frontiers in Genetics

METHODS
published: 17 March 2021
doi: 10.3389/fgene.2021.618829

Mendelian Randomization With Refined Instrumental Variables From Genetic Score Improves Accuracy and Reduces Bias

Lijuan Lin[1]†, Ruyang Zhang[1,2,3,4]†, Hui Huang[1], Ying Zhu[1], Yi Li[5], Xuesi Dong[1,6], Sipeng Shen[1], Liangmin Wei[1], Xin Chen[1], David C. Christiani[2,7,8], Yongyue Wei[1,2,3,4] and Feng Chen[1,2,3,4]*‡*

[1] Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China, [2] China International Cooperation Center for Environment and Human Health, School of Public Health, Nanjing Medical University, Nanjing, China, [3] Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China, [4] State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, China, [5] Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States, [6] Department of Epidemiology and Biostatistics, School of Public Health, Southeast University, Nanjing, China, [7] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, United States, [8] Division of Pulmonary and Critical Care Medicine, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States

Edited by:
*Yaozhong Liu,
Tulane University, United States*

Reviewed by:
*Yuehua Cui,
Michigan State University,
United States
Paola Sebastiani,
Boston University, United States*

***Correspondence:**
*Yongyue Wei
ywei@njmu.edu.cn
Feng Chen
fengchen@njmu.edu.cn*

†These authors have contributed equally to this work

‡Senior author

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 October 2020*
Accepted: *01 February 2021*
Published: *17 March 2021*

Citation:
Lin L, Zhang R, Huang H, Zhu Y, Li Y, Dong X, Shen S, Wei L, Chen X, Christiani DC, Wei Y and Chen F (2021) Mendelian Randomization With Refined Instrumental Variables From Genetic Score Improves Accuracy and Reduces Bias. Front. Genet. 12:618829. doi: 10.3389/fgene.2021.618829

Mendelian randomization (MR) can estimate the causal effect for a risk factor on a complex disease using genetic variants as instrument variables (IVs). A variety of generalized MR methods have been proposed to integrate results arising from multiple IVs in order to increase power. One of the methods constructs the genetic score (GS) by a linear combination of the multiple IVs using the multiple regression model, which was applied in medical researches broadly. However, GS-based MR requires individual-level data, which greatly limit its application in clinical research. We propose an alternative method called Mendelian Randomization with Refined Instrumental Variable from Genetic Score (MR-RIVER) to construct a genetic IV by integrating multiple genetic variants based on summarized results, rather than individual data. Compared with inverse-variance weighted (IVW) and generalized summary-data-based Mendelian randomization (GSMR), MR-RIVER maintained the type I error, while possessing more statistical power than the competing methods. MR-RIVER also presented smaller biases and mean squared errors, compared to the IVW and GSMR. We further applied the proposed method to estimate the effects of blood metabolites on educational attainment, by integrating results from several publicly available resources. MR-RIVER provided robust results under different LD prune criteria and identified three metabolites associated with years of schooling and additional 15 metabolites with indirect mediation effects through butyrylcarnitine. MR-RIVER, which extends score-based MR to summarized results in lieu of individual data and incorporates multiple correlated IVs, provided a more accurate and powerful means for the discovery of novel risk factors.

Keywords: Mendelian randomization, multiple correlated instrumental variables, genetic score, metabolomics, educational attainment

INTRODUCTION

Observational studies have long been utilized to detect associations between the exposures of interest and the risk of complex diseases. However, the estimated effects are typically biased and causality cannot be directly inferred because of unobserved confounders or reverse causality (Ebrahim and Davey Smith, 2008). Double-blind randomized controlled trials with perfect adherence, which use randomization allocation to avoid potential confounding, are often considered as the gold standard to infer causality (Bothwell et al., 2016). However, logistical difficulties limit the use in real-world studies.

Instrumental variable (IV) analysis provides unbiased causal estimates in the presence of observed and unobserved confounders under certain assumptions (Burgess et al., 2017). A valid IV should (1) be associated with the exposure of interest; (2) not be associated with any confounders of the exposure–outcome association; and (3) affect the outcome only through its impact on the exposure of interest (**Figure 1A**; Martens et al., 2006). Because human germline genetic variants usually form at fertilization and remain unchanged after birth (Ference et al., 2019), they are less likely to be correlated with the environmental or clinical factors but can be correlated with susceptibility to these factors that are associated with outcomes and thus are ideal candidates for IVs.

Mendelian randomization (MR), which uses genetic variants as IVs, has emerged recently as a powerful tool to estimate the causal effects of risk factors in observational settings (Smith and Ebrahim, 2003; Yavorska and Burgess, 2017; Burgess and Labrecque, 2018; Bowden et al., 2019) and has been increasingly used in genome-wide association studies (GWAS) (Welter et al., 2014; Burgess et al., 2015; Pickrell et al., 2016). However, as a single variant typically explains only a small proportion of variability, a large sample size is often required to power the traditional MR (Pierce et al., 2011). A variety of generalized MR methods have been proposed to integrate results arising from multiple IVs in order to increase power (Burgess and Thompson, 2013; Burgess et al., 2013). These methods include generalized summary-data-based Mendelian randomization (GSMR) (Zhu et al., 2018) and inverse-variance weighted method (IVW) (Burgess et al., 2013, 2016). GSMR integrates estimates from single IVs by using a generalized least-square approach (Zhu et al., 2018), whereas IVW combines estimates by using weights based on the variance–covariance matrix (Burgess et al., 2016). However, these existing methods are based on the summarized results of single-variant analysis and commonly prune IVs based on linkage disequilibrium to obtain relatively independent IVs, resulting in loss of information. Even with adjustment of the correlation structure, the results may still be inefficient. Notably, Burgess et al. (2017) introduced a multivariate regression method, which regresses the exposure factor on multiple IVs at the first stage to construct genetic scores (GSs). GS can be viewed as a linear combination of multiple IVs weighted by the strength of the association between an IV and the exposure, adjusted for all the other IVs. In the ensuing MR analysis, GS will be passed along as a single IV. The method was recently implemented in a study of *ACLY* and cardiovascular disease which incorporated multiple germline genetic variants (IVs) to construct GS as single IV and further inferred the causal relationship between *ACLY* inhibitors and the reduced risk of cardiovascular disease (Ference et al., 2019).

Thus, we propose an alternative method called Mendelian Randomization with Refined Instrumental Variable from Genetic Score (MR-RIVER) (**Figure 1B**) to construct a genetic score summarizing multiple genetic variants based on summarized results rather than individual-level data. Our method, which accounts for correlations among multiple genetic variants by borrowing linkage disequilibrium (LD) information from public databases (such as 1000 Genomes Project), provides a useful framework to integrate estimates obtained by using various genetic IVs and improves the performance of the summarized genetic score for the correlated genetic variants. Simulation studies suggested improved performance of our proposed method, compared to GSMR and IVW. We further applied the proposed method to estimate the effects of blood metabolites on educational attainment, by integrating results from several publicly available resources (Shin et al., 2014; Okbay et al., 2016).

METHOD

MR-RIVER Algorithm

We propose a method to infer the causal relationship between risk factor X (e.g., blood metabolites) and outcome Y (e.g., years of schooling) given a set of IVs, denoted by $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ (e.g., a set of genetic variants). The major components of our framework are depicted in **Figure 1A**. More specifically, we use b_{XZ_i}, along with standard error $se(b_{XZ_i})$, to quantify the association of each Z_i with the risk factor X from the traditional single-locus association analysis model, and likewise for b_{YZ_i} and $se\left(b_{YZ_i}\right)$ for each Z_i with the outcome Y.

The unified weighted GS incorporating multiple IVs could be estimated by the linear combination of multiple IVs:

$$GS = \sum_{i=1}^{p} \tilde{b}_{XZ_i} Z_i \qquad (1)$$

Where $\tilde{b}_{XZ_i}$ denotes the direct effect of Z_i on X after controlling for the other IVs that derived from multivariable regression. However, in practice, the published-available summarized data were derived from single-variant analysis; it is unlikely to get genetic association estimates from a multivariable regression model in a large independent dataset due to issues of practicality and confidentiality of data sharing on such a large scale. Here, we propose an estimator by borrowing the idea of coefficient decomposition to estimate $\tilde{b}_{XZ_i}$ by using summarized results rather than individual-level data.

Specifically, under the assumption that $(X, \mathbf{Z})$ follow a multivariate normal distribution, regressing X on each Z_i will yield an estimate of b_{XZ_i}. Without loss of generality, we assume that there is a linear relationship between X and $\mathbf{Z}$. As $E\left(X|Z_i\right) = b_0 + b_{XZ_i} Z_i$, b_{XZ_i} represents the total effect of Z_i on X.

