471.55(50.44)
	500	6.82(2.75)	7.61(2.30)	5.44(4.30)	0.15(0.02)	0.0007(0.0005)	12.17(1.04)	876.06(92.90)
	1052	24.92(6.28)	23.59(3.21)	5.91(4.79)	0.81(0.09)	0.0008(0.0004)	16.05(2.38)	2133.03(77.56)
	2605	186.14(28.45)	178.68(16.75)	5.37(0.73)	8.11(1.20)	0.0008(0.0004)	9.89(1.74)	6949.72(245.75)

Computation is carried out on a single thread of a Xeon Gold 6138 CPU. The computation time is averaged across 20 replicates, with values inside parentheses denoting the standard deviation. #SNP denotes the number of cis-SNPs for four exemplary genes in each study. The computational time for MR-PRESSO is based on 10,000 permutations.

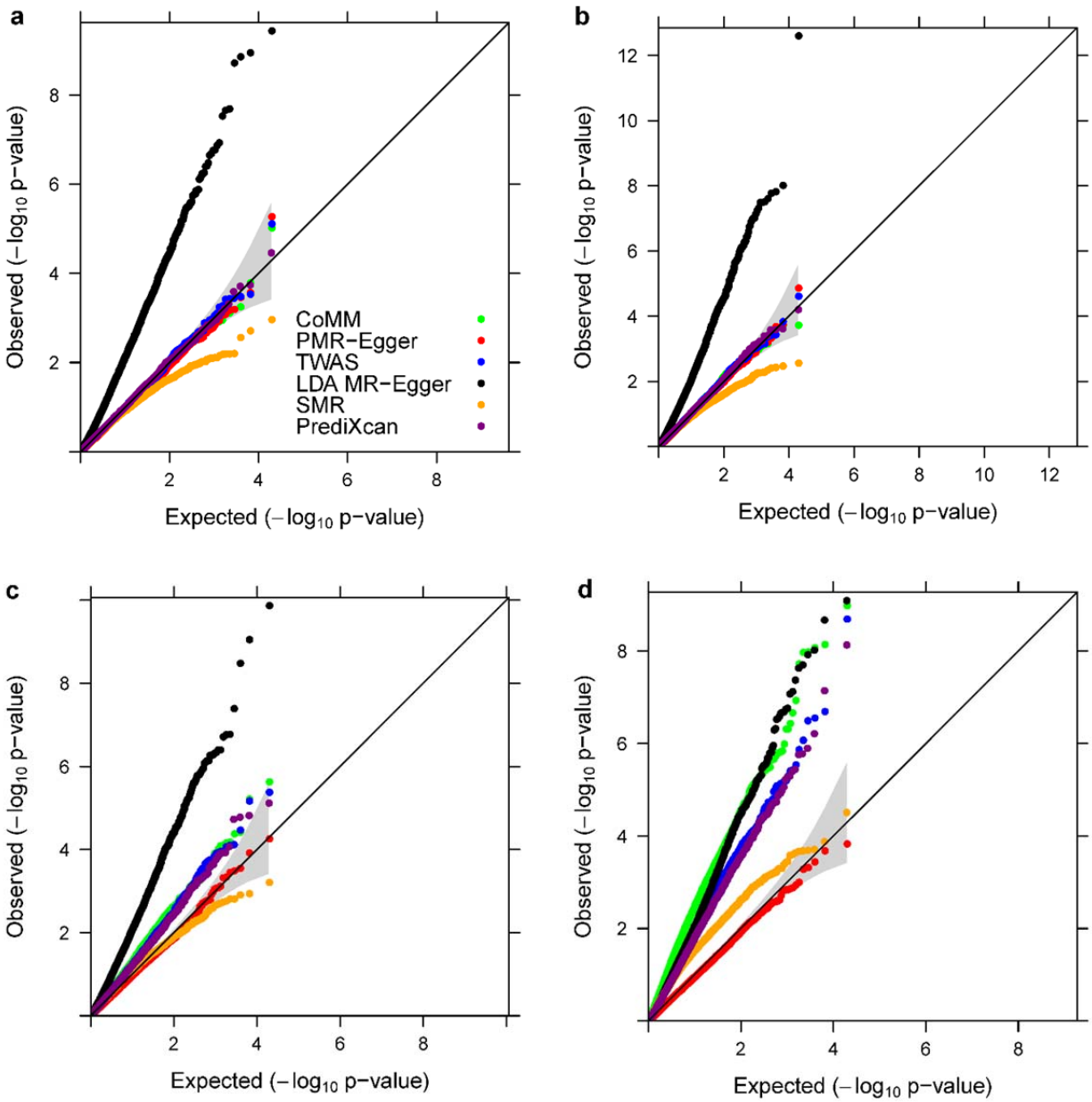

Fig. 1 Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect either in the absence or in the presence of horizontal pleiotropic effect under null simulations. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Null simulations are performed under different horizontal pleiotropic effect sizes: (**a**) $\gamma = 0$; (**b**) $\gamma = 0.0001$; (**c**) $\gamma = 0.0005$; (**d**) $\gamma = 0.001$. Only p-values from PMR-Egger adhere to the expected diagonal line across a range of horizontal pleiotropic effect sizes.

