TEMR: Trans-ethnic Mendelian Randomization Method using Large-scale GWAS Summary Datasets

Lei Hou[1†], Sijia Wu[2,3†], Zhongshang Yuan[2,3*], Hongkai Li[2,3*], Fuzhong Xue[2,3,4*]

[†] Lei Hou and Sijia Wu contributed equally to this work.

[*] Fuzhong Xue, Hongkai Li and Zhongshang Yuan contributed equally to this work.

Author affiliations:

1. Beijing International Center for Mathematical Research, Peking University, Beijing, People's Republic of China, 100871

2. Department of Epidemiology and Health Statistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, People's Republic of China, 250000

3. Institute for Medical Dataology, Cheeloo College of Medicine, Shandong University, Jinan, People's Republic of China, 250000

4. Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, People's Republic of China, 250000

Corresponding author:

1. Fuzhong Xue,

E-mail: xuefzh@sdu.edu.cn,

Telephone: 13906405997,

Address: School of public health, Shandong University, 44 Wenhua West Road, Jinan, Shandong province, China

2. Hongkai Li,

E-mail: lihongkaiyouxiang@163.com,

Telephone: 18310601363,

Address: School of public health, Shandong University, 44 Wenhua West Road, Jinan, Shandong province, China

3. Zhongshang Yuan,

E-mail: yuanzhongshang@sdu.edu.cn,

Telephone: 15069095790,

Address: School of public health, Shandong University, 44 Wenhua West Road, Jinan, Shandong province, China

Abstract

Available large-scale GWAS summary datasets predominantly stem from European populations, while sample sizes for other ethnicities, notably Central/South Asian, East Asian, African, Hispanic, etc. remain comparatively limited, which induces the low precision of causal effect estimation within these ethnicities using Mendelian Randomization (MR). In this paper, we propose a Trans-ethnic MR method called TEMR to improve statistical power and estimation precision of MR in the target population using trans-ethnic large-scale GWAS summary datasets. TEMR incorporates trans-ethnic genetic correlation coefficients through a conditional likelihood-based inference framework, producing calibrated p-values with substantially improved MR power. In the simulation study, TEMR exhibited superior precision and statistical power in the causal effects estimation within the target populations than other existing MR methods. Finally, we applied TEMR to infer causal relationships from 17 blood biomarkers to four diseases (hypertension, ischemic stroke, type 2 diabetes and schizophrenia) in East Asian, African and Hispanic/Latino populations leveraging the biobank-scale GWAS summary data from European. We found that causal biomarkers were mostly validated by previous MR methods, and we also discovered 13 new causal relationships that were not identified using previously published MR methods.

Keywords: trans-ethnic mendelian randomization, genetic correlation, GWAS summary data, statistical power

Introduction

In recent years, the evolving landscape has witnessed a progressive expansion of large-scale Genome-Wide Association Studies (GWAS), leading to the widespread release and utilization of GWAS summary data among researchers. At the forefront of these developments is Mendelian Randomization (MR) [1-2], a method that hinges on the use of publicly available GWAS summary data for causal inference. MR uses genetic variants as instrumental variables (IVs) to infer causal effect of an exposure on an outcome. It requires three assumptions: Relevance, IVs are strongly associated with the exposure; Exchangeability, IVs are independent with confounders among the exposure and outcome; Exclusion restriction, IVs affect the outcome only through the exposure. However, a noteworthy challenge surfaces as the bulk of available large-scale datasets predominantly stem from European populations, such as the UK Biobank (UKB) [3-7] and FinnGen consortium [8], while sample sizes for other ethnicities, notably Central/South Asian, East Asian, African, Hispanic, etc. remain comparatively limited [9-12]. Take the East Asian population as an example, despite the substantial data provided by the BioBank Japan (BBJ) [11-12], Taiwan Biobank (TWB) [13] and China Kadoorie Biobank (CKB) [14] for the East Asian population (> 100,000 individuals), it falls short of the extensive dataset available from the UKB (> 500,000 individuals) or FinnGen consortium (> 620,000 individuals). Moreover, the UKB incorporates a substantial amount of omics data, including imaging omics [4], exomes [5], proteomics [6] and metabolomics [7]—BBJ, TWB and CKB have significantly smaller sample sizes and may also lack some omics data [12]. Furthermore, omics databases dedicated to other ethnicities tend to exhibit relatively smaller sample sizes [15-21]. The potential inadequacy of GWAS summary data from smaller samples to furnish robust causal evidence for MR becomes apparent. Additionally, causal evidence derived from a substantial European population cannot be directly extrapolated to other ethnic groups due to diversity in the genetic structure between different ethnicities [22-23]. The unbalanced sample makeup across global populations may exacerbate the disparities in genetic studies of non-Europeans. Therefore, it is crucial to propose a methodology that leverages the genetic correlations [24-25] among different ethnicities, harnessing the advantages of large European datasets to enhance the accuracy and statistical power of MR in estimating causal effects within smaller populations.

A number of trans-ethnic MR analyses has been published, predominantly featured in applied research articles [26-28]. The common approach in these studies involves conducting separate MR analyses within distinct ethnic groups and subsequently comparing the nuances in the MR results between these groups. This is unfair for ethnicities with small sample sizes, as its statistical power of MR is much lower than that of large sample sizes. For methodology, advancements have been

made in cross-ethnic approaches within GWAS meta-analysis and Polygenic Risk Score (PRS). The published trans-ethnic meta-analysis approaches take into account the similarity in allelic effects between the most closely related populations while allowing for heterogeneity between more diverse ethnic group [29-32]. While trans-ethnic GWAS meta-analysis has the potential to improve the efficiency of identifying new loci by merging populations of different ethnicities, it operates at a mixed-population level and may not necessarily contribute to the discovery of genetic loci specific to particular ethnic groups. Tans-ethnic PRS prediction methods leverage shared genetic effects across ancestries to increase the accuracy of predicting the genetic predisposition of complex phenotypes in non-European populations [33-35]. However, these methods highlight that improving the power to discovery new loci or disease prediction, a noticeable gap in the current literature lies in the lack of attention to methods facilitating the transfer of causal effects in MR across different ethnicities. Despite progress in various methodological aspects of trans-ethnic analysis, there remains an unexplored avenue concerning the migration of causal effects across ethnic groups in the context of MR.

In this paper, we propose a MR method based on Trans-ethnic Population called TEMR to improve statistical power and estimation precision of MR in target population using trans-ethnic large-scale GWAS summary datasets. Under the framework of conditional likelihood-based inference framework, TEMR bridges the causal effects of different ethnics using a trans-ethnic genetic correlation coefficient, which is the correlation of Wald ratios for shared SNPs in different ethnic populations. In the simulation study, TEMR showed superior precision and power of causal effect estimation in the target population relative to other seven methods in the case of continuous and binary outcome variables. Finally, we apply TEMR to infer causal relationships from 17 blood biomarkers to four diseases (hypertension, ischemic stroke, type 2 diabetes (T2D) and schizophrenia) in East Asian, African and Hispanic/Latino populations leveraging the biobank-scale GWAS summary data from European.

Results

TEMR Method overview

[please insert Figure 1 here]

We consider a target dataset $\{G_1, X_1, Y_1\}$ from an under-represented ancestry (e.g. East Asian, African and Hispanic/Latino population, etc) with small sample size, where G_1, X_1 and Y_1 represent the Single Nucleotide Polymorphisms (SNPs), exposure and outcome, respectively. Now

suppose we have an auxiliary dataset $\{G_2, X_2, Y_2\}$ (e.g. European population) with a biobank-scale sample size available. We assume the sample sizes of two datasets satisfy the condition $N_2 \gg N_1$. We choose the *p* independent SNPs as IVs, which are associated with at least one of exposures in two ancestries (X_1 and X_2). The workflow of TEMR is shown in Figure 1.

When the three core assumptions of MR are all satisfied, we can obtain the Wald ratio estimation for each SNP: $\hat{\beta}_{1j}$ in the target population and $\hat{\beta}_{2j}$ in the target population, as well as their variances $\hat{\sigma}^2_{\beta_{1j}}$ and $\hat{\sigma}^2_{\beta_{2j}}$, respectively, using the summary-level data of *p* SNPs, including beta-coefficients ($\hat{\beta}_{Y_{1j}}, \hat{\beta}_{X_{1j}}$ and $\hat{\beta}_{Y_{2j}}, \hat{\beta}_{X_{2j}}$) and their standard error ($\hat{\sigma}^2_{Y_{1j}}, \hat{\sigma}^2_{X_{1j}}$ and $\hat{\sigma}^2_{Y_{2j}}, \hat{\sigma}^2_{X_{2j}}$). We set up the following multivariable normal distribution model for Wald ratios from two populations:

$$\begin{pmatrix} \hat{\beta}_{1j} \\ \hat{\beta}_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\beta_{1j}} & \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{2j}} \\ \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{2j}} & \hat{\sigma}^2_{\beta_{2j}} \end{pmatrix} \right) \tag{1}$$

where β_1 and β_2 are the causal effect of exposure on outcome in the target and auxiliary populations, respectively. ρ_β is the trans-ethnic genetic correlation, which represents the correlation of the causal effects of one exposure on one outcome in two ancestries (e.g. East Asian and European). It bridges the causal effects of two ethnics to achieve our aim of improving the statistical power of causal effect (β_1) estimation in the target population. In terms of the conditional normal distribution of $\hat{\beta}_{1j}$ given $\hat{\beta}_{2j}$,

$$\hat{\beta}_{1j} \mid \hat{\beta}_{2j} \sim N(\beta_1 + \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}^{-1}_{\beta_{2j}} (\hat{\beta}_{2j} - \beta_2), \hat{\sigma}^2_{\beta_{1j}} - \rho_\beta{}^2 \hat{\sigma}_{\beta_{1j}}{}^2) \tag{2}$$

we found that the variance of *j*-th Wald ratio estimation $\hat{\beta}_{1j}$ conditional on $\hat{\beta}_{2j}$ is smaller than its original variance as the trans-ethnic genetic correlation ρ_β increasing ($var(\hat{\beta}_{1j} \mid \hat{\beta}_{2j}) = \hat{\sigma}^2_{\beta_{1j}}(1 - \rho_\beta{}^2) < \hat{\sigma}^2_{\beta_{1j}}$). Then we obtain the causal effect (β_1) estimation in the target population by maximizing the log conditional likelihood function using Nelder-Mead method [38]. We use Likelihood-ratio test to perform hypothesis testing.

When there is horizontal pleiotropy, the third assumption of MR is violated, the causal effect estimation using the traditional Wald ratio is biased and we model a new TEMR-Wald ratio by removing the impact of horizontal pleiotropy (α_{1j}) from $\hat{\beta}_{Y_{1j}}$. We propose a two-step process to estimate TEMR-Wald ratios for each SNP leveraging MR-Egger regression. Next, we use the new TEMR-Wald ratio to set up model (1), then infer the causal effect (β_1) in the target population, and

the remaining steps are the same as in the case of no pleiotropy.

If there are multiple ancestries (*E* ancestries), the target dataset is $\{G_T, X_T, Y_T\}$ and the auxiliary datasets are $\{G_a, X_a, Y_a\}(a = 2,...,E)$, we can also set up the multivariable normal distribution model using Wald ratios from *E* ancestries

$$\begin{pmatrix} \hat{\beta}_{Tj} \\ \hat{\beta}_{Aj} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_T \\ \beta_A \end{pmatrix}, \begin{pmatrix} \sigma^2_{\beta_{Tj}} & \Sigma_{A1} \\ \Sigma_{1A} & \Sigma_{AA} \end{pmatrix} \right) \tag{3}$$

where $\hat{\beta}_{Aj} = \begin{pmatrix} \hat{\beta}_{2j} \\ \vdots \\ \hat{\beta}_{Ej} \end{pmatrix}, \beta_A = \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_E \end{pmatrix}, \Sigma_{AA} = \begin{pmatrix} \sigma^2_{\beta_{2j}} & \cdots & \rho_{\beta_{(2,E)}}\sigma_{\beta_{2j}}\sigma_{\beta_{Ej}} \\ \vdots & \cdots & \vdots \\ \rho_{\beta_{(E2)}}\sigma_{\beta_{Ej}}\sigma_{\beta_{2j}} & \cdots & \sigma^2_{\beta_{Ej}} \end{pmatrix},$

$\Sigma_{A1} = \begin{pmatrix} \rho_{\beta_{(1,2)}}\sigma_{\beta_{1j}}\sigma_{\beta_{2j}} & \cdots & \rho_{\beta_{(1,E)}}\sigma_{\beta_{1j}}\sigma_{\beta_{Ej}} \end{pmatrix}, \Sigma_{1A} = \Sigma_{A1}^{\ T}$ and $\rho_{\beta_{(m,n)}} = \rho_{\beta_{(n,m)}}$. Then we derive the conditional model and obtain the estimation of β_T by Maximum Likelihood Estimation.

Simulation

We conducted a series of simulation studies to evaluate the performance of TEMR, comprising with seven published MR methods. We vary with magnitudes of parameters: causal effect, trans-ethnic genetic correlation, sample size and the number of SNPs, in the scenarios of no pleiotropy and horizontal pleiotropy. We utilized boxplots to demonstrate the results of estimation bias and standard error, Q-Q plots to showcase the results of Type I error, and bar charts to depict the results of statistical power.