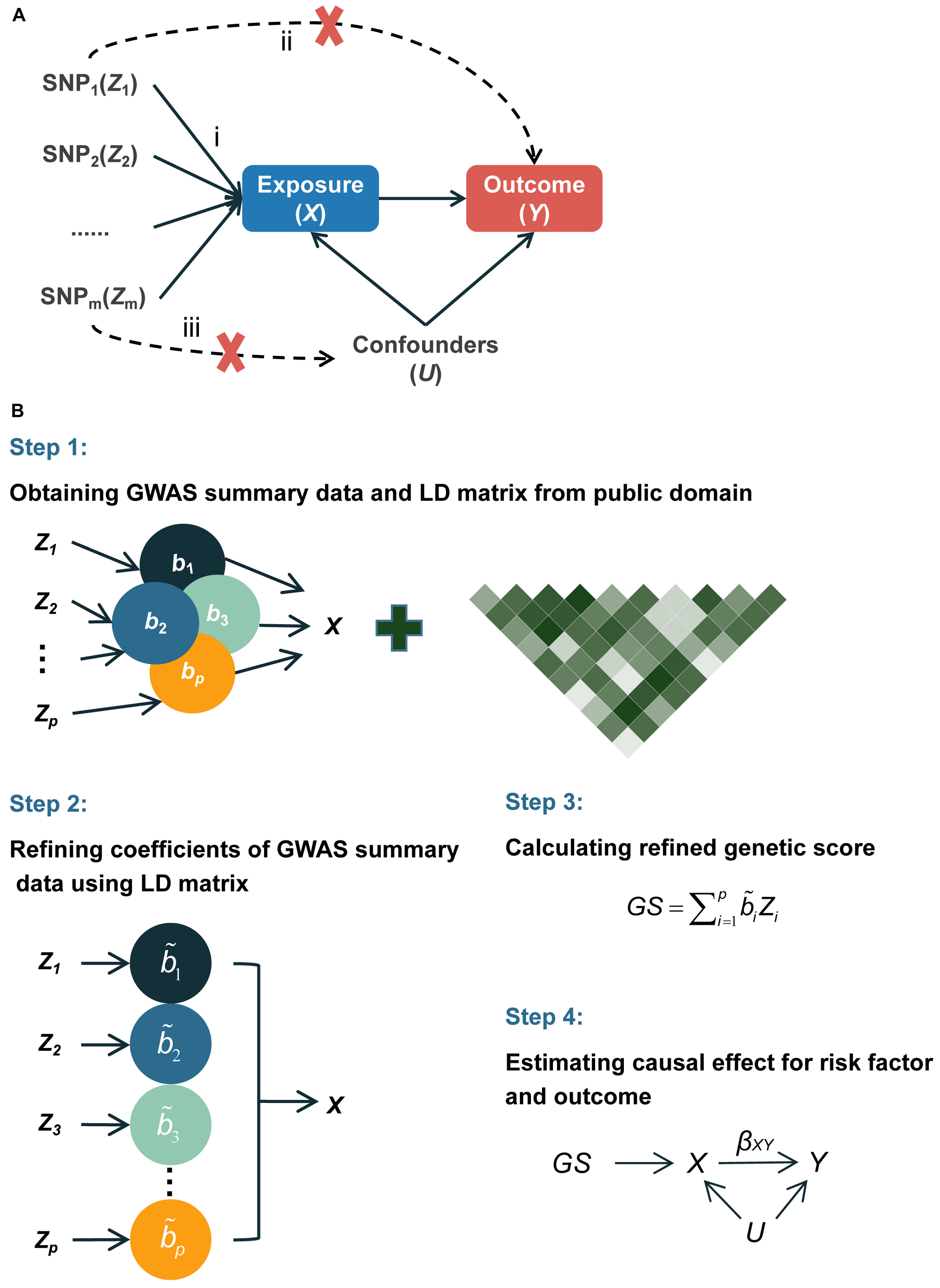


FIGURE 1 | Diagram of Mendelian randomization and flowchart of the proposed MR-RIVER method. **(A)** Mendelian randomization inferring the causal association of the exposure and outcome: (i) IVs are associated with the exposure X; (ii) IVs and outcome *Y* are independent, conditional on exposure *X* and unmeasured confounders *U*; (iii) IVs and confounders *U* are independent. **(B)** Flowchart of the proposed MR-RIVER method for multiple genetic variants in causal inference.

After adjusting the effect of all the other IVs, the relationship between X and Z_i can be expressed as $E\left(X|Z_1, \cdots, Z_p\right) = b_0 + \tilde{b}_{XZ_1} Z_1 + \cdots + \tilde{b}_{XZ_p} Z_p$, where $\tilde{b}_{XZ_i}$ is the direct effect of Z_i on X under the control of other IVs. Therefore, b_{XZ_i} can be decomposed into the direct effect and indirect effect via other correlated IVs:

$$b_{XZ_i} = \tilde{b}_{XZ_i} + \sum_{j \neq i}^{p} \tilde{b}_{XZ_j} \theta_{Z_j Z_i} \qquad (2)$$

Here, $\theta_{Z_jZ_i}$ is the regression coefficient of Z_j on Z_i, and $\tilde{b}_{XZ_i}$ is the direct effect of Z_i on X, after controlling for the other IVs. Equation 2 can be rewritten as:

$$\tilde{\mathbf{b}}_{XZ} = \boldsymbol{\theta}^{-1}\mathbf{b}_{XZ} \tag{3}$$

where $\mathbf{b}_{XZ}$ is the p-length vector containing b_{XZ_i}, $\tilde{\mathbf{b}}_{XZ}$ is the vector of refined coefficients $\tilde{b}_{XZ_i}$, and θ is a $p \times p$ matrix with $\theta_{Z_j\,Z_i}$ being the ij-th entry. It follows that $\theta_{Z_iZ_j} = \rho_{Z_jZ_i}\sqrt{var\left(Z_j\right)/var\left(Z_i\right)}$ where $\rho_{Z_j\,Z_i}$ is the correlation between Z_j and Z_i, $var(Z_i)$ is the variance of Z_i. $var(Z_i)$ and $\rho_{Z_jZ_i}$ can be obtained from the public GWAS resources (e.g., 1000 Genomes Project).

We note that Eq. 3 is crucial as it enables us to compute GS defined in Eq. 1 with only summary data, in lieu of individual-level data. With GS as a single IV, we can estimate the association between the risk factor X and outcome Y with:

$$\begin{aligned}\hat{\beta}_{XY} &= \frac{\beta_{YGS}}{\beta_{XGS}} \\ &= \frac{cov(Y,GS)}{cov(X,GS)} = \frac{cov(Y,\sum_i^p \tilde{b}_{XZ_i}Z_i)}{cov(X,\sum_i^p \tilde{b}_{XZ_i}Z_i)} \\ &= \frac{\sum_i^p \tilde{b}_{XZ_i}\,cov(Y,Z_i)}{\sum_i^p \tilde{b}_{XZ_i}\,cov(X,Z_i)} = \frac{\sum_i^p \tilde{b}_{XZ_i}b_{YZ_i}\,var(Z_i)}{\sum_i^p \tilde{b}_{XZ_i}b_{XZ_i}\,var(Z_i)}\end{aligned} \tag{4}$$

As mentioned by Burgess et al. (2016), $var\,(Z_i)$ is approximately proportional to $1/var\left(b_{YZ_i}\right)$; thus, Eq. 4 can be simplified as:

$$\hat{\beta}_{XY} = \frac{\sum_i^p \tilde{b}_{XZ_i}b_{YZ_i}\Big/var\left(b_{YZ_i}\right)}{\sum_i^p \tilde{b}_{XZ_i}b_{XZ_i}\Big/var\left(b_{YZ_i}\right)} \tag{5}$$

The asymptotic standard error for $\hat{\beta}_{XY}$ can be estimated by the delta method (Thomas et al., 2007):

$$se\left(\hat{\beta}_{XY}\right) = \sqrt{\frac{\sum_i^p\sum_j^p \rho_{Z_iZ_j}\tilde{b}_{XZ_i}\tilde{b}_{XZ_j}\Big/\left(se\left(b_{YZ_i}\right)se\left(b_{YZ_j}\right)\right)}{\left(\sum_i^p \tilde{b}_{XZ_iXZ_j}\Big/var\left(b_{YZ_i}\right)\right)^2}} \tag{6}$$

The association between X and Y can be further tested by using the Wald test statistic $u = \hat{\beta}_{XY}\Big/se\left(\hat{\upbeta}_{XY}\right)$, which asymptotically follows a standard normal distribution under the null hypothesis.