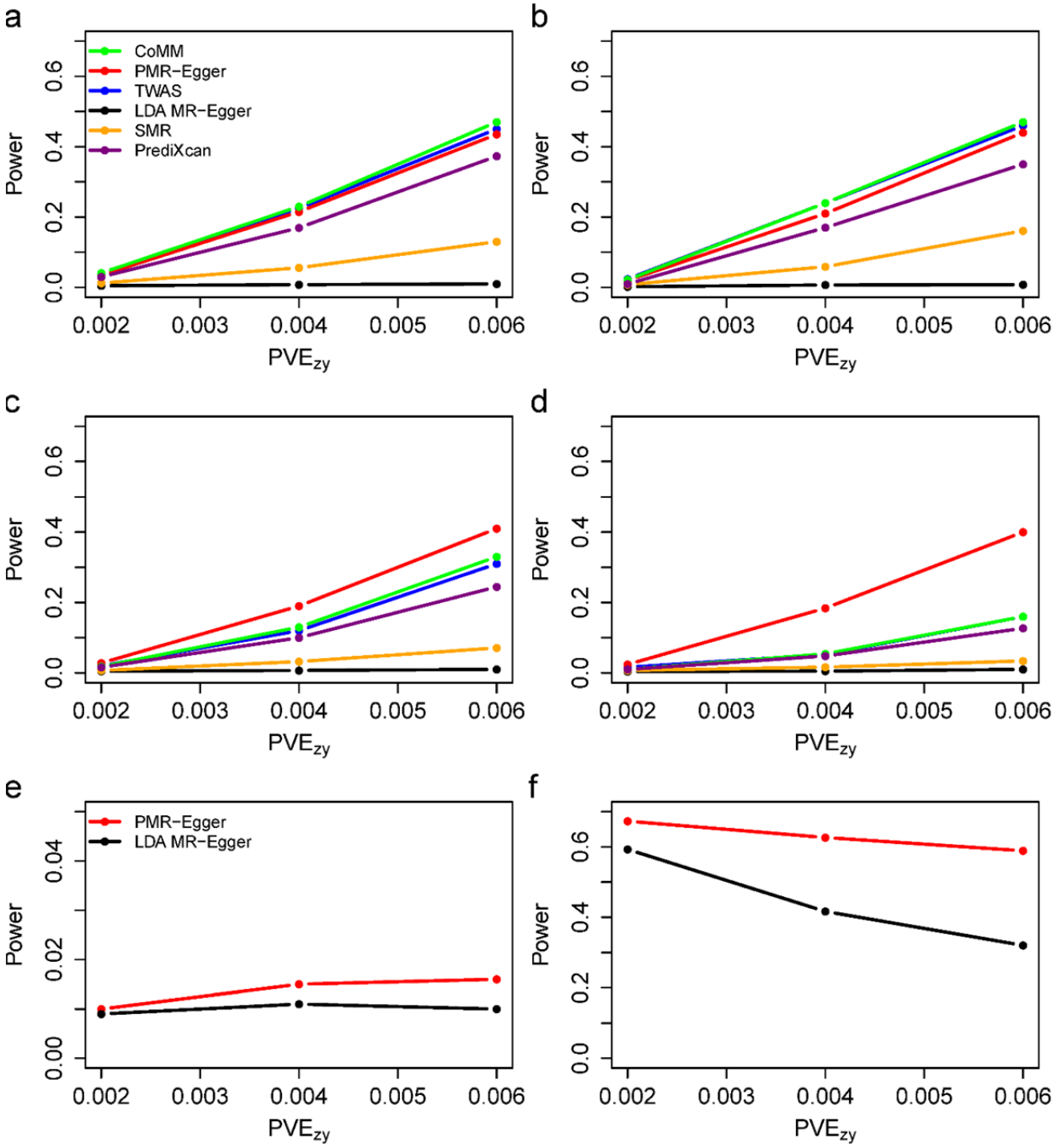

Fig. 2 Power of different methods under various simulation scenarios. Power (y-axis) at a false discovery rate of 0.1 to detect the causal effect (**a**-**d**) or the horizontal pleiotropic effect (**e**-**f**) is plotted against different causal effect size characterized by PVE_{zy} (x-axis). Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Simulations are performed under different horizontal pleiotropic effect sizes: (**a**) $\gamma = 0$; (**b**) $\gamma = 0.0001$; (**c**, **e**) $\gamma = 0.0005$; (**d**, **f**) $\gamma = 0.001$.