When there is no pleiotropy, Figure 2 shows the simulation results of causal effect estimation in the target population when there is one auxiliary population in the case of continuous outcome. Simulation results demonstrated that TEMR showed nearly unbiased estimates of causal effects regardless of the alignment between causal effects in the auxiliary and target populations. TEMR also showed superior precision and power across a broad spectrum of scenarios relative to other seven methods, which was consistently observed for both continuous and binary outcome variables. The precision of TEMR incrementally improved as the ρ_β increasing. When $\rho_\beta < 0.4$, the precision of TEMR was similar with the Inverse-variance weighted method (IVW) and the Weighted Median Estimation (WME) method. However, when $\rho_\beta \geq 0.4$, the precision of TEMR surpassed that of other seven methods (Figure 2A, Figure S1). Additionally, TEMR exhibited stable Type I errors, unaffected by variations in ρ_β or causal effects in the auxiliary population (Figure 2B-C, Figure S2). Moreover, the statistical power of TEMR significantly increased with ρ_β rising,

especially outperforming other seven methods when $\rho_\beta \geq 0.4$ (Figure 2D-E, Figure S3-S6). Specifically, when $\rho_\beta = 0.6$, there was a notable decrease in the standard error by approximately 15-20%, and an increase in power by about 20%. At a higher ρ_β, the standard error could decrease by up to 50%, while the power could improve by 40% (Table S1).

[please insert Figure 2 here]

Then we extended our simulation to scenario where there is horizontal pleiotropy, both balance and directional, are present. In addition to achieving unbiased estimates of causal effects and stable Type I errors, TEMR also maintained the precision and power advantages as described above (Figure 3, Figure S7-18). In cases involving categorical outcome, we obtained results consistent with those for continuous variables (Figure S19-36, Table S2). Additionally, we also observed that the target population can also enhance the precision and test power of causal effect estimates in the auxiliary population, exemplified by scenarios directional horizontal pleiotropy (Figure S37, Table S3). These observations underscore the robustness and effectiveness of TEMR in various genetic correlation contexts.

[please insert Figure 3 here]

In order to investigate the impact of the number of SNPs (Figure 4, Table S4) and sample size (Figure S38-43, Table S5) on the causal effect estimates, firstly, we conducted a thorough exploration by varying the number of SNPs while maintaining other parameters at their initial settings. Simulation results indicated that an increase in the number of SNPs leads to higher precision and greater test power in the causal effect estimates derived from the TEMR. While other methods also demonstrated improvements with more SNPs, the enhancement was not as pronounced as that observed with TEMR.

[please insert Figure 4 here]

Additionally, we explored the causal effect estimates using the TEMR when there is a negative genetic correlation between ethnic groups. The results indicated that the precision and the statistical power significantly improved as the absolute value of the genetic correlation increased, which was consistent with the aforementioned findings (Figure S44-45, Table S6).

Subsequently, we consider the case of multiple auxiliary populations, taking three auxiliary populations and one target population as an example, assuming uniform causal effects across different populations. In the absence of horizontal pleiotropy, TEMR produced nearly unbiased estimates of causal effects. And the precision and statistical power of TEMR also incrementally improved as the ρ_β increased, when $\rho_\beta \geq 0.4$, the precision and power of TEMR surpassed that

of other methods (Figure 5A). Specifically, when $\rho_\beta = 0.6$, there was a notable decrease in the standard error by approximately 15-20%, and an increase in power by about 20%. At a higher ρ_β, the standard error could decrease by up to 50%, while the power could improve by 40% (Figure 5C). Furthermore, TEMR exhibited stable Type I errors, unaffected by variations in ρ_β (Figure 5B). In cases involving horizontal pleiotropy, we obtained results consistent with those (Figure S46-54, Table S7). And we also obtained consistent results when the genetic correlations were negative (Figure S55-56, Table S8). Moreover, compared to having only one auxiliary population, the precision (standard error) of the causal effect estimate obtained from three auxiliary populations also improved with the increase in genetic correlations, with an approximate 15% improvement when $\rho_\beta = 0.9$.

[please insert Figure 5 here]

Application

In this section, we applied TEMR to infer the causal relationships between different biomarkers and four diseases (hypertension, ischemic stroke, T2D and schizophrenia) in the East Asian, African, Hispanic/Latino population, leveraging GWAS summary data from large European cohorts (Table S9). Initially, we identified 17 specific biomarkers that were significantly associated with at least 2 SNPs from a multitude of biomarkers. Then we calculated the trans-ethnic genetic correlation for all pairs of biomarkers, results are shown in Figure 6 (Table S10). The results showed that there were trans-ethnic genetic correlations between the causal effects of all biomarkers and diseases in four populations, which could be analyzed by TEMR. Among these, the absolute value of correlation of 14 pairs exhibited 0.5 (total 17×4=68 pairs), including basophil count to hypertension (between East Asian and Hispanic/Latino), neutrophil count to hypertension (between African, Hispanic/Latino and East Asian), mean corpuscular hemoglobin concentration (MCHC) to ischemic stroke (between East Asian, European and Hispanic/Latino, African and East Asian), eosinophil count to schizophrenia (between European and Hispanic/Latino), neutrophil count to schizophrenia (except between East Asian and Hispanic/Latino), platelet count to schizophrenia (between East Asian, Hispanic/Latino and African), body mass index (BMI) to schizophrenia (between European and East Asian), triglyceride (TG) to schizophrenia (between European and East Asian), lymphocyte count to T2D (between European, Hispanic/Latino and East Asian), BMI to T2D (between European and Hispanic/Latino), glucose to T2D (between European Hispanic/Latino and East Asian), neutrophil count to T2D (between European, African and East Asian), Total cholesterol

(TC) to T2D (between European and East Asian), TG to T2D (between European and East Asian).

[please insert Figure 6 here]

Then we perform trans-ethnic MR analysis using TEMR and other seven methods. The results indicated that TEMR identified a greater number of biomarker pairs with significant causal associations compared to the other seven methods (Figure 7). Among these, TEMR emerged as the method identifying the most significant biomarker pairs in each target population, with IVW following closely behind (Table S11), and most of the significant biomarker pairs identified by the other methods were also detected by TEMR with smaller *P*-values. Notably, there were several significant relationships across different ethnic groups that only TEMR identified as significant ($P<0.0007(0.05/68)$): three new causal relationships in East Asian, four new causal relationships in African and six in Hispanic/Latino population.

[please insert Figure 7 here]

In the East Asian population, significant causal associations including TC to schizophrenia (*OR*=2.30, *P*=0.019), HDL-cholesterol (HDL-C) to schizophrenia (*OR*=0.50, *P*=0.044) and neutrophil count to T2D (*OR*=0.89, *P*=0.038) were detected (Table S12). The association between TC and schizophrenia suggested that higher TC levels might influence the risk of developing schizophrenia. This connection could be through the alteration of cell membrane fluidity, which in turn may impact neurotransmitter signaling. Many studies supported that elevated serum TC levels could be linked to enhanced cognitive function in individuals with schizophrenia [39-41]. For the relationship between HDL-C and schizophrenia, the inverse association could indicate that higher levels of HDL-C, often considered good cholesterol, might have a protective effect against the development of schizophrenia. Studies corroborated these findings, indicating that patients with schizophrenia often have lower levels of HDL-C compared to those without the condition [40]. Finally, while the relationship between neutrophil count and the risk of T2D is complex and not fully understood, some studies have suggested that increased neutrophil activity may be associated with a reduced risk of the disease, potentially due to their role in modulating inflammatory responses [42-43].

Similarly, in the African population, the TEMR analysis revealed notable associations such as lymphocyte count to schizophrenia (*OR*=0.01, $P<0.001$) which might suggest a substantial protective role of higher lymphocyte counts against schizophrenia, potentially through immune regulation mechanisms [44]. The link between glucose levels and schizophrenia (*OR*=0.84, $P<0.001$) could reflect the metabolic disturbances that are often observed in patients with schizophrenia and might indicate a broader metabolic syndrome component of the disorder [45]. The causal relationship between TC and schizophrenia (*OR*=0.78, *P*=0.007) in this demographic implies a protective effect

of lower cholesterol levels, which contrasts with findings in the Asian population, suggesting the influence of genetic and environmental factors in different populations. Cholesterol is involved in the production of steroid hormones and neurosteroids, which neurosteroids have been found to modulate the central nervous system's activity and may influence symptoms of schizophrenia. They can also affect the immune system, which has been implicated in the pathophysiology of schizophrenia [46-47]. Higher TC levels could reflect more robust synthesis of such compounds, potentially contributing to more stable cellular functions and better disease outcomes [48]. Additionally, the significant association between TC and hypertension (*OR*=0.74, *P*<0.001) could hint at the complex interplay between lipid metabolism and blood pressure regulation. There have been suggestions that certain lipid components might have a role in immune defense systems and that some aspects of the inflammatory response may be influenced by lipid levels [49]. Cholesterol may also modulate cell membrane properties, affecting the reactivity of blood vessels and contributing to the regulation of blood pressure (BP) [50] (Table S13).

In the Hispanic/Latino population, the TEMR method unveiled notable significant causal relationships, indicative of unique pathophysiological pathways. The association of TG to schizophrenia (*OR*=0.80, *P*<0.001) and TC to schizophrenia (*OR*=0.61, *P*<0.001) also suggested a link between metabolic dysregulation and the development of schizophrenia. The significant causal associations between basophil count (*OR*<0.01, *P*<0.001) and platelet count (*OR*=0.06, *P*<0.001) emphasized the role of the immune system in schizophrenia [51]. The association of LDL-cholesterol (LDL-C) with hypertension (*OR*=0.97, *P*<0.001) parallelled the findings of TC in the African population, adding to the evidence that lipid metabolism plays a complex role in BP regulation across different ethnicities. Finally, the associations between TG (*OR*=0.96, *P*<0.001) and ischemic stroke suggested a pathogenic role for blood components and lipid metabolism in vascular health. For TG, while high levels are typically considered a risk factor for atherosclerosis and thus ischemic strokes, However, very low TG levels might also not be ideal, as they can indicate an insufficient energy reserve for normal cellular functions, which could potentially affect overall health and cellular processes, including those in vascular cells [52] (Table S14).

Discussion

In this paper, we propose a trans-ethnic MR method called TEMR to improve statistical power and estimation accuracy of MR in the target population only using trans-ethnic large-scale GWAS summary datasets. TEMR showed superior precision and power of causal effect estimation in the target population relative to other published MR methods in the simulation study. Leveraging the

biobank-scale GWAS summary data from European, application of inferring causal relationships from 17 blood biomarkers to four diseases in East Asian, African population and Hispanic populations discover 13 new causal relationships that not found using published MR methods.

TEMR bridges the causal effects of multiple ethnics using a trans-ethnic genetic correlation coefficient. With the increase of trans-ethnic genetic association, the statistical power of causal effect in the non-European population is significantly improved. Trans-ethnic genetic correlation measures the extent to which genetic variants influence phenotypes similarly across different populations. With the advent of genomic technologies, researchers were able to conduct genome-wide studies of large cohorts from different ethnicities. These studies revealed that while there is substantial genetic variation between different populations, certain variants have similar frequencies and effects across groups. Numerous studies showed that the genetic variants for many traits were highly correlated across different populations. Trans-ethnic genetic correlation is assessed using various methods, such as multi-ancestry GWAS, TWAS and PRS prediction, etc. It can be estimated by abundant methods including LD score regression [53], HDL [54], GCTA-GREML [55], BOLT-REML [56] and PAINTOR [57], etc. They can achieve much higher accuracy than z-score based method. In this paper, TEMR uses a simple Z-score method to get results quickly, and using these methods will make TEMR perform better. TEMR is suitable for traits with high genetic association between different ethnics. When the genetic association between traits is nearly zero, TEMR method behaves similar to traditional MR Method.

There are several limitations in our study. The impact of pleiotropy is an important topic in MR study. Here we consider the case of no pleiotropy and horizontal pleiotropy. For the latter, we propose a two-step process to remove the pleiotropy effect from traditional Wald ratio using MR-Egger regression, and obtain the TEMR-Wald ratio estimation. The limitation of this process is that it also requires the InSIDE assumption and cannot remove the influence of correlated pleiotropy. Available solution is that detect outliers using published methods such as MR Radial and MR-PRESSO, etc, and then remove them before conducting TEMR. In addition, when there are multiple ethnics, TEMR can improve the statistical power of causal effect estimation only in one target population leveraging other target populations and European population. In the future, we will extend TEMR to implement that improving the statistical power of causal effect estimation in multiple target population leveraging only European population. The degree of improvement in statistical power is closely related to the number of IVs and the magnitude of trans-ethnic genetic correlations.

In conclusion, we proposed a new method TEMR to improve statistical power and estimation accuracy of MR in the target population only using trans-ethnic large-scale GWAS summary dataset.

It has important guiding significance for the discovery of new disease-related factors.