We stress that, though Eqs.5, 6 resemble the estimator proposed in Burgess et al. (2017), our estimator differs from that in Burgess et al.'s (2017) required individual data, while our estimator, with the introduction of the refined estimates in Eq. 3, can be computed even with the summary data. Therefore, our estimator is applicable in more broad settings, where only summary data are available. Simulations have confirmed the utility of our method.

Design of Statistical Simulations

Two sets of simulation studies were designed to investigate MR-RIVER.

Evaluation of the Estimates of the Refined Coefficients of IVs on *X*

We generated six IVs, $Z_1, Z_2, \ldots, Z_6$, from a multivariate normal distribution $MVN\,(\mathbf{0},\,\Sigma)$, where Σ is a correlation matrix with an equal correlation structure. We varied the correlation coefficient and set it to be 0, 0.1, 0.3, 0.5, 0.7, and 0.9, corresponding to various scenarios: from the independent case to the highly correlated case. We generated X using the following models:

$$\begin{aligned}&X_i = \textstyle\sum_{j=1}^{6} Z_{ij}\tilde{b}_j + \varepsilon_{Xi} \\ &\tilde{b}_j \sim N\,(\mu,\ 1)\,, \quad \mu = -1,\ -0.5,\ 0.5,\ 1,\ 1.5,\ 2 \\ &\varepsilon_{X_i} \sim N\,(0,1)\end{aligned} \tag{7}$$

The sample size was fixed at 1,000. In addition, we simulated 5,000 additional individuals to provide an external correlation structure for IVs. For each simulation configuration, 2,000 datasets were produced.

We first regressed X on each Z_j separately to obtain the summarized effect of Z_j on X, and based on these results, we applied Eq. 2 to obtain the estimates of the refined coefficients. The estimated refined coefficients, along with the corresponding standard errors, were compared to those from traditional GWAS summarized results under different correlation structures and effect sizes of Z.

Investigation of the Statistical Properties of MR-RIVER

Let X_i and Y_i denote the exposure and outcome variables of the ith subject, and Z_{ij} the jth IV (j = 1, . . ., J). The data were generated from the following model:

$$\begin{aligned}&Z_i \sim MVN\,(0,\ \Sigma)\,, \quad b_j \sim U\,(0,\ 0.5) \\ &X_i = \textstyle\sum_{j=1}^{J} Z_{ij}b_j + \upvarepsilon_{Xi} \\ &Y_i = X_ib_{XY} + \upvarepsilon_{Yi} \\ &where \quad \upvarepsilon_{Xi} \sim N\left(0, var\left(\textstyle\sum_{j=1}^{J} Z_{ij}b_j\right)\left(R_{ZX}^{-2} - 1\right)\right) \\ &and \quad \upvarepsilon_{Yi} \sim N\left(0, var\left(X_ib_{XY}\right)\left(R_{XY}^{-2} - 1\right)\right)\end{aligned} \tag{8}$$

where Σ is the correlation matrix of IVs with an equal correlation structure. We varied the correlation parameter from 0 to 0.9 by 0.1. Each IV explains 0.005 of the variance of X, and we considered J = 5, 10, 15, 20. Moreover, R_{ZX}^2 is the proportion of variance of X explained by all IVs, which was set to be 0.025, 0.05, 0.075, and 0.1, while R_{XY}^2 is the proportion of variance of Y explained by X, which was set to be 0.05, 0.1, 0.15, and 0.2. Sample sizes for the IV-exposure association study (N_1) and the IV-outcome association study (N_2) were set to be 1,000 and 1,500, respectively. In addition, 5,000 (N_3) individuals were generated to provide an external correlation structure for genetic variants.

For each parameter configuration, a total of 2,000 datasets were produced. Under all the scenarios examined, MR-RIVER was found to outperform GSMR and IVW by maintaining the Type I error, possessing more statistical power, as well as having smaller biases and mean squared errors.

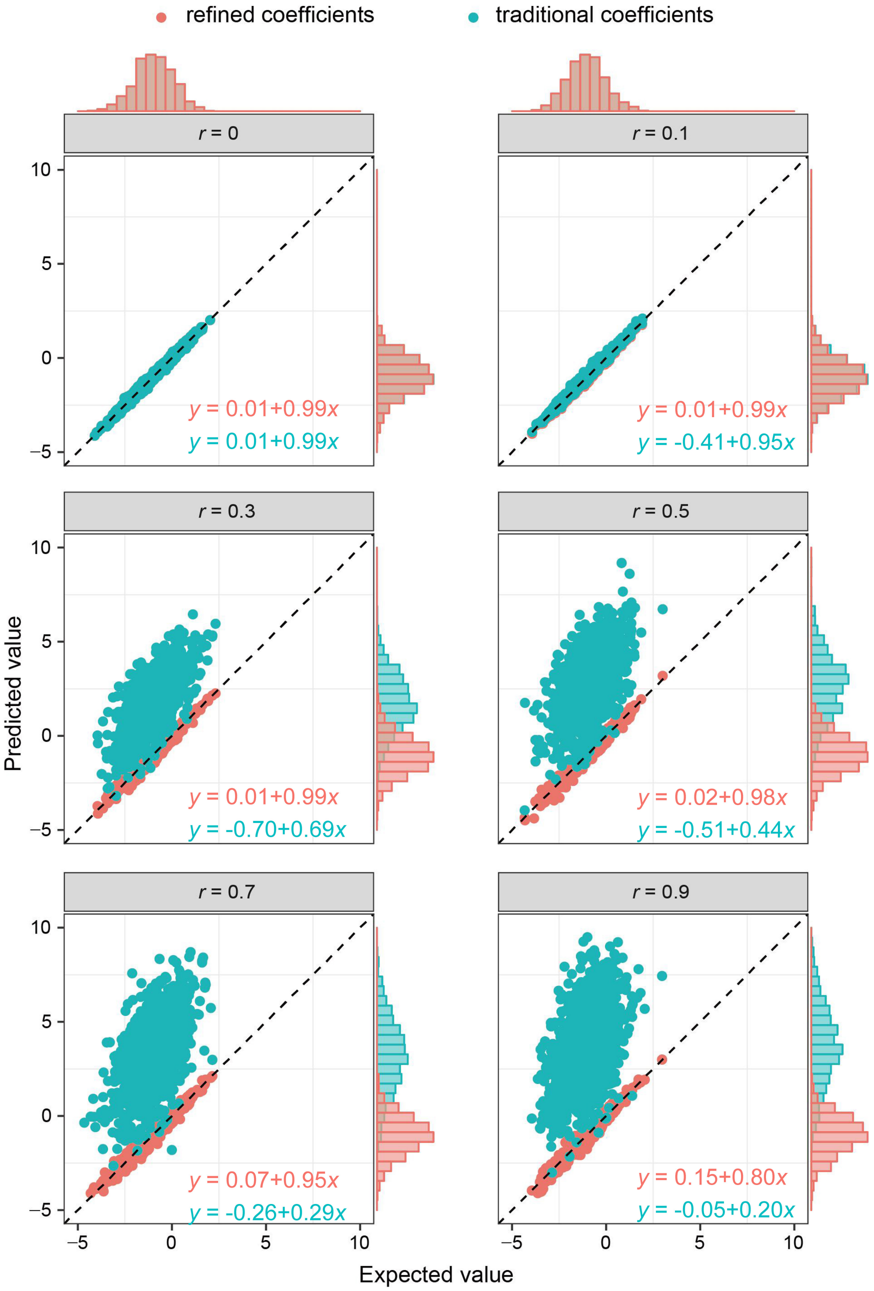


FIGURE 2 | Comparison of refined and traditional coefficients under different correlation structures. Expected values are the regression coefficients obtained from the multivariable regression model with all the variables used to generate dependent variable Y plotted against predicted values obtained from the refined method (refined coefficients) and traditional single-locus analyses (traditional coefficients). Refined and traditional coefficients were compared with the bias from expected coefficients under different correlation structures through a regression model. Red equation represents the relationship between expected coefficients and refined coefficients, and green equation represents traditional coefficients.

RESULTS

Statistical Properties of Refined Coefficients

We investigated the accuracy of refined coefficients. With the obtained correlation structure of IVs from the internal analysis set, the estimated refined coefficients (along with the standard errors) based on the summarized results were in consistent with the corresponding estimates from multivariable regression (**Supplementary Figures 1A,B**), suggesting that the estimates of the refined coefficients were unbiased.