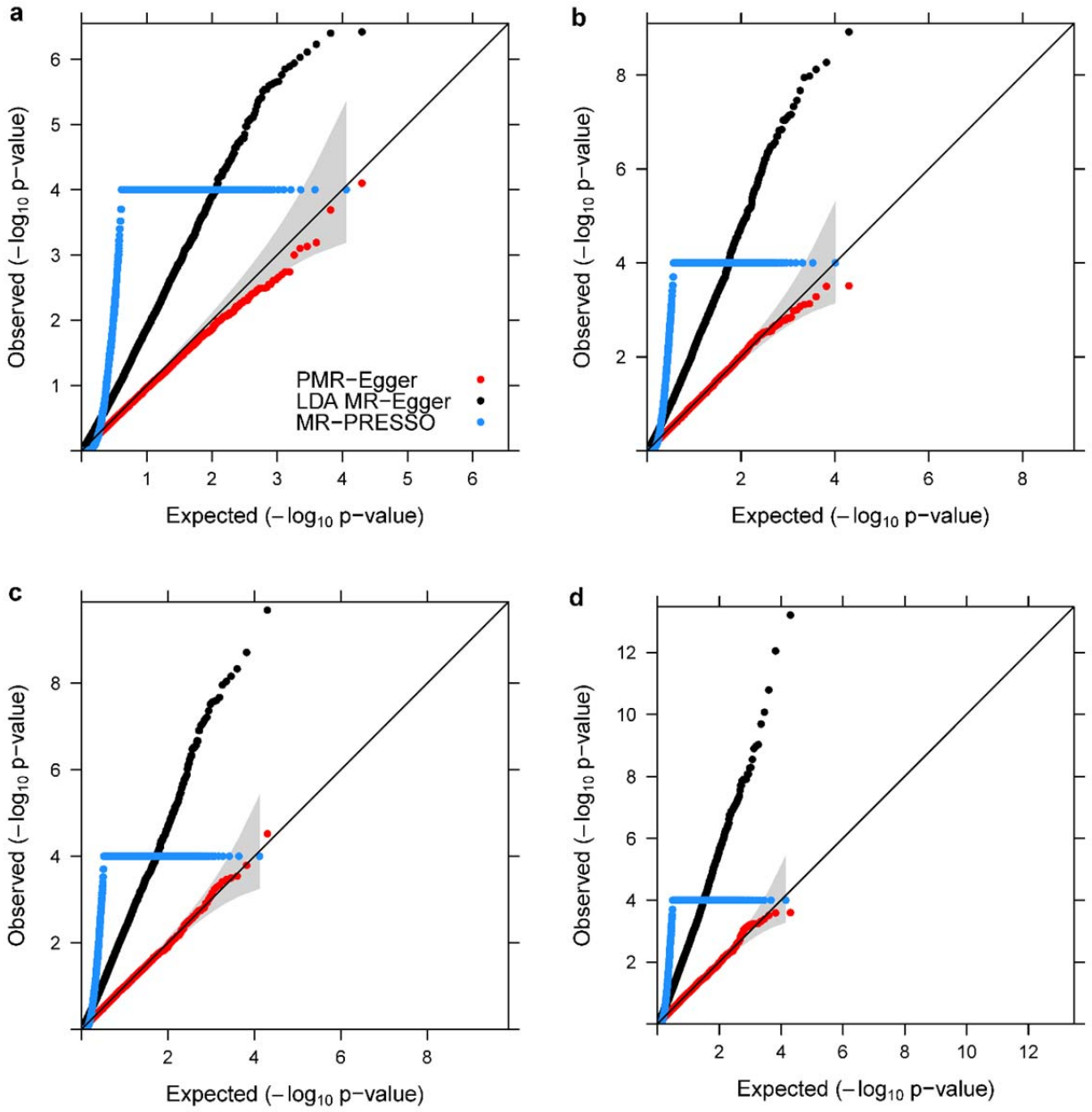

Fig. 3 Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect either in the absence or in the presence of causal effect under null simulations. Compared methods include PMR-Egger (red), LDA MR-Egger (black), and MR-PRESSO (dodger blue). Null simulations are performed under different causal effect sizes characterized by $\mathrm{PVE}_{\mathrm{zy}}$: (**a**) $\mathrm{PVE}_{\mathrm{zy}} = 0$; (**b**) $\mathrm{PVE}_{\mathrm{zy}} = 0.2\%$; (**c**) $\mathrm{PVE}_{\mathrm{zy}} = 0.4\%$; and (**d**) $\mathrm{PVE}_{\mathrm{zy}} = 0.6\%$. Only p-values from PMR-Egger adhere to the expected diagonal line across a range of horizontal pleiotropic effect sizes. Due to heavy computational burden, we are only able to run 10,000 permutations for MR-PRESSO. Therefore, the minimal p-value from MR-PRESSO is 10^{-4}.