Methods

TEMR model based on two ancestries

Consider a target dataset $\{G_1, X_1, Y_1\}$ from an under-represented ancestry (e.g. East Asian ancestry) with small sample size, where G_1 is an $N_1 \times p$ genotype matrix, X_1 and Y_1 are $N_1 \times 1$ phenotype/disease vectors, represent the exposure and outcome, respectively. Now we suppose a biobank-scale dataset $\{G_2, X_2, Y_2\}$ (e.g. European ancestry) is also available, where G_2 is an $N_2 \times p$ genotype matrix, X_2 and Y_2 are $N_2 \times 1$ phenotype/disease vectors, represent exposure and outcome, respectively. We assume $N_2 \square N_1$. Since we are mainly interested in improving the statistical power of causal effect estimation in the target population leveraging biobank scale datasets from another ancestry. We choose the *p* independent IVs (SNPs) which are associated with at least one of exposures in two ancestries (X_1 and X_2). We can obtain the summary-level data of *p* SNPs from published GWAS studies, including the beta-coefficients ($\hat{\beta}_{Y_{1j}}, \hat{\beta}_{X_{1j}}$ and $\hat{\beta}_{Y_{2j}}, \hat{\beta}_{X_{2j}}$) and their standard error ($\hat{\sigma}^2_{Y_{1j}}, \hat{\sigma}^2_{X_{1j}}$ and $\hat{\sigma}^2_{Y_{2j}}, \hat{\sigma}^2_{X_{2j}}$).

When the three core assumptions of MR are all satisfied, we can obtain the causal effect estimation using the Wald ratio for each SNP

$$\hat{\beta}_{1j} = \frac{\hat{\beta}_{Y_{1j}}}{\hat{\beta}_{X_{1j}}}, \hat{\beta}_{2j} = \frac{\hat{\beta}_{Y_{2j}}}{\hat{\beta}_{X_{2j}}}, \; j = 1, \ldots, p, \tag{4}$$

with their variances

$$\hat{\sigma}^2_{\beta_{1j}} = \frac{\hat{\beta}^2_{Y_{1j}} \times \hat{\sigma}^2_{X_{1j}}}{\hat{\sigma}^4_{X_{1j}}} + \frac{\hat{\sigma}^2_{Y_{1j}}}{\hat{\sigma}^2_{X_{1j}}}, \quad \hat{\sigma}^2_{\beta_{2j}} = \frac{\hat{\beta}^2_{Y_{2j}} \times \hat{\sigma}^2_{X_{2j}}}{\hat{\sigma}^4_{X_{2j}}} + \frac{\hat{\sigma}^2_{Y_{2j}}}{\hat{\sigma}^2_{X_{2j}}}.$$

The $\hat{\beta}_{1j}$ and $\hat{\beta}_{2j}$ are the causal effect estimation of exposure on outcome using j -th SNPs in the target and auxiliary populations, respectively. We set up the following multivariable normal distribution model for Wald ratios from two populations

$$\begin{pmatrix} \hat{\beta}_{1j} \\ \hat{\beta}_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\beta_{1j}} & \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{2j}} \\ \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{2j}} & \hat{\sigma}^2_{\beta_{2j}} \end{pmatrix} \right),$$

where β_1 and β_2 are the causal effect of exposure on outcome in the target and auxiliary

populations, respectively. They can be the same or different. ρ_β is the trans-ethnic genetic correlation, which represents the correlation of the causal effects of one exposure on one outcome in two ancestries (e.g. Chinese and European), and it can be calculated by the Pearson correlation

$$\rho_\beta = \frac{cov(z_1, z_2)}{\sqrt{var(z_1) \cdot var(z_2)}},$$

where z_1 and z_2 are the *z*-scores of *p*-dimensional Wald ratio vectors in two ancestries, respectively

$$z_{1j} = \frac{\hat{\beta}_{1j}}{\hat{\sigma}_{\beta_{1j}}}, z_{2j} = \frac{\hat{\beta}_{2j}}{\hat{\sigma}_{\beta_{2j}}}, j = 1, ..., p.$$

We aim to improve the statistical power of causal effect (β_1) estimation in the target population using the trans-ethnic genetic correlation ρ_β, which connect the causal effects of two ethnics. Based on the model (1) and the conditional normal distribution formula [58], we have

$$\hat{\beta}_{1j} \mid \hat{\beta}_{2j} \sim N(\beta_1 + \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{2j}}^{-1} (\hat{\beta}_{2j} - \beta_2), \hat{\sigma}_{\beta_{1j}}^2 - \rho_\beta{}^2 \hat{\sigma}_{\beta_{1j}}{}^2)$$

with its variance

$$var(\hat{\beta}_{1j} \mid \hat{\beta}_{2j}) = \hat{\sigma}_{\beta_{1j}}^2 (1 - \rho_\beta{}^2) < \hat{\sigma}_{\beta_{1j}}^2$$

Therefore, the variance of *j*-th Wald ratio estimation $\hat{\beta}_{1j}$ conditional on $\hat{\beta}_{2j}$ is smaller than its original variance as the trans-ethnic genetic correlation ρ_β increasing. Then we obtain the conditional log-likelihood function of model (2)

$$Q(\beta_1) = \sum\nolimits_j -p \ln(2\pi) - \frac{1}{2} \ln(\hat{\sigma}_{\beta_{1j}}^2 - \rho_\beta{}^2 \hat{\sigma}_{\beta_{1j}}{}^2) - \frac{1}{2} \frac{(\hat{\beta}_{1j} - \beta_1 - \rho_\beta \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{2j}}^{-1} (\hat{\beta}_{2j} - \hat{\beta}_2))^2}{(1 - \rho_\beta{}^2) \hat{\sigma}_{\beta_{1j}}{}^2}. \quad (5)$$

where $\hat{\beta}_2$ is obtained by IVW or other effective MR methods using large-scale dataset in the auxiliary population. We aim to maximize the log conditional likelihood function using Nelder-Mead method [38] to obtain the estimation of β_1. Then we use Likelihood-ratio test to perform hypothesis testing,

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

the testing statistics is

$$\chi^2 = -2 \times \frac{Q(\hat{\beta}_1)}{Q(0)} \sim \chi^2(1)$$

When there is horizontal pleiotropy, the third assumption of MR is violated, the causal effect

estimation using the traditional Wald ratio is biased and we model the TEMR-Wald ratio as following

$$\beta_{1j}=\frac{\hat{\beta}_{Y_{1j}}-\alpha_{1j}}{\hat{\beta}_{X_{1j}}},\ \beta_{2j}=\frac{\hat{\beta}_{Y_{2j}}-\alpha_{2j}}{\hat{\beta}_{X_{2j}}} \tag{6}$$

where α_{1j} represent the horizontal pleiotropy and it is unknown. Therefore, in the first step, we need to estimate α_{1j} and α_{2j} using MR-Egger regression

(1) seperately estimate causal effect β_1^{Egger} and β_2^{Egger} in each ancestry using MR-Egger regression

$$\hat{\beta}_{Y_{1j}}=\hat{\beta}_{X_{1j}}\cdot\beta_1^{Egger}+\alpha_1+\varepsilon_{1j},\varepsilon_{1j}\sim N(0,\hat{\sigma}^2_{Y_{1j}})$$
$$\hat{\beta}_{Y_{2j}}=\hat{\beta}_{X_{2j}}\cdot\beta_2^{Egger}+\alpha_2+\varepsilon_{2j},\varepsilon_{2j}\sim N(0,\hat{\sigma}^2_{Y_{1j}})$$

(2) seperately estimate horizontal pleiotropy α_{1j} and α_{2j} in each ancestry using

$$\hat{\alpha}_{1j}=\hat{\beta}_{Y_{1j}}-\hat{\beta}_{X_{1j}}\cdot\hat{\beta}_1^{Egger}$$
$$\hat{\alpha}_{2j}=\hat{\beta}_{Y_{2j}}-\hat{\beta}_{X_{2j}}\cdot\hat{\beta}_2^{Egger}$$

Then we can obtain the estimations of new Wald ratio $\hat{\beta}_{1j}$ and $\hat{\beta}_{2j}$ by substituting $\hat{\alpha}_{1j}$ and $\hat{\alpha}_{2j}$ into equation (6). Following we use models (1,2,4,5) to obtain the estimation of β_1. The difference is that the $\hat{\beta}_2$ in model (5) is obtained by horizontal pleiotropy-robust MR methods using large-scale dataset in the auxiliary population.

TEMR model based on multiple ancestries

If there are $E>2$ ancestries, the target dataset is $\{G_T,X_T,Y_T\}$ and the auxiliary datasets are $\{G_a,X_a,Y_a\}(a=2,...,E)$, we set up the following multivariable normal distribution model using Wald ratios from *E* ancestries

$$\begin{pmatrix}\hat{\beta}_{Tj}\\ \hat{\beta}_{Aj}\end{pmatrix}\sim N\left(\begin{pmatrix}\beta_T\\ \beta_A\end{pmatrix},\begin{pmatrix}\hat{\sigma}^2_{\beta_{Tj}} & \Sigma_{A1}\\ \Sigma_{1A} & \Sigma_{AA}\end{pmatrix}\right)$$

where $\hat{\beta}_{Aj}=\begin{pmatrix}\hat{\beta}_{2j}\\ \vdots\\ \hat{\beta}_{Ej}\end{pmatrix},\beta_A=\begin{pmatrix}\beta_2\\ \vdots\\ \beta_E\end{pmatrix},\Sigma_{AA}=\begin{pmatrix}\hat{\sigma}^2_{\beta_{2j}} & \cdots & \rho_{\beta_{(2,E)}}\hat{\sigma}_{\beta_{2j}}\hat{\sigma}_{\beta_{Ej}}\\ \vdots & \cdots & \vdots\\ \rho_{\beta_{(E2)}}\hat{\sigma}_{\beta_{Ej}}\hat{\sigma}_{\beta_{2j}} & \cdots & \hat{\sigma}^2_{\beta_{Ej}}\end{pmatrix}$,

$\Sigma_{A1}=\begin{pmatrix}\rho_{\beta_{(1,2)}}\hat{\sigma}_{\beta_{1j}}\hat{\sigma}_{\beta_{2j}} & \cdots & \rho_{\beta_{(1,E)}}\hat{\sigma}_{\beta_{1j}}\hat{\sigma}_{\beta_{Ej}}\end{pmatrix},\Sigma_{1A}=\Sigma_{A1}^T$ and $\rho_{\beta_{(m,n)}}=\rho_{\beta_{(n,m)}}$. The conditional distribution

of $\hat{\beta}_{Tj}$ given $\hat{\beta}_{Aj}$ is

$$\hat{\beta}_{Tj} \mid \hat{\beta}_{Aj} \sim N(\beta_T + \Sigma_{A1}\Sigma_{AA}^{-1}(\hat{\beta}_{Aj} - \beta_A), \hat{\sigma}_{\beta_{Tj}}^{-2} - \Sigma_{A1}\Sigma_{AA}^{-1}\Sigma_{1A})$$

Then we obtain the estimation of β_T by Maximum Likelihood Estimation using Nelder-Mead method.

Due to the predominant representation of European individuals in public GWAS summary dataset, with smaller sample sizes for other ethnicities, our aim is to utilize the information from the European population to improve the causal effect estimation precision and testing efficacy for smaller sample populations. Furthermore, if our focus is exclusively on the Asian population, the inclusion of other small-sample ethnicities could still contribute to enhancing the estimation performance of causal effects on the Asian population, although the contribution may not be as substantial as that from the European population.

Simulation settings

In our simulation study, we systematically evaluated the performance of TEMR through several key steps. Initially, we generated individual data for exposure (X_e), outcome (Y_e) and genotypes (G_{je}, $j=1,...,p$) in multiple ethnicities ($e=1,...,E$):

$$\begin{aligned} &G_{je} \sim B(n_e, 2, 0.3), \\ &X_e = \sum\nolimits_j \alpha_{je} G_{je} + \varepsilon_e, \quad \varepsilon_e \sim N(n_e, 0, 1) \\ &Y_e = \beta_e X_e + \sum\nolimits_j \gamma_{je} G_{je} + \xi_e, \quad \xi_e \sim N(n_e, 0, 1) \end{aligned}$$

Subsequently, we obtain GWAS summary data (including the regression coefficients ($\hat{\beta}_{X_{ej}}$ and $\hat{\beta}_{Y_{ej}}$) and their standard errors ($\hat{\sigma}^2_{X_{ej}}$ and $\hat{\sigma}^2_{Y_{ej}}$)) by linear regressions of continuous variables on each SNP and logistic regressions of binary variables on each SNP, enabling the calculation of the Wald ratios' standard errors for each SNP:

$$\hat{\sigma}^2_{\beta_{ej}} = \frac{(\hat{\beta}_{Y_{ej}} / \hat{\beta}_{X_{ej}})^2 \times \hat{\sigma}^2_{X_{ej}}}{\hat{\sigma}^4_{X_{ej}}} + \frac{\hat{\sigma}^2_{Y_{ej}}}{\hat{\sigma}^2_{X_{ej}}}.$$