As the key of the approach lies in borrowing the correlation information from public resources, we further evaluated

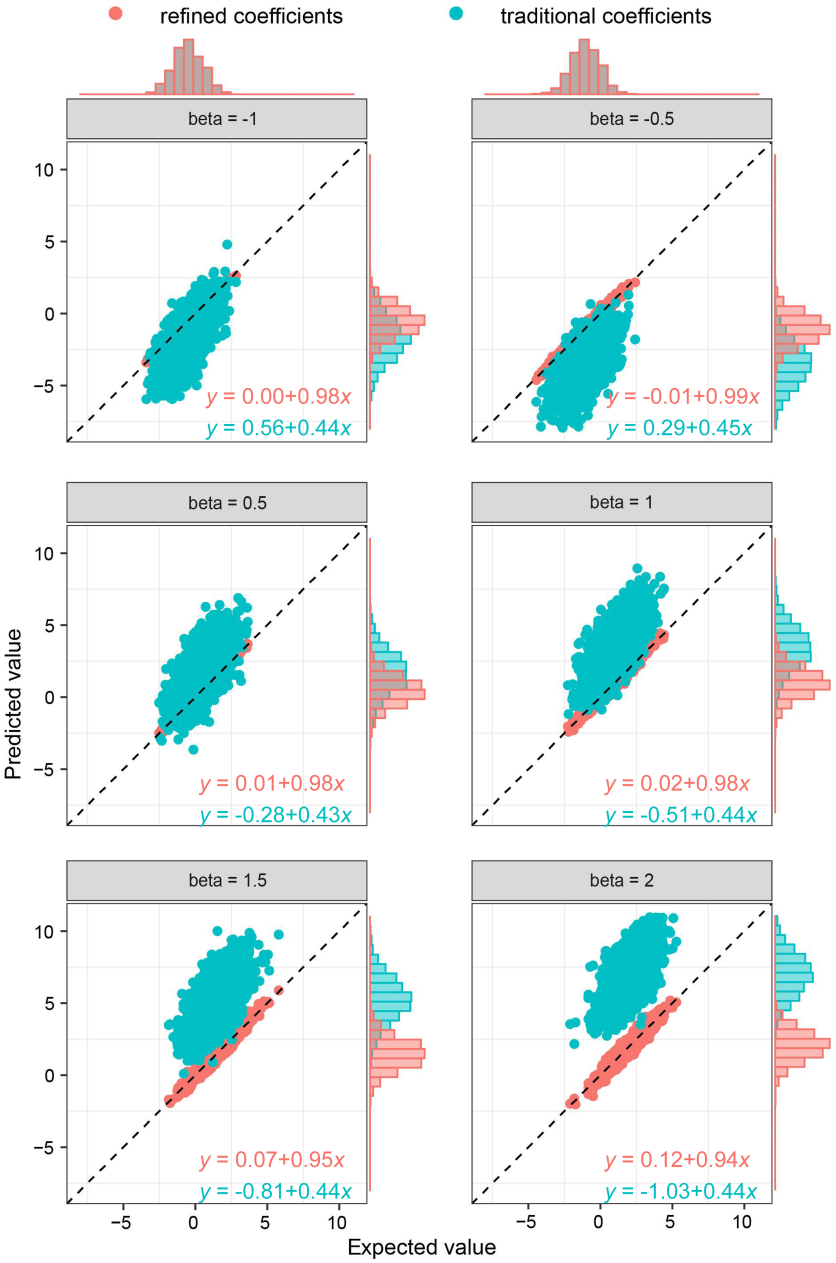


FIGURE 3 | Comparison of refined and traditional coefficients under different effect sizes. Expected values are the regression coefficients obtained from the multivariable regression model with all the variables used to generate dependent variable *Y* plotted against predicted values obtained from the refined method (refined coefficients) and traditional single-locus analyses (traditional coefficients). Refined and traditional coefficients were compared with the bias from expected coefficients under different effect sizes through a regression model. Red equation represents the relationship between expected coefficients and refined coefficients, and green equation represents traditional coefficients.

the method by obtaining the correlation structure from the simulated external samples. According to different levels of correlation among IVs, refined coefficients outperformed traditional coefficients obtained from single-locus analysis, especially when the correlations among IVs were relatively high (**Figure 2**). Similarly, under the specific correlation structure (with a correlation coefficient of 0.5), refined coefficients remained approximately unbiased, while traditional coefficients showed increased biases with increased effect sizes (**Figure 3**).

Statistical Properties of MR-RIVER

With various strengths of correlations among IVs, MR-RIVER maintained the type I error at the 0.05 level, compared to the IVW (with type I error around 0.04) and GSMR (with the most conservative control of the type I error) (**Figure 4A**). The results

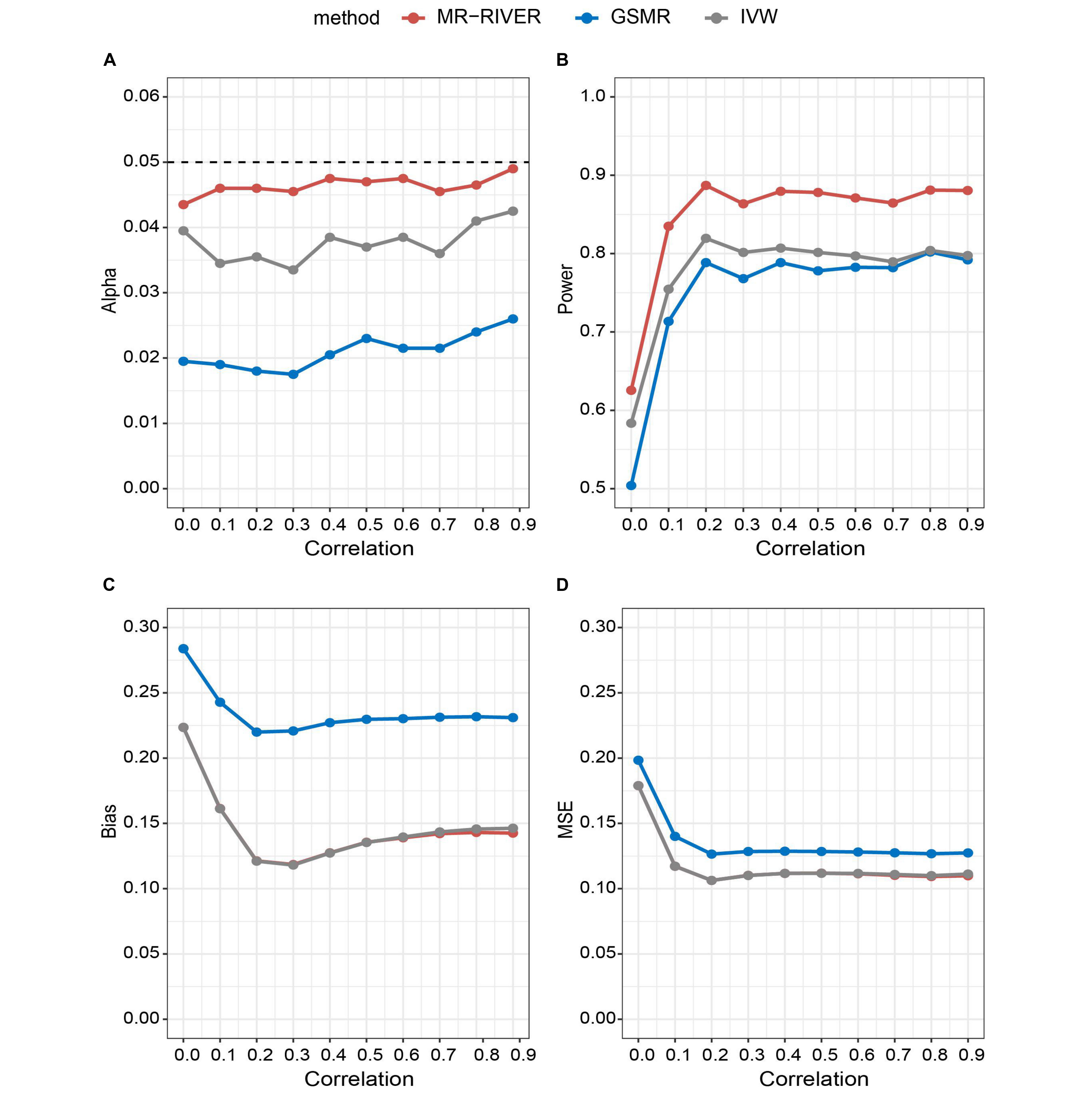


FIGURE 4 | Statistical properties of MR methods under different correlations. Correlation between IVs plotted against: **(A)** type I error under the null hypothesis; **(B)** performance of power under the alternative hypothesis with b_{xy} = 1; **(C)** bias under the alternative hypothesis; and **(D)** mean square error.

held when we varied the sample size (**Supplementary Figure 2A**) or the number of IVs (**Supplementary Figure 3A**). Further, increasing correlation strengths among IVs (**Figure 4B**), or increasing sample size (**Supplementary Figure 2B**), or increasing the numbers of IVs (**Supplementary Figure 3B**) led to increased power for all MR methods. Overall, the power of MR-RIVER was higher than that of GSMR and IVW under different parameter settings.