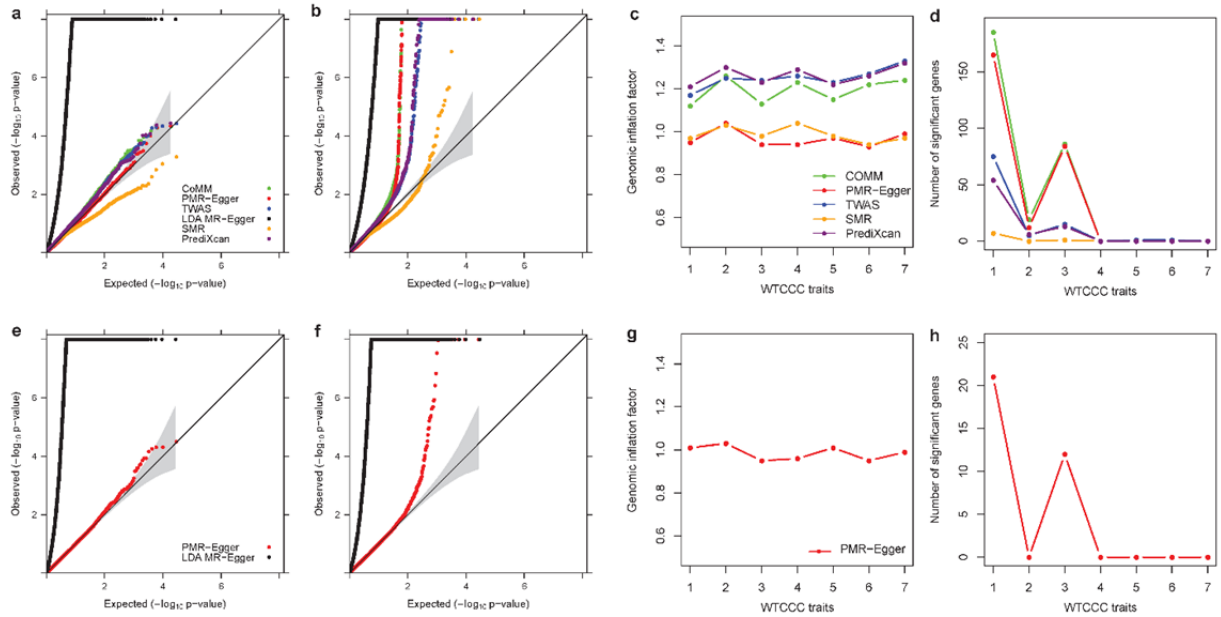

Fig. 4 TWAS analysis results by different methods for traits in the WTCCC data. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). (**a**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for an exemplary trait BD. (**b**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for another exemplary trait T1D. (**c**) Genomic inflation factor for testing the causal effect for each of the 7 traits by different methods. (**d**) Number of causal genes identified for each of the 7 traits by different methods. (**e**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for an exemplary trait BD. (**f**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for another exemplary trait T1D. (**g**) Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 7 traits by different methods. (**h**) Number of genes identified to have significant horizontal pleiotropic effect for each of the 7 traits by different methods. For **c**, **d**, **g**, **h**, the number on the x-axis represents seven traits in order: T1D, CD, RA, BD, T2D, CAD, HT.