The reason for initially generating individual-level data was to simulate the variation in the estimates and precision of the Wald ratio obtained with different sample sizes in real-world applications. While it is possible to directly simulate based on the observed Wald ratio from practical data, the limitation lies in the finite range of sample sizes in public GWAS summary datasets. This constraint prevents a comprehensive simulation of the performance of TEMR under

various sample size scenarios. Next, we generate the Wald ratios ($\hat{\beta}_{ej}$) for different ethnicities using trans-ethnic genetic correlation $\rho_{\beta_{(e_1 e_2)}}$:

$$\begin{pmatrix} \hat{\beta}_{1j} \\ \dots \\ \hat{\beta}_{Ej} \end{pmatrix} \sim Mvrnorm\left(\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_E \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\beta_{1j}} & \dots & \rho_{\beta_{(1,E)}} \hat{\sigma}_{\beta_{1j}} \hat{\sigma}_{\beta_{Ej}} \\ \vdots & \dots & \vdots \\ \rho_{\beta_{(E1)}} \hat{\sigma}_{\beta_{Ej}} \hat{\sigma}_{\beta_{1j}} & \dots & \hat{\sigma}^2_{\beta_{Ej}} \end{pmatrix} \right)$$

We considered scenarios where the causal effects are the same or different across different ethnicities, as well as situations where the causal effects are either zero ($\beta_e = 0$) or non-zero ($\beta_e = 0.05, 0.1, 0.15, 0.2$). We also explore various scenarios, including different trans-ethnic genetic correlation coefficients between ethnicities. We considered the number of ethnicities is *E*=2 or *E*=4, $\rho_{\beta_{(e_1 e_2)}}$ vary $\rho_{\beta_{(e_1 e_2)}}$ with 0.1-0.9 as well as consider the trans-ethnic genetic correlations are the same or different across different race-pairs. Acknowledging the potential influence of genetic factors across diverse racial backgrounds, this exploration aimed to account for variations in genetic correlation. Furthermore, in an effort to optimize precision and statistical power, we systematically varied the number of SNPs (*p*=25, 50, 100 and 200) while keeping other parameters constant. This process allows us to determine how much the precision and statistical power of causal effect estimates can be significantly improved under different numbers of IVs. Finally, our simulation study was designed to encompass three distinct scenarios: one where pleiotropy was absent ($\gamma_{je} = 0$), another where balanced horizontal pleiotropy was present($\gamma_{je} \sim U(-0.2, 0.2)$), and a third scenario where directional horizontal pleiotropy was present ($\gamma_{je} \sim U(0, 0.2)$). We then applied our method TEMR to estimate causal effects in the target populations. To benchmark the performance of our approach, we conducted a comparative analysis with previously published MR methods [2] based on the Wald ratio, including IVW method [59], MR-Egger [60], Simple Median [61], Weighted Median [62], Simple Mode [63], Weighted Mode [64]. By thoroughly examining these scenarios, we aimed to provide a comprehensive assessment of TEMR's performance and robustness under diverse genetic and phenotypic conditions.

The evaluation metrics include estimation bias, standard error, type I error for testing null causal effect and statistical power for testing non-null causal effect. We utilized boxplots to demonstrate the results of bias and standard error, Q-Q plots to showcase the results of Type I error, and bar charts to depict the results of statistical power.

Application

We applied TEMR to estimate the causal effects between different biomarkers and four diseases (hypertension, ischemic stroke, T2D and schizophrenia) in the East Aisan, African, Hispanic/Latino population, leveraging data from large European cohorts. These diseases were chosen for their significant public health impact, high prevalence, and representative nature of the complex interplay between genetics and environment. The GWAS summary data for the Asian population was mainly sourced from the BBJ with a sample size of 170,000. For the African population, the data was mainly obtained from the Pan-UKB with a sample size of 6,000, and for the Hispanic population from the GWAS Catalog with a sample size of. The GWAS summary data of European population was derived from the UKB with a sample size of 500,000. Details of datasets information are shown in Supplementary Table S7. Firstly, for each trait, we chose SNPs based on the criterion of *P*-value less than 5×10^{-8} in at least one ethnicity and no linkage disequilibrium (r^2<0.001). The SNP satisfying above criterion in at least one of the ethnics are selected as IVs. Then, we applied TEMR and other six MR methods for trans-ethnic MR analysis using the 17 biomarkers and four diseases. For each target population, we use the other three datasets as auxiliary datasets.

Acknowledgements

We thank Haoran Xue for his constructive suggestions.

Author Contributions

HL and FX conceived the study. LH contributed to theoretical derivation with assistance from ZY. SW contributed to the data simulation. LH and SW contributed to the application. LH and SW wrote the manuscript with input from all other authors. All authors reviewed and approved the final manuscript.

Competing Interests statement

The authors declare no competing interests.

Data and code availability

The GWAS summary data in UK Biobank are publicly available at http://www.nealelab.is/uk-biobank. The GWAS summary data in Biobank Japan are publicly available at https://pheweb.jp/. The GWAS summary data in Pan-UKB are publicly available at

https://pan.ukbb.broadinstitute.org/. Other GWAS summary data are publicly available at IEU OpenGWAS project (https://gwas.mrcieu.ac.uk/) and GWAS Catalog (https://www.ebi.ac.uk/gwas/). All the analysis in our article were implemented by R software (version 4.3.2). R packages used in our analysis include *TwoSampleMR*, *MendelianRandomization*, *ggplot*2, *plinkbinr* and *ieugwasr*. TEMR package can be implemented by https://github.com/hhoulei/TEMR. All the codes for simulation are uploaded in https://github.com/hhoulei/TEMR_Simul.

Ethics approval and consent to participate

The data used in our study was all publicly available and obtained written informed consent from all participants.

Source of Funding

HL was supported by the National Key Research and Development Program of Chinaunder (Grant 2022YFC3502100), National Natural Science Foundation of China (Grant 82003557) and Shandong Province Key R&D Program Project (2021SFGC0504). FX was supported by the National Natural Science Foundation of China (Grant 82173625).

References

[1]. Emdin, C. A., Khera, A. V., & Kathiresan, S. (2017). Mendelian randomization. Jama, 318(19), 1925-1926.

[2]. Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., ... & Davey Smith, G. (2022). Mendelian randomization. Nature Reviews Methods Primers, 2(1), 6.

[3]. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Marchini, J. (2018). The UK Biobank resource with deep phenoty** and genomic data. Nature, 562(7726), 203-209.

[4]. Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., ... & Allen, N. E. (2020). The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature communications, 11(1), 2624.

[5]. Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., ... & Ferreira, M. A. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature, 599(7886), 628-634.

[6]. Sun, B. B., Chiou, J., Traylor, M., Benner, C., Hsu, Y. H., Richardson, T. G., ... & Whelan, C. D. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature, 622(7982), 329-338.

[7]. Julkunen, H., Cichońska, A., Tiainen, M., Koskela, H., Nybo, K., Mäkelä, V., ... & Würtz, P. (2023). Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nature Communications, 14(1), 604.

[8]. Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., ... & Waring, J. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature, 613(7944), 508-518.

[9]. Pan-UKB team. https://pan.ukbb.broadinstitute.org. 2020.

[10]. Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., **ao, C., Toneva, I., ... & 1000 Genomes Project Consortium. (2012). The 1000 Genomes Project: data management and community access. Nature methods, 9(5), 459-462.

[11]. Kubo, M. (2017). BioBank Japan project: epidemiological study. Journal of epidemiology, 27(3 Suppl), S1.

[12]. He, Y., Koido, M., Sutoh, Y., Shi, M., Otsuka-Yamasaki, Y., Munter, H. M., ... & Kamatani, Y. (2023). East Asian-specific and cross-ancestry genome-wide meta-analyses provide mechanistic insights into peptic ulcer disease. Nature Genetics, 1-10.

[13]. Feng, Y. A., Chen, C. Y., Chen, T. T., Kuo, P. H., Hsu, Y. H., Yang, H. I., Chen, W. J., Su, M.

W., Chu, H. W., Shen, C. Y., Ge, T., Huang, H., & Lin, Y. F. (2022). Taiwan Biobank: A rich biomedical research database of the Taiwanese population. Cell genomics, 2(11), 100197. https://doi.org/10.1016/j.xgen.2022.100197

[14]. Walters, R. G., Millwood, I. Y., Lin, K., Schmidt Valle, D., McDonnell, P., Hacker, A., Avery, D., Edris, A., Fry, H., Cai, N., Kretzschmar, W. W., Ansari, M. A., Lyons, P. A., Collins, R., Donnelly, P., Hill, M., Peto, R., Shen, H., Jin, X., Nie, C., … China Kadoorie Biobank Collaborative Group (2023). Genotyping and population characteristics of the China Kadoorie Biobank. Cell genomics, 3(8), 100361. https://doi.org/10.1016/j.xgen.2023.100361

[15]. Tahir, U. A., Katz, D. H., Avila-Pachecho, J., Bick, A. G., Pampana, A., Robbins, J. M., ... & Gerszten, R. E. (2022). Whole genome association study of the plasma metabolome identifies metabolites linked to cardiometabolic disease in black individuals. Nature Communications, 13(1), 4923.

[16]. Zhao, H., Rasheed, H., Nøst, T. H., Cho, Y., Liu, Y., Bhatta, L., ... & Zheng, J. (2022). Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. Cell Genomics, 2(11).

[17]. Vargas, L. B., Lange, L. A., Ferrier, K., Aguet, F., Ardlie, K., Gabriel, S., ... & Konigsberg, I. R. (2023). Gene expression associations with body mass index in the Multi-Ethnic Study of Atherosclerosis. International Journal of Obesity, 47(2), 109-116.

[18]. Feofanova, E. V., Chen, H., Dai, Y., Jia, P., Grove, M. L., Morrison, A. C., ... & Yu, B. (2020). A genome-wide association study discovers 46 loci of the human metabolome in the Hispanic Community Health Study/Study of Latinos. The American Journal of Human Genetics, 107(5), 849-863.

[19]. Yang, G., Mishra, M., & Perera, M. A. (2023). Multi-Omics Studies in Historically Excluded Populations: The Road to Equity. Clinical Pharmacology & Therapeutics, 113(3), 541-556.

[20]. Kim, W., Cho, M. H., Sakornsakolpat, P., Lynch, D. A., Coxson, H. O., Tal-Singer, R., ... & Beaty, T. H. (2019). DSP variants may be associated with longitudinal change in quantitative emphysema. Respiratory research, 20, 1-10.

[21]. Manichaikul, A., Wang, X. Q., Sun, L., Dupuis, J., Borczuk, A. C., Nguyen, J. N., ... & Lederer, D. J. (2017). Genome-wide association study of subclinical interstitial lung disease in MESA. Respiratory research, 18(1), 1-11.

[22]. Lewis, A. C., Molina, S. J., Appelbaum, P. S., Dauda, B., Di Rienzo, A., Fuentes, A., ... & Allen, D. S. (2022). Getting genetic ancestry right for science and society. Science, 376(6590), 250-252.

[23]. Petrovski, S., & Goldstein, D. B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome biology, 17, 1-3.

[24]. Shi, H., Gazal, S., Kanai, M., Koch, E. M., Schoech, A. P., Siewert, K. M., ... & Price, A. L. (2021). Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nature communications, 12(1), 1098.

[25]. Brown, B. C., Ye, C. J., Price, A. L., & Zaitlen, N. (2016). Transethnic genetic-correlation estimates from summary statistics. The American Journal of Human Genetics, 99(1), 76-88.

[26]. Zheng J, Zhang Y, Rasheed H, et al. Trans-ethnic Mendelian-randomization study reveals causal relationships between cardiometabolic factors and chronic kidney disease. Int J Epidemiol. 2022;50(6):1995-2010.

[27]. Wu, S., Kong, M., Song, Y., & Peng, A. (2023). Ethnic disparities in bidirectional causal effects between serum uric acid concentrations and kidney function: Trans-ethnic Mendelian randomization study. Heliyon, 9(11).

[28]. Xiu X, Zhang H, Xue A, et al. Genetic evidence for a causal relationship between type 2 diabetes and peripheral artery disease in both Europeans and East Asians. BMC Med. 2022;20(1):300. Published 2022 Aug 31. doi:10.1186/s12916-022-02476-0

[29]. Morris AP. Transethnic meta-analysis of genomewide association studies. Genet Epidemiol. 2011;35(8):809-822. doi:10.1002/gepi.20630

[30]. Mägi R, Horikoshi M, Sofer T, et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum Mol Genet. 2017;26(18):3639-3650. doi:10.1093/hmg/ddx280

[31]. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. 2014;6(10):91. Published 2014 Oct 31. doi:10.1186/s13073-014-0091-5

[32]. Cordero, R. Y., Cordero, J. B., Stiemke, A. B., Datta, L. W., Buyske, S., Kugathasan, S., ... & Simpson, C. L. (2023). Trans-ancestry, Bayesian meta-analysis discovers 20 novel risk loci for inflammatory bowel disease in an African American, East Asian and European cohort. Human molecular genetics, 32(5), 873-882.

[33]. Cai M, Xiao J, Zhang S, et al. A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am J Hum Genet. 2021;108(4):632-655. doi:10.1016/j.ajhg.2021.03.002

[34]. Hoggart C, Choi SW, García-González J, Souaiaia T, Preuss M, O'Reilly P. BridgePRS: A powerful trans-ancestry Polygenic Risk Score method. Preprint. bioRxiv. 2023;2023.02.17.528938.

Published 2023 Feb 21. doi:10.1101/2023.02.17.528938

[35]. Zhang H, Zhan J, Jin J, et al. A new method for multiancestry polygenic prediction improves performance across diverse populations. Nat Genet. 2023;55(10):1757-1768. doi:10.1038/s41588-023-01501-z

[36]. Li, Z., Zhao, W., Shang, L., Mosley, T. H., Kardia, S. L., Smith, J. A., & Zhou, X. (2022). METRO: Multi-ancestry transcriptome-wide association studies for powerful gene-trait association detection. The American Journal of Human Genetics, 109(5), 783-801.