Estimates of b_{xy} from the three MR methods were approximately unbiased, while the biases of the MR-RIVER and IVW estimates were lower than that of the GSMR estimate (**Figure 4C**). The bias increased with the increased effect size (**Supplementary Figure 4**) and so was true for the MSE (**Figure 4D**). MR-RIVER and IVW had lower biases and MSEs, compared to GSMR.

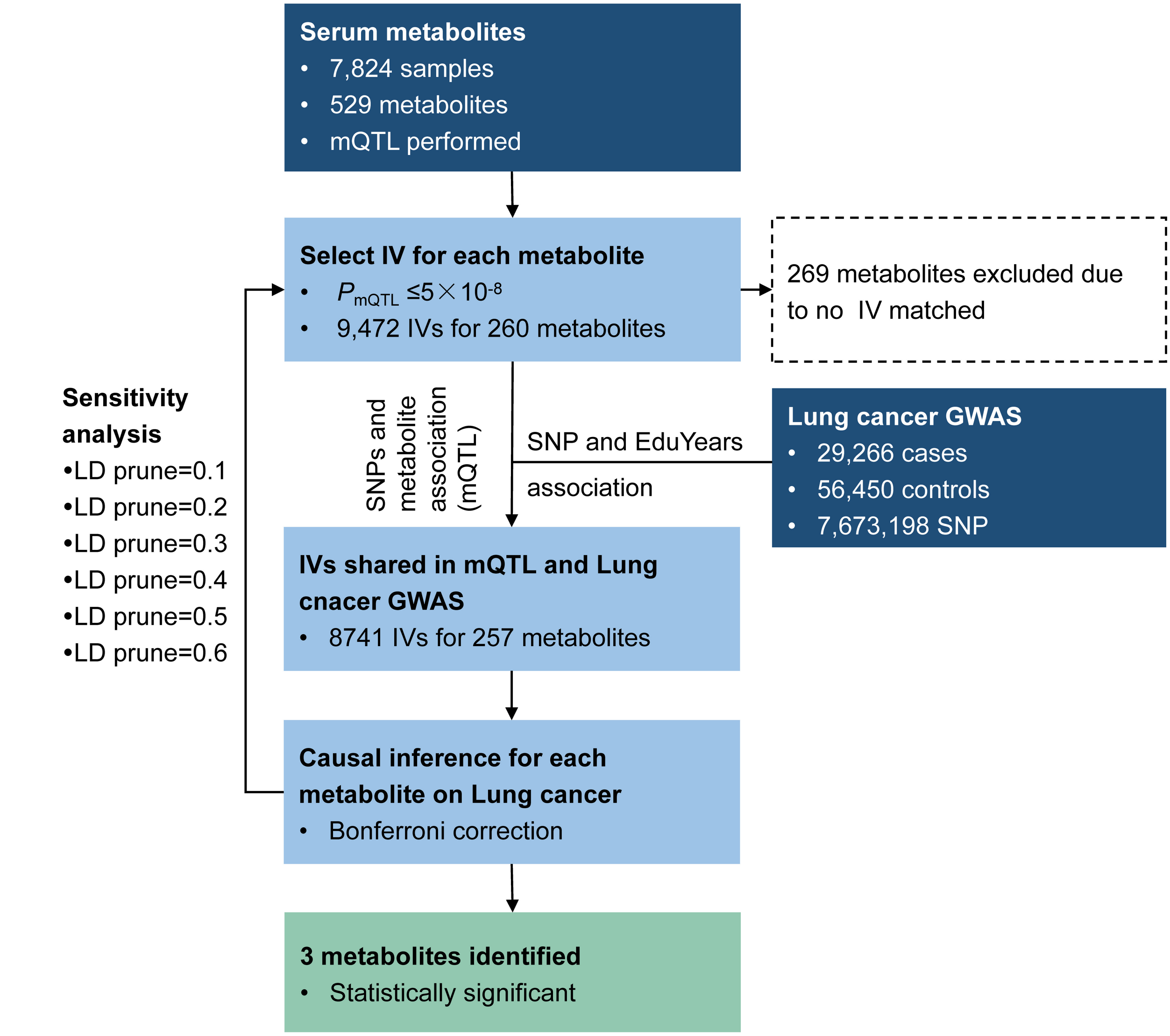

Serum metabolites
• 7,824 samples
• 529 metabolites
• mQTL performed
Select IV for each metabolite
• $P_{\text{mQTL}} \leq 5 \times 10^{-8}$
• 9,472 IVs for 260 metabolites
269 metabolites excluded due to no IV matched
Sensitivity analysis
•LD prune=0.1
•LD prune=0.2
•LD prune=0.3
•LD prune=0.4
•LD prune=0.5
•LD prune=0.6
SNPs and metabolite association (mQTL)
SNP and EduYears association
Lung cancer GWAS
• 29,266 cases
• 56,450 controls
• 7,673,198 SNP
IVs shared in mQTL and Lung cnacer GWAS
• 8741 IVs for 257 metabolites
Causal inference for each metabolite on Lung cancer
• Bonferroni correction
3 metabolites identified
• Statistically significant

FIGURE 5 | Study workflow for educational attainment MR analysis.

REAL DATA APPLICATION

Motivation

Educational attainment is moderately heritable and has been recognized as a proxy phenotype for intelligence, cognition, and neuropsychiatric disorders (Berry et al., 2006; Esch et al., 2014). Discovery of the causal factors linking to the educational attainment could shed light on the biological pathways underlying human behavioral and health-related outcomes (Rietveld et al., 2013). Blood metabolites, which closely represent the physiological status of an organism, have garnered significant interest in biomedical research (Simpson et al., 2016). However, few studies have focused on a causal relationship between metabolites and educational attainment in the presence of multiple IV variables. Taking advantage of the proposed MR-RIVER, this application aims to systematically evaluate the causal relationship between blood metabolites and educational attainment using multiple GWAS summary results.

Materials

Genome-wide association studies summary results for educational attainment were obtained based on various studies from the Social Science Genetic Association Consortium[1] (Berry et al., 2006; Rietveld et al., 2013). Educational attainment was measured as the year of schooling completed (EduYears) among 293,723 individuals (with a mean of 14.3 years) (**Supplementary Table 1**). Approximately, 9.3 million SNPs were included in the association analysis, and minor allele frequencies were obtained from the 1000 Genomes Project. Details of the SNPs included in our analysis are displayed in **Supplementary Table 2**.