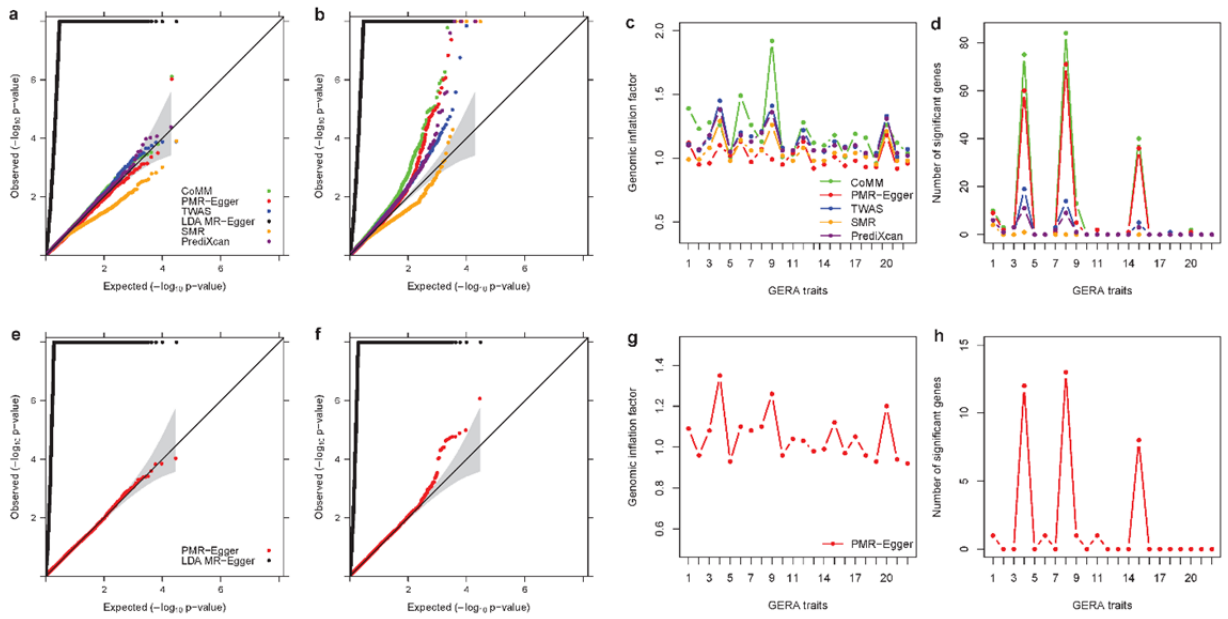

Fig. 5 TWAS analysis results by different methods for traits in the GERA data. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). (**a**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for an exemplary trait Irritable Bowel Syndrome. (**b**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for another exemplary trait Asthma. (**c**) Genomic inflation factor for testing the causal effect for each of the 22 traits by different methods. (**d**) Number of causal genes identified for each of the 22 traits by different methods. (**e**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for an exemplary trait Irritable Bowel Syndrome. (**f**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for another exemplary trait Asthma. (**g**) Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 22 traits by different methods. (**h**) Number of genes identified to have significant horizontal pleiotropic effect for each of the 22 traits by different methods. For **c**, **d**, **g**, **h**, the number on the x-axis represents 22 traits in order: Asthma, Allergic Rhinitis, CARD, Cancers, Depressive Disorder, Dermatophytosis, T2D, Dyslipidemia, HT, Hemorrhoids, Abdominal Hernia, Insomnia, Iron Deficiency, Irritable Bowel Syndrome, Macular Degeneration, Osteoarthritis, Osteoporosis, PVD, Peptic Ulcer, Psychiatric disorders, Stress Disorders, Varicose Veins.