[37]. Chen, F., Wang, X., Jang, S. K., Quach, B. C., Weissenkampen, J. D., Khunsriraksakul, C., ... & Liu, D. J. (2023). Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. Nature genetics, 55(2), 291-300.

[38]. Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. Computer Journal, 7, 308–313. doi:10.1093/comjnl/7.4.308.

[39]. Ryan MC, Collins P, Thakore JH. Impaired fasting glucose tolerance in first-episode, drug-naive patients with schizophrenia. Am J Psychiatry. 2003 Feb;160(2):284-9.

[40]. Huang TL, Chen JF. Serum lipid profiles and schizophrenia: effects of conventional or atypical antipsychotic drugs in Taiwan. Schizophr Res. 2005 Dec 1;80(1):55-9. doi: 10.1016/j.schres.2005.05.001

[41]. Zhang G, Ye X, Wang X, et al. Serum total cholesterol levels associated with immediate memory performance in patients with chronic schizophrenia. Schizophr Res. 2023 May;255:256-260.

[42]. Bhat, Tariq, Sumaya Teli, et al. (2013). Neutrophil to Lymphocyte Ratio and Cardiovascular Diseases: A Review. Expert Review of Cardiovascular Therapy, 11(1), 55-59.

[43]. Guo W, Song Y, Sun Y, Du H, Cai Y, You Q, Fu H, Shao L. Systemic immune-inflammation index is associated with diabetic kidney disease in Type 2 diabetes mellitus patients: Evidence from NHANES 2011-2018. Front Endocrinol (Lausanne). 2022 Dec 6;13:1071465.

[44]. Zhou X, Wang X, Li R, Yan J, Xiao Y, Li W, Shen H. Neutrophil-to-Lymphocyte Ratio Is Independently Associated With Severe Psychopathology in Schizophrenia and Is Changed by Antipsychotic Administration: A Large-Scale Cross-Sectional Retrospective Study. Front Psychiatry. 2020 Oct 30;11:581061.

[45]. Steiner J, Bernstein HG, Schiltz K, Müller UJ, Westphal S, Drexhage HA, Bogerts B. Immune system and glucose metabolism interaction in schizophrenia: a chicken-egg dilemma. Prog Neuropsychopharmacol Biol Psychiatry. 2014 Jan 3;48:287-94.

[46]. CAO Ting, LI Nana, CAI Hualin. Progress in neurosteroids related to the pathogenesis and

treatment of schizophrenia. Chinese Journal of Clinical Pharmacology and Therapeutics, 2018, 23(6): 709-714.

[47]. Mohammadi A，Rashidi E，Amooeian VG. Brain，blood，cerebrospinal fluid，and serum biomarkers in schizophrenia. Psychiatry Res, 2018. 265:25-38.

[48]. Zhang J, Liu Q. Cholesterol metabolism and homeostasis in the brain. Protein Cell. 2015 Apr;6(4):254-64.

[49]. Garcia, C.; Andersen, C.J.; Blesso, C.N. The Role of Lipids in the Regulation of Immune Responses. Nutrients 2023, 15, 3899.

[50]. Yan J, Horng T. Lipid Metabolism in Regulation of Macrophage Functions. Trends Cell Biol. 2020 Dec;30(12):979-989.

[51]. Koskuvi M, Pörsti E, Hewitt T, Räsänen N, Wu YC, Trontti K, McQuade A, Kalyanaraman S, Ojansuu I, Vaurio O, Cannon TD, Lönnqvist J, Therman S, Suvisaari J, Kaprio J, Blurton-Jones M, Hovatta I, Lähteenvuo M, Rolova T, Lehtonen Š, Tiihonen J, Koistinaho J. Genetic contribution to microglial activation in schizophrenia. Mol Psychiatry. 2024 Mar 22.

[52]. Hansen DL, Maquet J, Lafaurie M, Möller S, Berentsen S, Frederiksen H, Moulis G, Gaist D. Primary autoimmune haemolytic anaemia is associated with increased risk of ischaemic stroke: A binational cohort study from Denmark and France. Br J Haematol. 2024 Mar;204(3):1072-1081.

[53]. Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., ... & Price, A. L. (2016). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics, 47(3), 291-295.

[54]. Ning, Z., Pawitan, Y., & Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. Nature genetics, 52(8), 859-864.

[55]. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2016). GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. Proceedings of the National Academy of Sciences, 113(32), E4579-E4580.

[56]. Loh, P. R. (2018). BOLT-LMM v2. 3.2 user manual. Available oniline at: https://data. broadinstitute. org/alkesgroup/BOLT-LMM/(accessed May 2, 2019).

[57]. Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., & Schaid, D. J. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics, 200(3), 719-736.

[58]. Bischoff, W., & Fieger, W. (1991). Characterization of the multivariate normal distribution by conditional normal distributions. Metrika, 38, 239-248.

[59]. Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. Computer

Journal, 7, 308–313. doi:10.1093/comjnl/7.4.308.

[60]. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015 Apr;44(2):512-25. doi: 10.1093/ije/dyv080.

[61]. Lawlor DA, Harbord RM, Sterne JAC, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statist Med, 2008, 27:1133-1163. doi: 10.1002/sim.3034.

[62]. Bowden J, Davey Smith G, Haycock PC, et al. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol, 2016, 40(4):304-314. doi: 10.1002/gepi.21965.

[63]. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol 2017;46:1985–98. doi: 10.1093/ije/dyx102.

[64]. Burgess S, Zuber V, Gkatzionis A, Foley CN. Modal-based estimation via heterogeneity -penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. Int J Epidemiol. 2018 Aug 1;47(4):1242-1254. doi: 10.1093/ije/dyy080.

Figure legends and Tables

Figure 1. TEMR flowchart

(A) The example of multiple ethnics, which also are the ethnics that we are interested in the applied example. (B) The aim of TEMR is to improve the statistical power and estimation accuracy of MR in target population only using trans-ethnic large-scale auxiliary dataset. (C) The flowchart of TEMR model, take two ethnics as example, one target population and one auxiliary population.

Figure 2. Simulation results for causal effect estimation in the target population when there is one auxiliary population (no pleiotropy).

Sample size of target population is 3,000 and the sample size of auxiliary population is 300,000. IVs include 100 common SNPs. **A)** Boxplots show the performances of causal effect estimation in target population; **B-C)** Q-Q plots show the performances of Type I error rates of zero causal effect estimation in target population when the causal effect of auxiliary population is 0 and 0.05, respectively; **D-E)** Bar chart plots show the performances of statistical power of non-zero causal effect estimation in target population when the causal effect of auxiliary population is 0 and 0.05, respectively. IVW, Inverse-variance weighted method.

Figure 3. Simulation results for causal effect estimation in the target population when there is one auxiliary population (directional horizontal pleiotropy).

Sample size of target population is 3,000, and the sample size of auxiliary population is 300,000. IVs include 100 common SNPs. **A)** Boxplots show the performances of causal effect estimation in target population; **B-C)** Q-Q plots show the performances of Type I error rates of zero causal effect estimation in target population when the causal effect of auxiliary population is 0 and 0.05, respectively; **D-E)** Bar chart plots show the performances of statistical power of non-zero causal effect estimation in target population when the causal effect of auxiliary population is 0 and 0.05, respectively. IVW, Inverse-variance weighted method.

Figure 4. **Simulation results for causal effect estimation in the target population with different number of SNPs.**

Continuous outcome, no horizontal pleiotropy. Sample size of target population is 3,000, and the sample size of auxiliary population is 300,000. **A)** Boxplots show the performances of causal effect estimation in target population; **B-C)** Bar chart plots show the performances of statistical power of non-zero causal effect estimation in target population. IVW, Inverse-variance weighted method.

Figure 5. Simulation results for causal effect estimation in the target population when there are multiple auxiliary populations.

Continuous outcome. No pleiotropy. Sample size of target population is 3,000, and the sample size

of auxiliary populations are 3,000, 3,000, 300,000. IVs include 100 common SNPs. A) Boxplots show the performances of causal effect estimation in target population; B) Q-Q plots show the performances of Type I error rates of zero causal effect estimation in target population; C) Bar chart plots show the performances of statistical power of non-zero causal effect estimation in target population. IVW, Inverse-variance weighted method.

Figure 6. Heatmap of trans-ethnic genetic correlation for 22 biomarkers and four diseases.

The color intensity indicates the strength of the correlation. Warmer colors, tending towards red, signify a correlation coefficient approaching 1, indicating a strong positive correlation. Conversely, cooler colors, leaning towards blue, denote a correlation coefficient nearing -1, suggesting a strong negative correlation.

Figure 7. Results of trans-ethnic MR analysis for causal relationships from 22 biomarkers to four diseases.

Different colors represent the $-\log_{10}(P)$ calculated by different methods. The triangle points represent the significant relationships in TEMR results but not significant in other methods. The solid or dashed points indicate whether the causal effects are significant ($P<0.05$). In cases where the MR-Egger test suggests the presence of horizontal pleiotropy between biomarker pairs, the P-values presented are those adjusted for such pleiotropy.

Supplementary File

Supplementary Materials and Methods

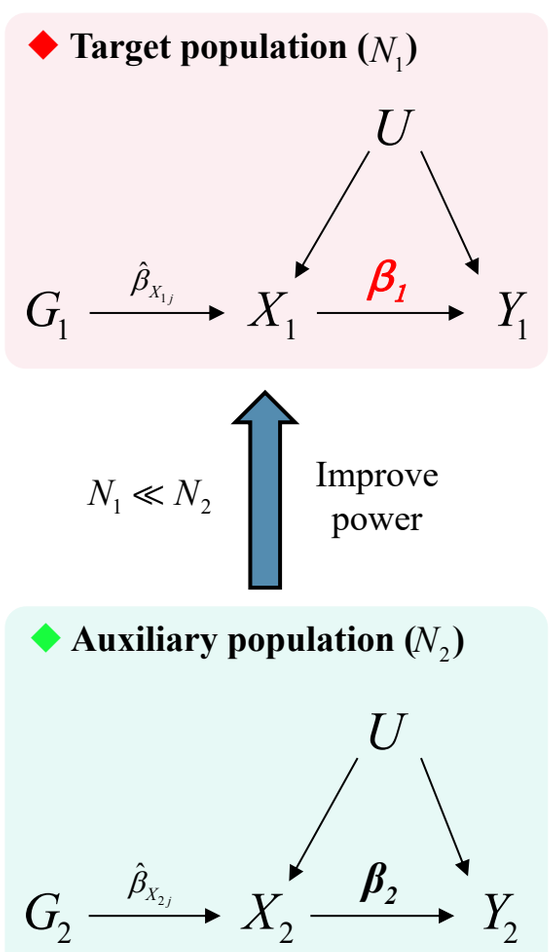
Details of methods and results of simulation.

Supplementary Table

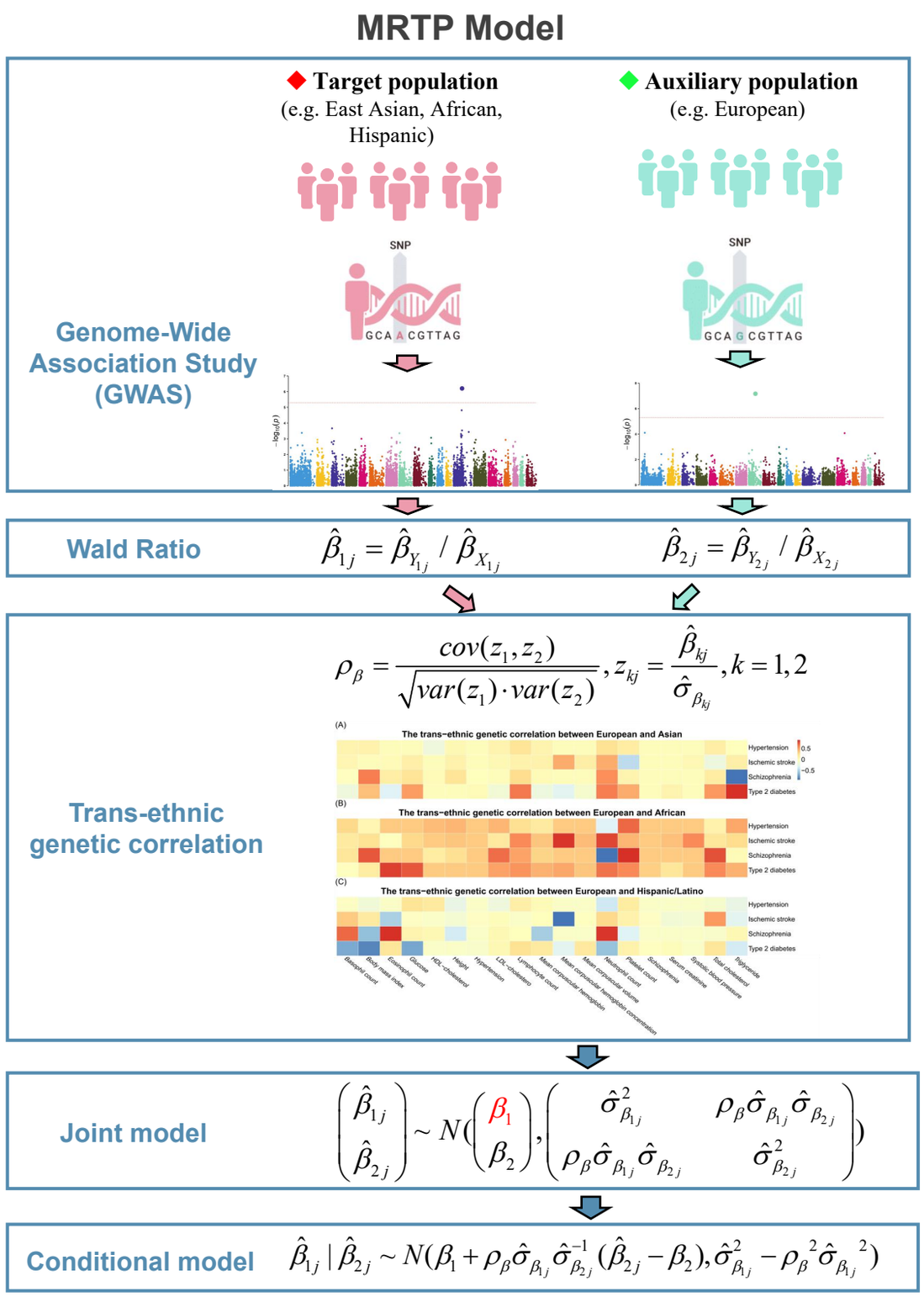
GWAS summary datasets information and results of application.