[1] https://www.thessgac.org/data

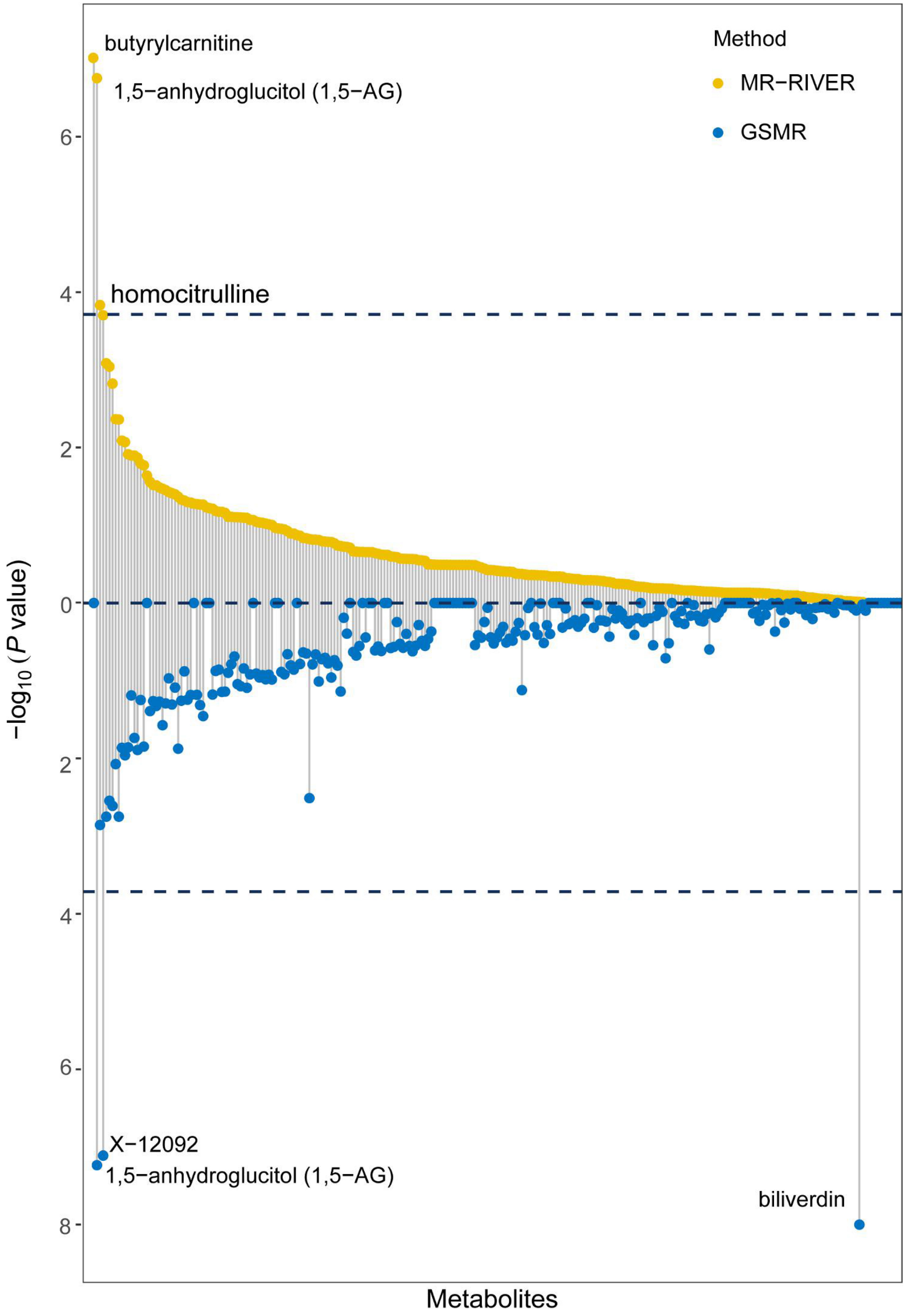


FIGURE 6 | MR-RIVER and GSMR analysis for causal association between metabolites and educational attainment. Relationship between individual metabolites with $-\log_{10}$ (*P*-value) of the association. Upper yellow values represent MR-RIVER results, and lower blue values represent GSMR results. Associated metabolites are annotated.

Summary results of quantitative trait locus (QTL) analysis of SNPs on corresponding metabolites were obtained from 7,824 European adult individuals (**Supplementary Table 3**) (Shin et al., 2014). Specifically, the metabolite QTL (mQTL) data contained all of the summarized association statistics for 529 metabolites with P-values less than 1×10^{-52}.A total of 196 metabolites out of 529 (37%) were unknown because their chemical identity was not yet determined at the time of analysis. Detailed information of metabolites can be found in **Supplementary Table 4**.

MR Analysis Results

We applied the method to explore the causal effect of blood metabolites on educational attainment as depicted in **Figure 5**. Based on assumption (1) of IV, SNPs were required to have an mQTL relationship with the corresponding metabolites with *P-values* less than 5×10^{-8}. As a result, 9,472 SNPs were selected as IVs, matched with 260 metabolites. Among these, 9,329 SNPs were available in the educational attainment GWAS.

Causal inference for each metabolite on quantitative education years was evaluated through MR-RIVER and GSMR. To obtain sufficient IVs to increase the power of MR, IVs were pruned by LD at 0.5; The HEIDI-outlier test was used to detect pleiotropic SNPs and remove them from the MR analysis; see **Figure 6**. Bonferroni correction was used to control for false positives. MR-RIVER identified three metabolites associated with education years: butyrylcarnitine (b_{xy} = −0.043, P = 1.08 × 10^{-7}), 1,5-anhydroglucitol (1,5-AG) (b_{xy} = −0.192, P = 1.77 × 10^{-7}), and homocitrulline (b_{xy} = −0.269, P = 1.47 × 10^{-4}). GSMR identified biliverdin (b_{xy} = −0.028, P = 2.92 × 10^{-15}), 1,5-AG (b_{xy} = −0.183, P = 5.83 × 10^{-8}), and an unknown metabolite, X-12092 (retention time, 1.130; mass-to-charge ratio, 144.2; spectra, 84.2:0.8) (b_{xy} = 0.028, P = 3.85 × 10^{-7}) (**Table 1**). In addition, sensitivity analyses with different LD prune criteria (0.1–0.7, in 0.1 increments) showed robust results for MR-RIVER, but not for GSMR (**Supplementary Tables 5, 6**).

We performed additional analyses to explore whether the remaining metabolites affected education years through the above-identified candidate metabolites. SNPs associated with the remaining metabolites were treated as IVs to infer potential causal associations between the identified metabolites and remaining metabolites (**Figure 7A**). The results indicated 28 additional metabolites were associated with the three candidate metabolites. Among these, 24 metabolites (including six unknown metabolites) were associated with butyrylcarnitine, three unknown metabolites were associated with 1,5-AG, and one unknown metabolite was associated with homocitrulline (**Supplementary Table 7**).

Further, mediation analysis was used to evaluate potential metabolic regulatory pathways for education years by Sobel test (Baron and Kenny, 1986). The 15 metabolites indirectly mediated the effect on education years through butyrylcarnitine (**Figure 7B** and **Supplementary Table 7**). Most metabolites were located in the carnitine metabolism pathway (8/15, 53.0%). Blood metabolic biomarkers overall formed a potential causal network (**Figure 7C**).

TABLE 1 | Relative bias of imputed datasets with three imputation methods.

Method	Metabolite	b_{xy}	se of b_{xy}	P-value
MR-RIVER	Butyrylcarnitine	−0.0430	0.0081	1.08 × 10^{-07}
	1,5-Anhydroglucitol (1,5-AG)	−0.1916	0.0367	1.77 × 10^{-07}
	Homocitrulline	−0.2687	0.0708	1.47 × 10^{-04}
GSMR	Biliverdin	−0.0284	0.0036	2.92 × 10^{-15}
	1,5-Anhydroglucitol (1,5-AG)	−0.1838	0.0339	5.83 × 10^{-08}
	X-12092	0.0283	0.0056	3.85 × 10^{-07}

b_{xy}: causal effect of metabolite and educational attainment.
se of b_{xy}: standard error of causal effect.
P value: P-value of causal effect.
X-12092: unknown metabolite (retention time, 1.130; mass-to-charge ratio, 144.2; spectra, 84.2:0.8).