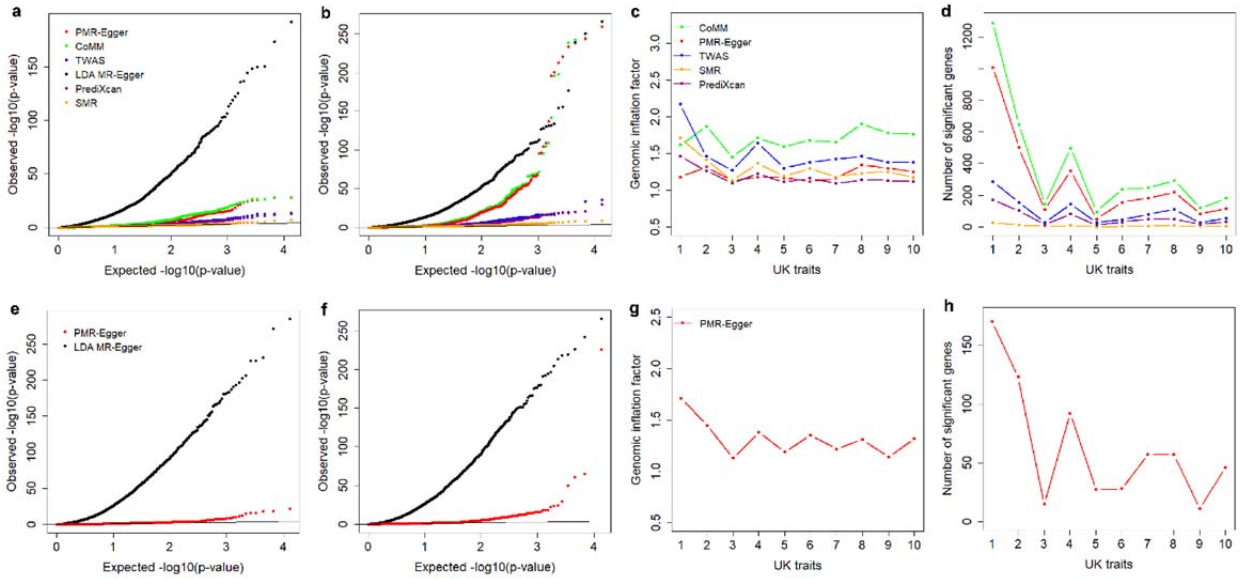

Fig. 6 TWAS analysis results by different methods for traits in the UK Biobank data. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). (**a**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for an exemplary trait BMI. (**b**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for another exemplary trait Platelet Count. (**c**) Genomic inflation factor for testing the causal effect for each of the 10 traits by different methods. (**d**) Number of causal genes identified for each of the 10 traits by different methods. (**e**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for an exemplary trait BMI. (**f**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for another exemplary trait Platelet Count. (**g**) Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 10 traits by different methods. (**h**) Number of genes identified to have significant horizontal pleiotropic effect for each of the 10 traits by different methods. For **c**, **d**, **g**, **h**, the number on the x-axis represents 10 traits in order: Height, Platelet count, Bone mineral density, Red blood cell count, FEV1-FVC ratio, BMI, RDW, Eosinophils count, Forced vital capacity, White blood cell count.