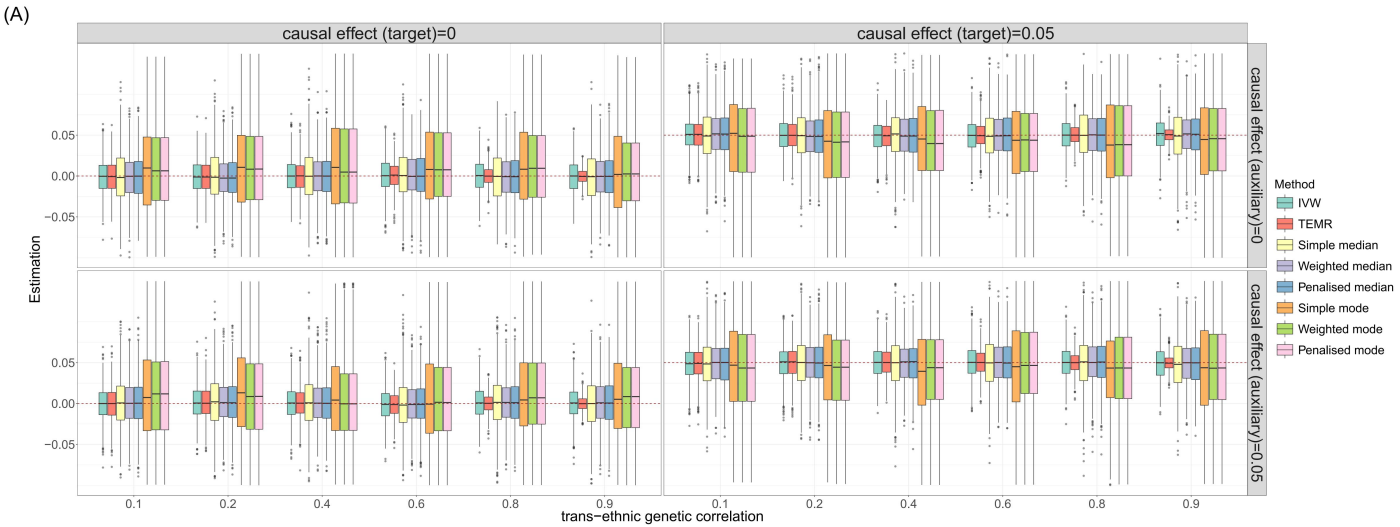
(A)

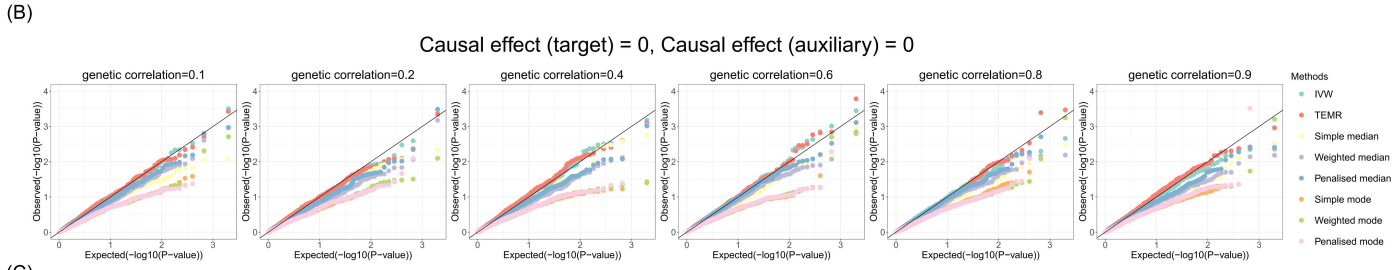
(B)

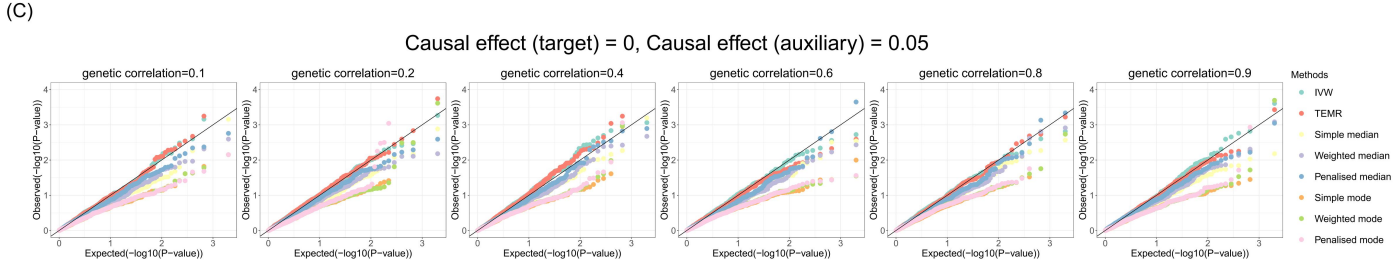


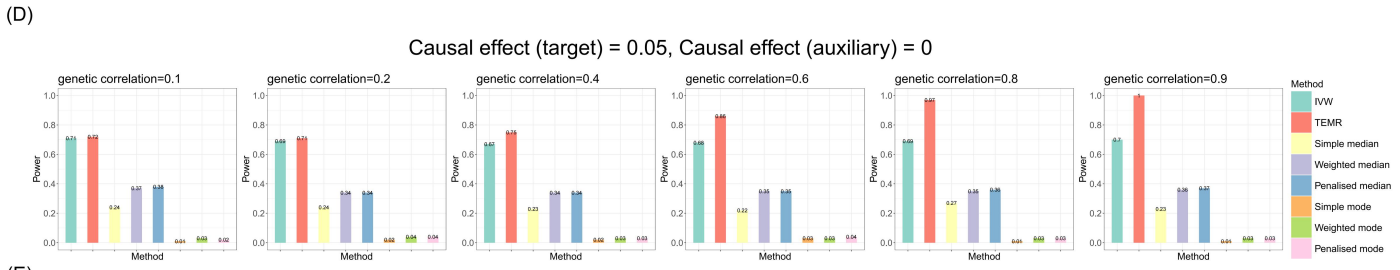
(C)

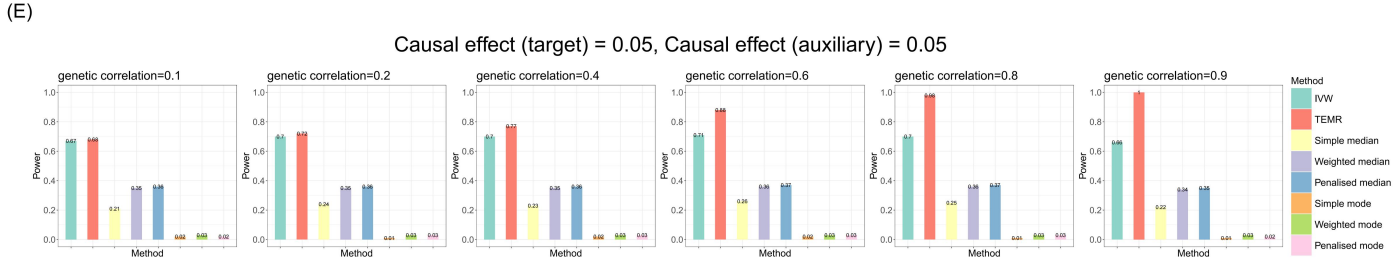


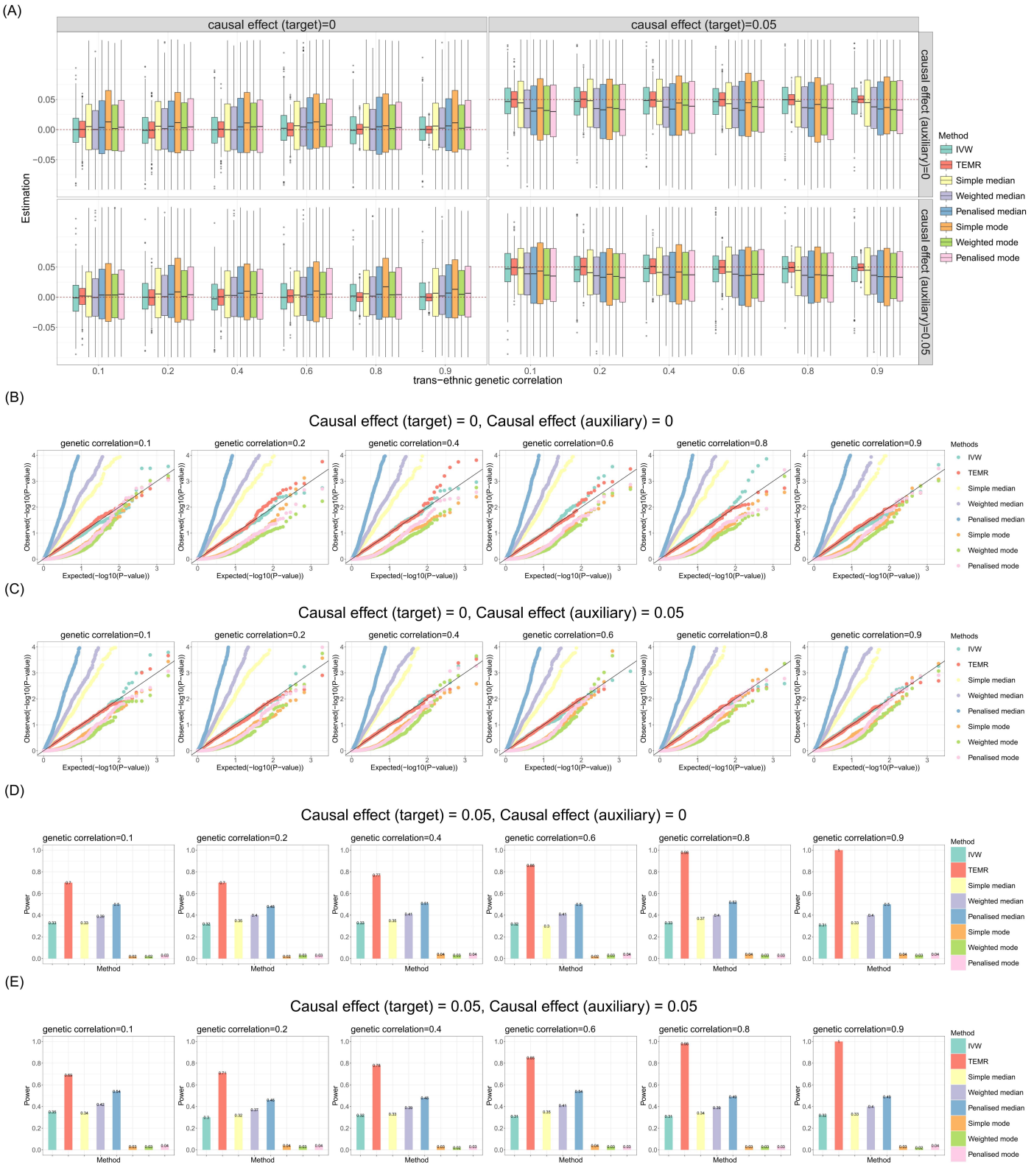

(A)
causal effect (target)=0
causal effect (target)=0.05
causal effect (auxiliary)=0
causal effect (auxiliary)=0.05
Estimation
trans−ethnic genetic correlation
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode


(B)
Causal effect (target) = 0, Causal effect (auxiliary) = 0
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Observed(−log10(P−value))
Expected(−log10(P−value))
Methods
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