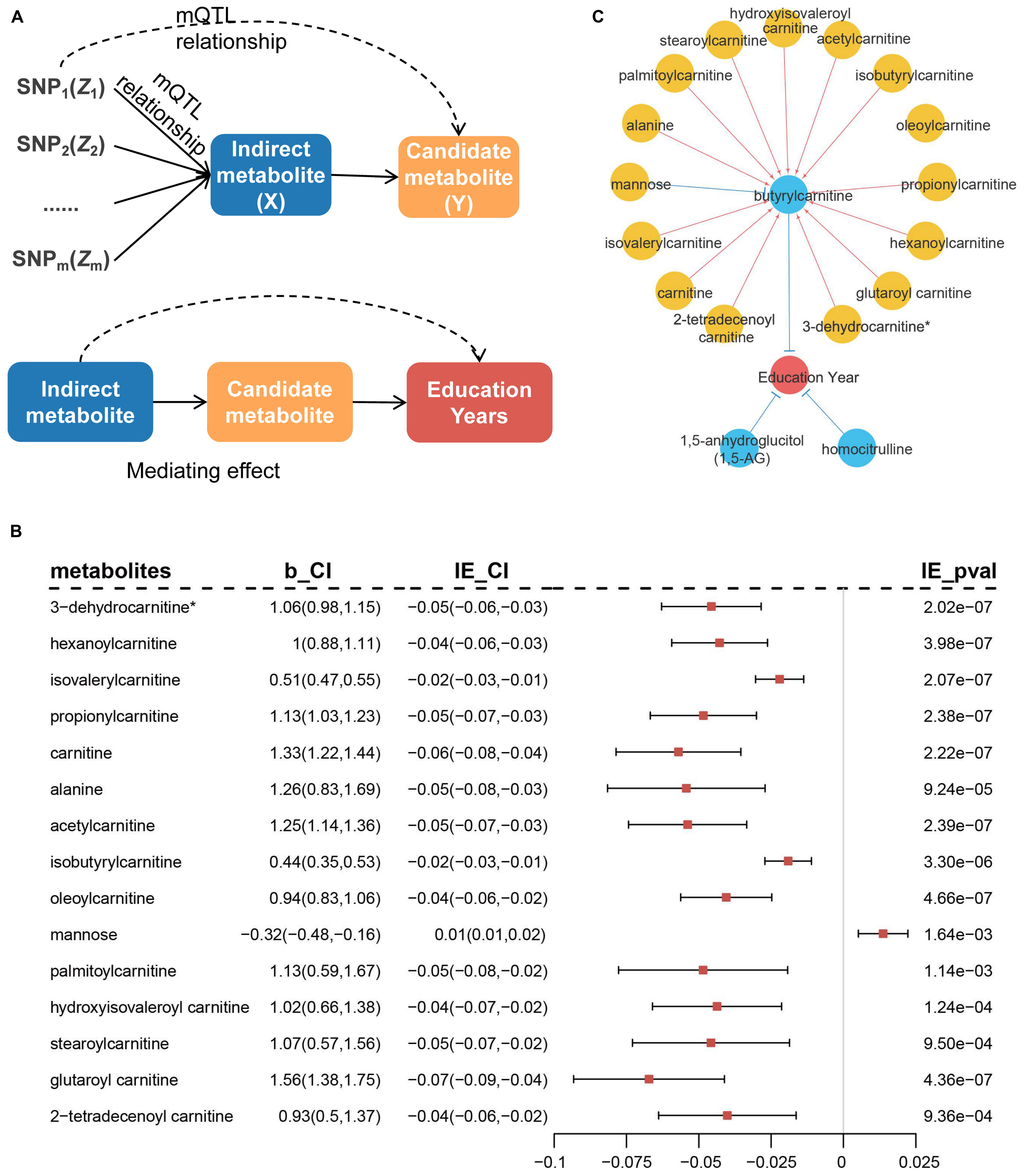
DISCUSSION

We proposed an improved MR approach, MR-RIVER, to combine summarized results of multiple IVs into a single GS and to estimate the unbiased causal effect of a risk factor on an outcome. The publicly accessible summary-level data were obtained from single-locus analyses without consideration of the correlation between IVs. MR-RIVER provides a novel way to refine the effect size of genetic variants account for the correlation based on summary data and makes it efficient to perform summarized data genetic score MR when the correlation between IVs are unignorable. MR-RIVER closely maintains the type I error around the nominal level while it has higher power, lower bias, and smaller variation compared to GSMR and IVW.

Genome-wide association studies uses original GWAS summarized results for IV exposure and IV outcome obtained from single-locus analyses and then derives the causal effect by the generalized least-square approach weighted by the variance–covariance matrix to adjust for correlations among IVs (Zhu et al., 2018). MR-RIVER instead first modifies the summarized results, accounting for correlations among IVs, and then integrates the results. Thus, there are several differences between MR-RIVER and GSMR. First, MR-RIVER adjusts summarized results for each genetic IV by borrowing external LD information to obtain more accurately estimate IV-exposure effect—therefore, MR-RIVER has an advantage in accuracy. Second, MR-RIVER aggregates multiple IVs by weighted linear combination weighted by refined coefficients, which reduces the dimension for IVs and simplifies the following calculation.

Interestingly, MR-RIVER and IVW showed similar performance in bias and MSE. If the weights used to aggregate multiple IVs are equal to the original GWAS summary results ($\tilde{b}_{XZ_i} = b_{XZ_i}$ in Eq. 5), then MR-RIVER is the same as IVW. On the one hand, estimates of MR-RIVER are approximately identical to IVW because point estimates are robust toward the weights (**Supplementary Figure 5A**). On the other hand, different weights result in different standard errors

[2] http://metabolomics.helmholtz-muenchen.de/gwas



metabolites	b_CI	IE_CI	IE_pval
3−dehydrocarnitine*	1.06(0.98,1.15)	−0.05(−0.06,−0.03)	2.02e−07
hexanoylcarnitine	1(0.88,1.11)	−0.04(−0.06,−0.03)	3.98e−07
isovalerylcarnitine	0.51(0.47,0.55)	−0.02(−0.03,−0.01)	2.07e−07
propionylcarnitine	1.13(1.03,1.23)	−0.05(−0.07,−0.03)	2.38e−07
carnitine	1.33(1.22,1.44)	−0.06(−0.08,−0.04)	2.22e−07
alanine	1.26(0.83,1.69)	−0.05(−0.08,−0.03)	9.24e−05
acetylcarnitine	1.25(1.14,1.36)	−0.05(−0.07,−0.03)	2.39e−07
isobutyrylcarnitine	0.44(0.35,0.53)	−0.02(−0.03,−0.01)	3.30e−06
oleoylcarnitine	0.94(0.83,1.06)	−0.04(−0.06,−0.02)	4.66e−07
mannose	−0.32(−0.48,−0.16)	0.01(0.01,0.02)	1.64e−03
palmitoylcarnitine	1.13(0.59,1.67)	−0.05(−0.08,−0.02)	1.14e−03
hydroxyisovaleroyl carnitine	1.02(0.66,1.38)	−0.04(−0.07,−0.02)	1.24e−04
stearoylcarnitine	1.07(0.57,1.56)	−0.05(−0.07,−0.02)	9.50e−04
glutaroyl carnitine	1.56(1.38,1.75)	−0.07(−0.09,−0.04)	4.36e−07
2−tetradecenoyl carnitine	0.93(0.5,1.37)	−0.04(−0.06,−0.02)	9.36e−04

FIGURE 7 | Diagram of MR analysis between metabolites and mediation analysis. **(A)** MR inferring the causal association of remaining metabolites (*X*) on previously identified metabolites (*Y*). Mediation analysis of the rest of metabolites on risk of education years through the identified metabolites. **(B)** Metabolites that indirectly mediate the effect on education years through butyrylcarnitine in mediation analysis. b_CI represents effect of metabolites on butyrylcarnitine and 95% confidence interval (95% CI). IE_CI represents indirect effect of metabolites on education years and 95% CI, and IE_pval represents *P*-values. **(C)** Causal network of blood metabolites on education years. Blue circles indicate metabolites that are directly identified, while yellow circles have indirect effect through blue metabolites. Red lines represent positive effects, and blue lines indicate negative effects.

(**Supplementary Figure 5B**), which in turn lead to different statistics (**Supplementary Figure 5C**). This may explain why the bias and MSE of MR-RIVER and IVW are similar, but the performance of power and type I error is different. To summarize, MR-RIVER improves upon IVW and is powerful to infer a causal relationship between an exposure and outcome.

There has been much discussion on the potentials and limitations of MR, as the IV assumptions cannot be fully tested (VanderWeele et al., 2014; Paternoster et al., 2017). Horizontal pleiotropy is a common phenomenon in the human genome that some genetic variants affect the outcome through other traits or pathways rather than exclusively through the risk factor (Solovieff et al., 2013). It is a violation of the instrumental variable assumptions and may induce a major source of potential bias in causal inference. There are several methods are proposed to detect pleiotropy (Slob and Burgess, 2020). The MR-Egger method is able to assess the pleiotropic effects as well as to provide a consistent estimate of the causal effect (Bowden et al., 2017), while the estimates were generally imprecise with low power (Slob and Burgess, 2020). The HEIDI-outlier test was proposed to detect heterogeneity at multiple correlated instruments (Zhu et al., 2018). It will be powerful and valuable when only some proportion of the SNPs have a horizontal pleiotropy effect. In our proposed method, we ensembled the HEIDI-outlier test to detect potential pleiotropy and then remove them from the MR-RIVER analysis.

Notably, after GWAS significant threshold screening, LD prune, and HEIDI-outlier filtering, MR-RIVER analysis suggested three causal metabolites that are associated with education years. The first metabolite is butyrylcarnitine, classified as an acylcarnitine. Previous studies have shown that abnormally increased levels of acylcarnitines, including butyrylcarnitine, are associated with fatty acid oxidation disorders (Jones et al., 2010). Elevated butyrylcarnitine concentration in plasma is associated with short-chain acyl-CoA dehydrogenase deficiency (van Maldegem et al., 2006), which may cause failure to thrive, developmental and cognitive delay, seizures, and neuromuscular (Corydon et al., 2001). Moreover, fatty acid oxidation disorders may lead to mitochondrial dysfunction and further affect the energy supply of the brain (Kölker et al., 2004; Wajner and Amaral, 2015). Therefore, high levels of acylcarnitines may be involved in potential metabolic regulatory pathways affecting cognitive status or brain energy supplement and, in turn, increased education years (mannose→butyrylcarnitine→education years). Mannose easily crosses the blood–brain barrier and is converted to fructose-6-phosphate that enters the glycolytic pathway (Sharma et al., 2014). Cerebral tissue can utilize mannose directly and rapidly from the blood to restore or maintain normal metabolic functions in the absence of glucose (Sloviter and Kamimoto, 1970). Taken altogether, mannose levels appear to be a potential beneficial factor for education years.