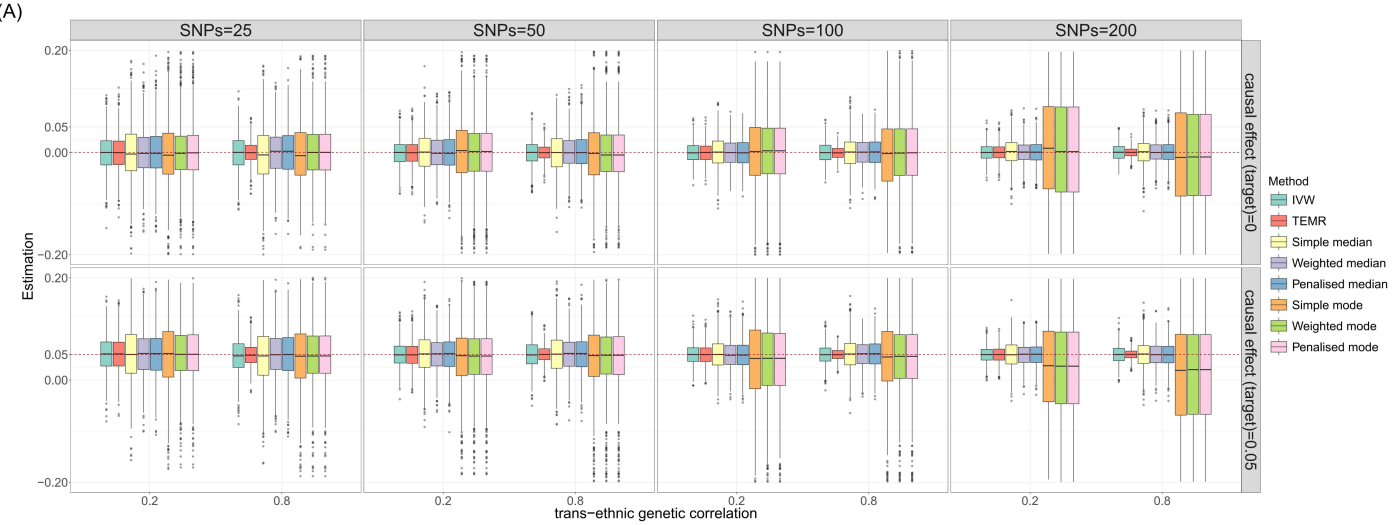

(C)
Causal effect (target) = 0, Causal effect (auxiliary) = 0.05
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Observed(−log10(P−value))
Expected(−log10(P−value))
Methods
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode


(D)
Causal effect (target) = 0.05, Causal effect (auxiliary) = 0
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Power
Method
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

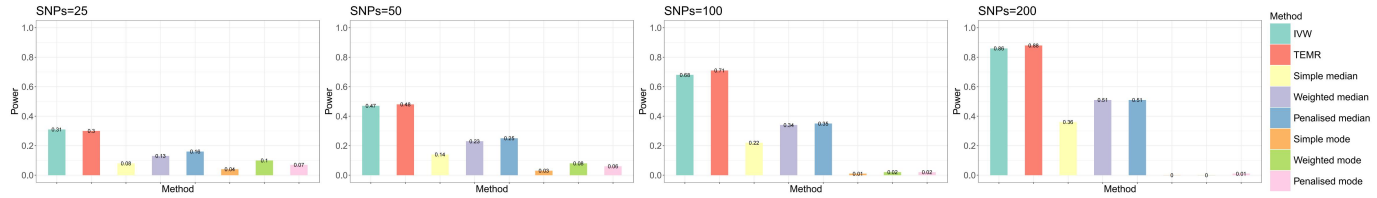

(E)
Causal effect (target) = 0.05, Causal effect (auxiliary) = 0.05
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Power
Method
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode


(A)
causal effect (target)=0
causal effect (target)=0.05
causal effect (auxiliary)=0
causal effect (auxiliary)=0.05
Estimation
trans−ethnic genetic correlation
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode
(B)
Causal effect (target) = 0, Causal effect (auxiliary) = 0
(C)
Causal effect (target) = 0, Causal effect (auxiliary) = 0.05
(D)
Causal effect (target) = 0.05, Causal effect (auxiliary) = 0
(E)
Causal effect (target) = 0.05, Causal effect (auxiliary) = 0.05
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Observed(−log10(P−value))
Expected(−log10(P−value))
Power
Methods

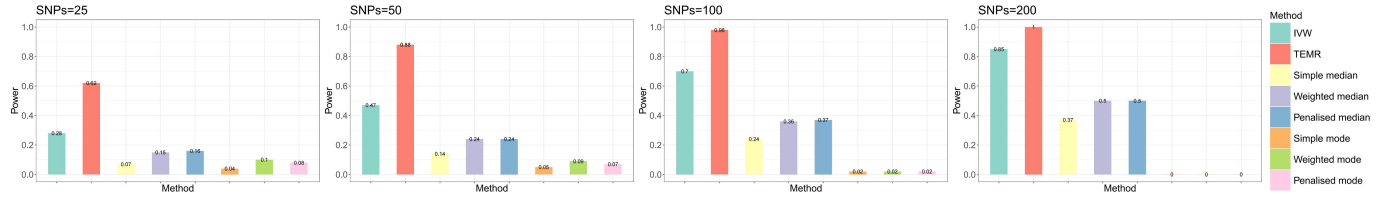
(A)


SNPs=25
SNPs=50
SNPs=100
SNPs=200
causal effect (target)=0
causal effect (target)=0.05
Estimation
0.20
0.05
0.00
−0.20
0.2
0.8
trans−ethnic genetic correlation
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

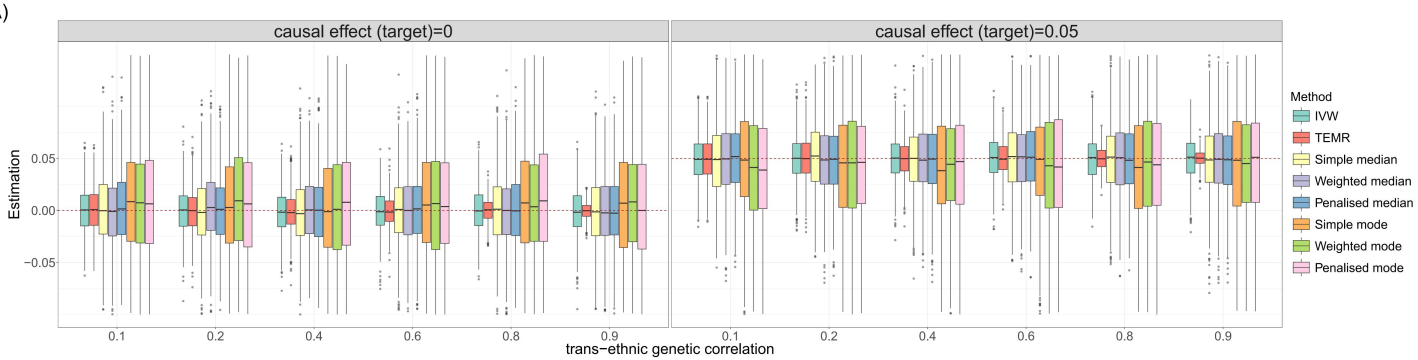
(B)


Causal effect (target) = Causal effect (auxiliary) = 0.05, trans−ethnic genetic correlation = 0.2
SNPs=25
SNPs=50
SNPs=100
SNPs=200
Power
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

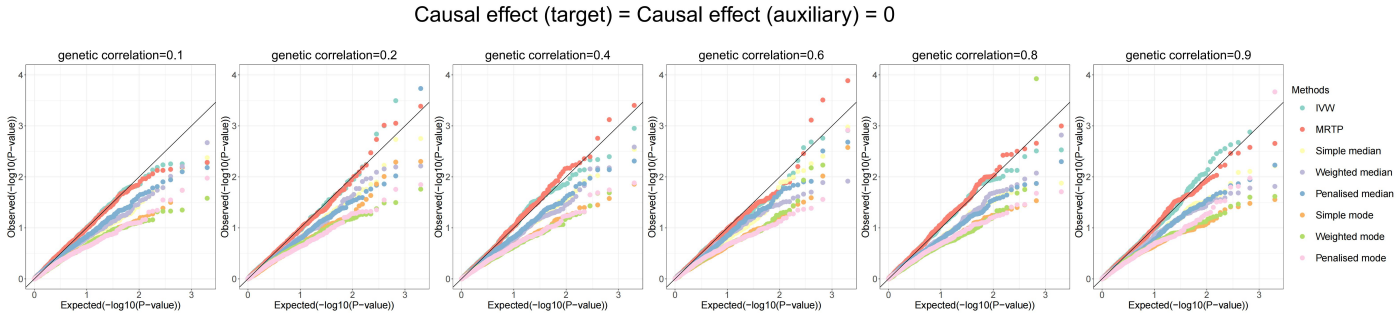
(C)


Causal effect (target) = Causal effect (auxiliary) = 0.05, trans−ethnic genetic correlation = 0.8
SNPs=25
SNPs=50
SNPs=100
SNPs=200
Power
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

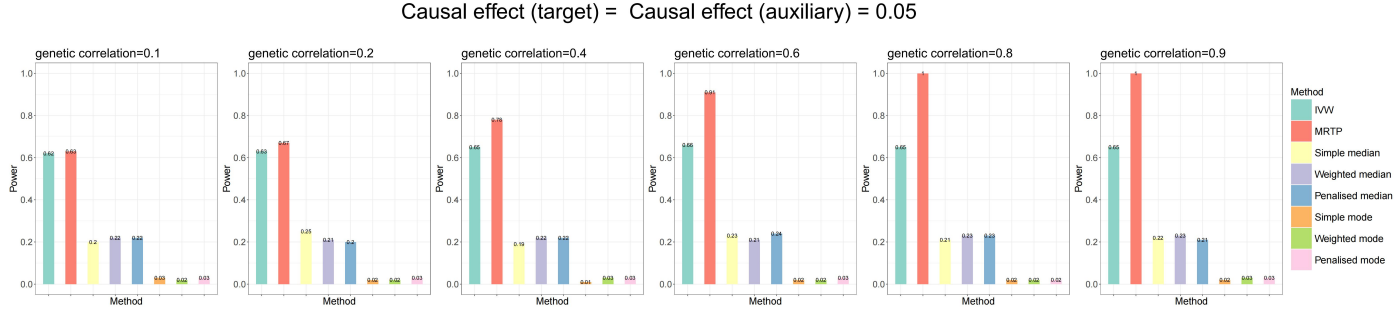
(A)


causal effect (target)=0
causal effect (target)=0.05
Estimation
0.05
0.00
−0.05
0.1
0.2
0.4
0.6
0.8
0.9
trans−ethnic genetic correlation
Method
IVW
TEMR
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

(B)

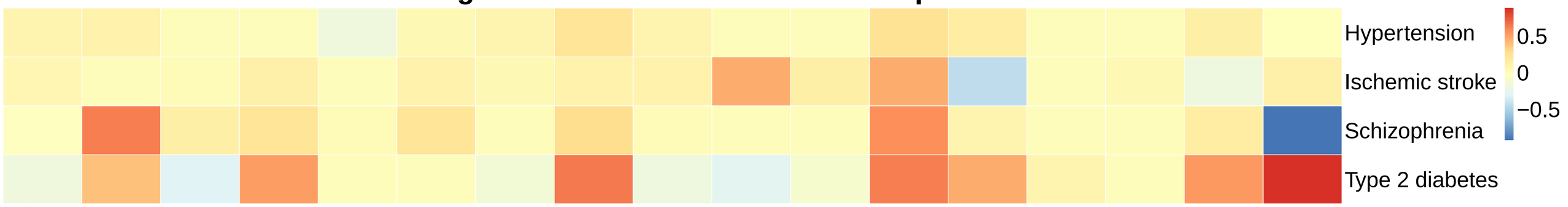

Causal effect (target) = Causal effect (auxiliary) = 0
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Observed(−log10(P−value))
Expected(−log10(P−value))
Methods
IVW
MRTP
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

(C)


Causal effect (target) = Causal effect (auxiliary) = 0.05
genetic correlation=0.1
genetic correlation=0.2
genetic correlation=0.4
genetic correlation=0.6
genetic correlation=0.8
genetic correlation=0.9
Power
Method
Method
IVW
MRTP
Simple median
Weighted median
Penalised median
Simple mode
Weighted mode
Penalised mode

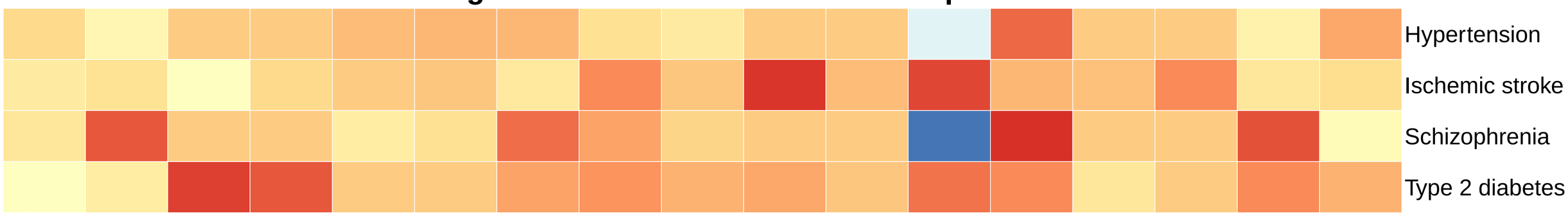
(A)

The trans−ethnic genetic correlation between European and Asian



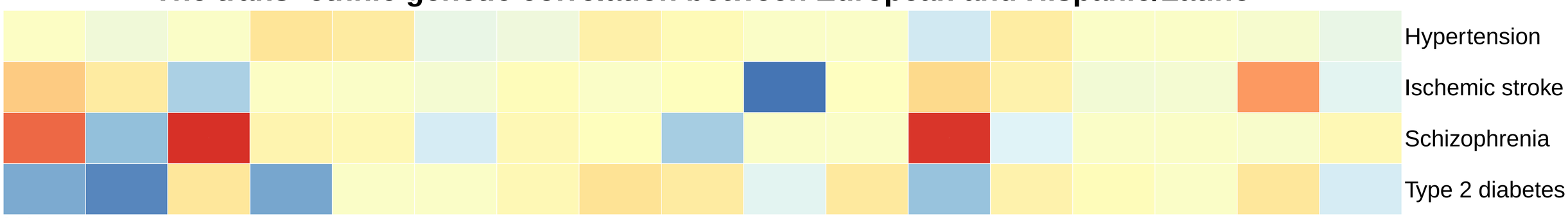
(B)

The trans−ethnic genetic correlation between European and African



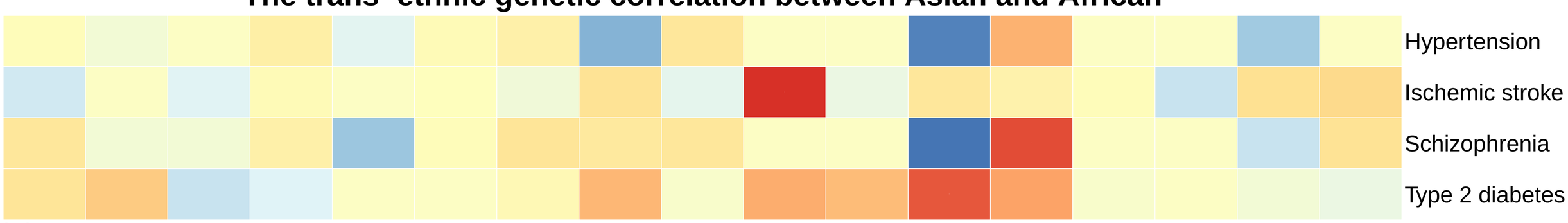
(C)

The trans−ethnic genetic correlation between European and Hispanic/Latino



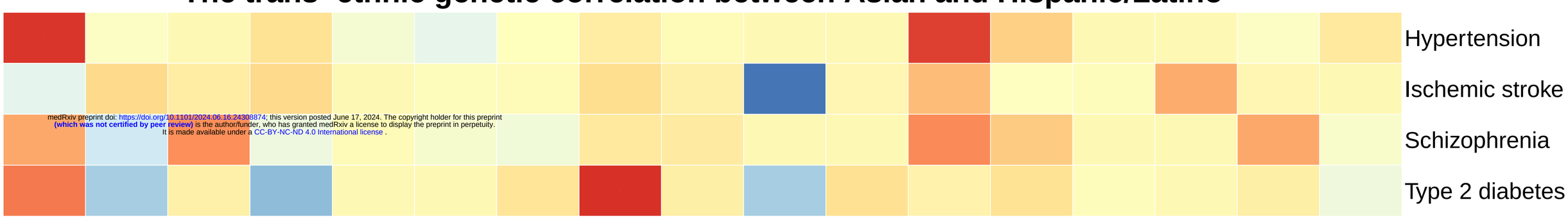
(D)

The trans−ethnic genetic correlation between Asian and African



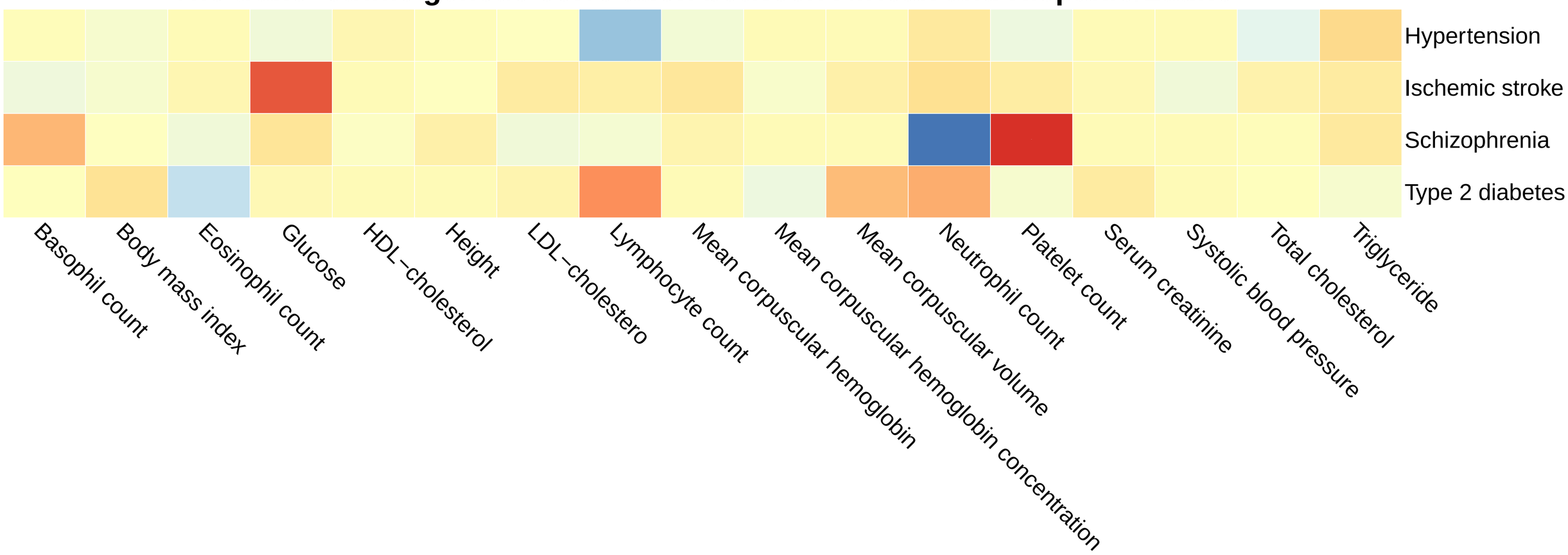
(E)

The trans−ethnic genetic correlation between Asian and Hispanic/Latino

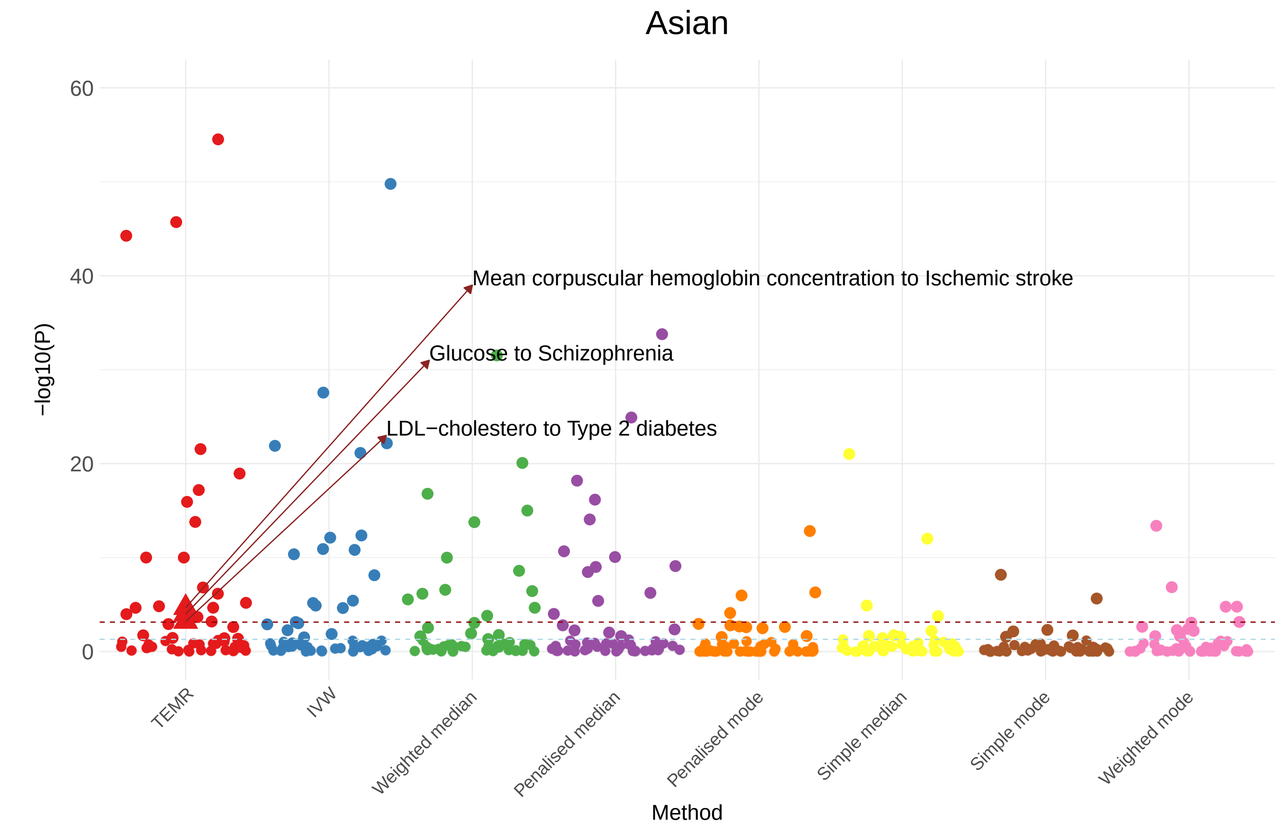


(F)

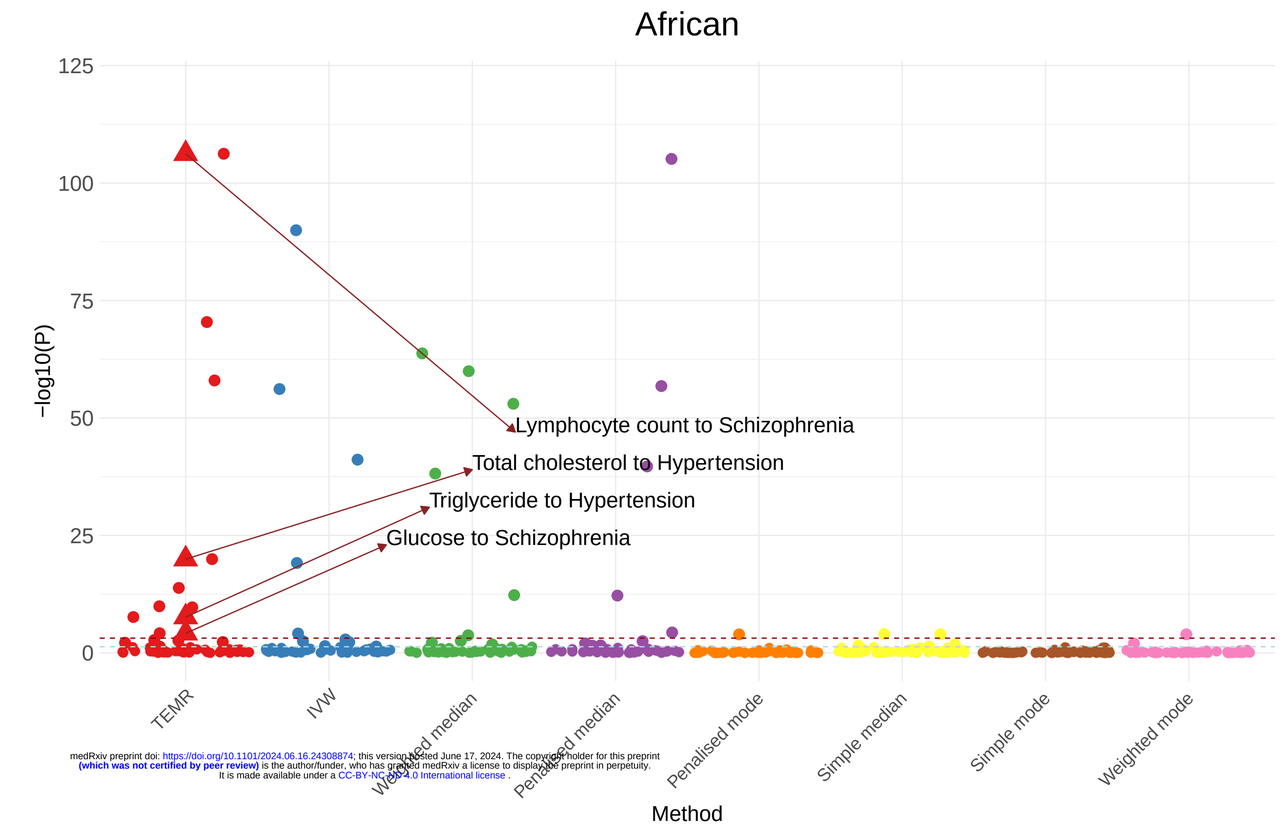
The trans−ethnic genetic correlation between African and Hispanic/Latino



(A)



(B)



(C)