The second metabolite, 1,5-AG, is a monosaccharide structurally similar to glucose and is a validated marker of short-term glycemic control (Buse et al., 2003). Low levels of 1,5-AG, indicative of glycemic peak, are associated with dementia and cognitive decline (Rawlings et al., 2017). Finally, elevated homocitrulline, the third metabolite, is structurally similar to but one methylene group longer than citrulline, and impairs bioenergetics in the brain cortex, by reducing velocity of the citric acid cycle and creatine kinase activity. Consequently, it decreases energy production and transfer (Viegas et al., 2009). Therefore, administration of 1,5-AG and homocitrulline may improve educational attainment.

In conclusion, the proposed MR-RIVER method appears to outperform the existing commonly used MR methods. With publicly accessible summary-level data, MR-RIVER provides a more accurate and powerful mean for novel discoveries and identifies several blood metabolites as biomarkers and interventional targets for educational attainment.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GWAS summary results for education attainment are available at https://www.thessgac.org/data; summary results of quantitative trait locus (QTL) analysis of SNPs on metabolites are available at http://metabolomics.helmholtz-muenchen.de/gwas.

AUTHOR CONTRIBUTIONS

LL, RZ, YW, and FC contributed the study design. LL, YW, and RZ performed the statistical analysis and interpretation and drafted the manuscript. YW, LL, RZ, YL, XD, SS, HH, YZ, LW, XC, and DC revised the manuscript. RZ, YW, and FC provided the financial support and study supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Key Research and Development Program of China (2016YFE0204900 to FC), National Natural Science Foundation of China (81530088 to FC and 81973142 to YW), Natural Science Foundation of Jiangsu Province (BK20191354 to RZ), Natural Science Foundation of Jiangsu Higher Education Institutions of China (18KJB310011 to RZ), and Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The sponsors had no role in design of the study, collection and analysis of data, or preparation of the manuscript.

ACKNOWLEDGMENTS

We would like to thank SSGAC for providing the educational attainment summary data and the Metabolomics GWAS Server for providing mQTL results.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.618829/full#supplementary-material

REFERENCES

Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173–1182.

Berry, C. M., Gruys, M. L., and Sackett, P. R. (2006). Educational attainment as a proxy for cognitive ability in selection: effects on levels of cognitive ability and adverse impact. *J. Appl. Psychol.* 91, 696–705. doi: 10.1037/0021-9010.91.3.696

Bothwell, L. E., Greene, J. A., Podolsky, S. H., and Jones, D. S. (2016). Assessing the Gold Standard–Lessons from the History of RCTs. *New Engl. J. Med.* 374, 2175–2181. doi: 10.1056/nejmms1604593

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N., Thompson, J., et al. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statist. Med.* 36, 1783–1802.

Bowden, J., Del Greco, M. F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., et al. (2019). Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Intern. J. Epidemiol.* 48, 728–742. doi: 10.1093/ije/dyy258

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37, 658–665. doi: 10.1002/gepi.21758

Burgess, S., Dudbridge, F., and Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statist. Med.* 35, 1880–1906. doi: 10.1002/sim.6835

Burgess, S., and Labrecque, J. A. (2018). Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *Eur. J. Epidemiol.* 33, 947–952. doi: 10.1007/s10654-018-0424-6

Burgess, S., Scott, R. A., Timpson, N. J., Davey Smith, G., Thompson, S. G., EPIC- Interact Consortium, et al. (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* 30, 543–552. doi: 10.1007/s10654-015-0011-z

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* 26, 2333–2355. doi: 10.1177/0962280215597579

Burgess, S., and Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *Intern. J. Epidemiol.* 42, 1134–1144. doi: 10.1093/ije/dyt093

Buse, J. B., Freeman, J. L., Edelman, S. V., Jovanovic, L., and McGill, J. B. (2003). Serum 1,5-anhydroglucitol (GlycoMark): a short-term glycemic marker. *Diabetes Technol. Therap.* 5, 355–363. doi: 10.1089/152091503765691839

Corydon, M. J., Vockley, J., Rinaldo, P., Rhead, W. J., Kjeldsen, M., Winter, V., et al. (2001). Role of common gene variations in the molecular pathogenesis of short-chain acyl-CoA dehydrogenase deficiency. *Pediatr. Res.* 49, 18–23.

Ebrahim, S., and Davey Smith, G. (2008). Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Hum. Genet.* 123, 15–33. doi: 10.1007/s00439-007-0448-6

Esch, P., Bocquet, V., Pull, C., Couffignal, S., Lehnert, T., Graas, M., et al. (2014). The downward spiral of mental disorders and educational attainment: a systematic review on early school leaving. *BMC Psychiatry* 14:237.

Ference, B. A., Ray, K. K., Catapano, A. L., Ference, T. B., Burgess, S., Neff, D. R., et al. (2019). Mendelian randomization study of ACLY and cardiovascular disease. *New Engl. J. Med.* 380, 1033–1042. doi: 10.1056/nejmoa1806747

Jones, L. L., McDonald, D. A., and Borum, P. R. (2010). Acylcarnitines: role in brain. *Progress Lipid Res.* 49, 61–75. doi: 10.1016/j.plipres.2009.08.004

Kölker, S., Koeller, D. M., Okun, J. G., and Hoffmann, G. F. (2004). Pathomechanisms of neurodegeneration in glutaryl-CoA dehydrogenase deficiency. *Ann. Neurol.* 55, 7–12. doi: 10.1002/ana.10784

Martens, E. P., Wiebe, P., de Boer, A., Svetlana, B. V., and Klungel, O. H. (2006). Instrumental variables application and limitations. *Epidemiology* 17, 260–267.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533, 539–542.

Paternoster, L., Tilling, K., and Davey Smith, G. (2017). Genetic epidemiology and Mendelian randomization for informing disease therapeutics: conceptual and methodological challenges. *PLoS Genet.* 13:e1006944. doi: 10.1371/journal.pgen.1006944

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., and Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* 48, 709–717. doi: 10.1038/ng.3570

Pierce, B. L., Ahsan, H., and Vanderweele, T. J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Intern. J. Epidemiol.* 40, 740–752. doi: 10.1093/ije/dyq151

Rawlings, A. M., Sharrett, A. R., Mosley, T. H., Ballew, S. H., Deal, J. A., Selvin, E., et al. (2017). Glucose peaks and the risk of dementia and 20-year cognitive decline. *Diabetes Care* 40, 879–886. doi: 10.2337/dc16-2203

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340, 1467–1471.

Sharma, V., Ichikawa, M., and Freeze, H. H. (2014). Mannose metabolism: more than meets the eye. *Biochem. Biophys. Res. Commun.* 453, 220–228. doi: 10.1016/j.bbrc.2014.06.021

Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550.

Simpson, B. N., Kim, M., Chuang, Y.-F., Beason-Held, L., Triolo, M. K., Kraut, M., et al. (2016). Blood metabolite markers of cognitive performance and brain function in aging. *J. Cereb. Blood Flow Metab.* 36, 1212–1223. doi: 10.1177/0271678x15611678

Slob, E. A. W., and Burgess, S. (2020). A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* 44, 313–329. doi: 10.1002/gepi.22295

Sloviter, H. A., and Kamimoto, T. (1970). The isolated, persed rat brain preparation metabolizes mannose but not maltose. *J. Neurochem.* 17, 1109–1111. doi: 10.1111/j.1471-4159.1970.tb02266.x

Smith, G. D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Intern. J. Epidemiol.* 32, 1–22. doi: 10.1093/ije/dyg070

Solovieff, N., Chris, C., Lee, P. H., and Purcell, S. M. (2013). Smolle Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483–495. doi: 10.1038/nrg3461

Thomas, D. C., Lawlor, D. A., and Thompson, J. R. (2007). Re: estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Ann. Epidemiol.* 17, 511–513. doi: 10.1016/j.annepidem.2006.12.005

van Maldegem, B. T., Duran, M., Wanders, R. J., Niezen-Koning, K. E., Hogeveen, M., Ijlst, L., et al. (2006). Clinical, biochemical, and genetic heterogeneity in short-chain acyl-coenzyme A dehydrogenase deficiency. *JAMA* 296, 943–952. doi: 10.1001/jama.296.8.943

VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., and Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology* 25, 427–435. doi: 10.1097/ede.0000000000000081

Viegas, C. M., Zanatta, A., Knebel, L. A., Schuck, P. F., Tonin, A. M., Ferreira Gda, C., et al. (2009). Experimental evidence that ornithine and homocitrulline disrupt energy metabolism in brain of young rats. *Brain Res.* 1291, 102–112. doi: 10.1016/j.brainres.2009.07.021

Wajner, M., and Amaral, A. U. (2015). Mitochondrial dysfunction in fatty acid oxidation disorders: insights from human and animal studies. *Biosci. Rep.* 36:e00281.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006.

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Intern. J. Epidemiol.* 46, 1734–1739. doi: 10.1093/ije/dyx034

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., et al. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* 9:224.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.