Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PCA-MR - A powerful framework for cis-Mendelian Randomization with correlated variants and sparse exposures

Christopher Lawrence

chrislaw133@gmail.com

Wake Forest University Information Systems

Research Article

Keywords:

Posted Date: October 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10095364/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PCA-MR - A powerful framework for cis-Mendelian Randomization with correlated variants and sparse exposures

Authors:

Christopher Lawrence[1, 2]

Corresponding Authors:

Christopher Lawrence: chrislaw133@gmail.com

Affiliations:

[1]Wake Forest Department of Biology, Winston-Salem, NC, USA

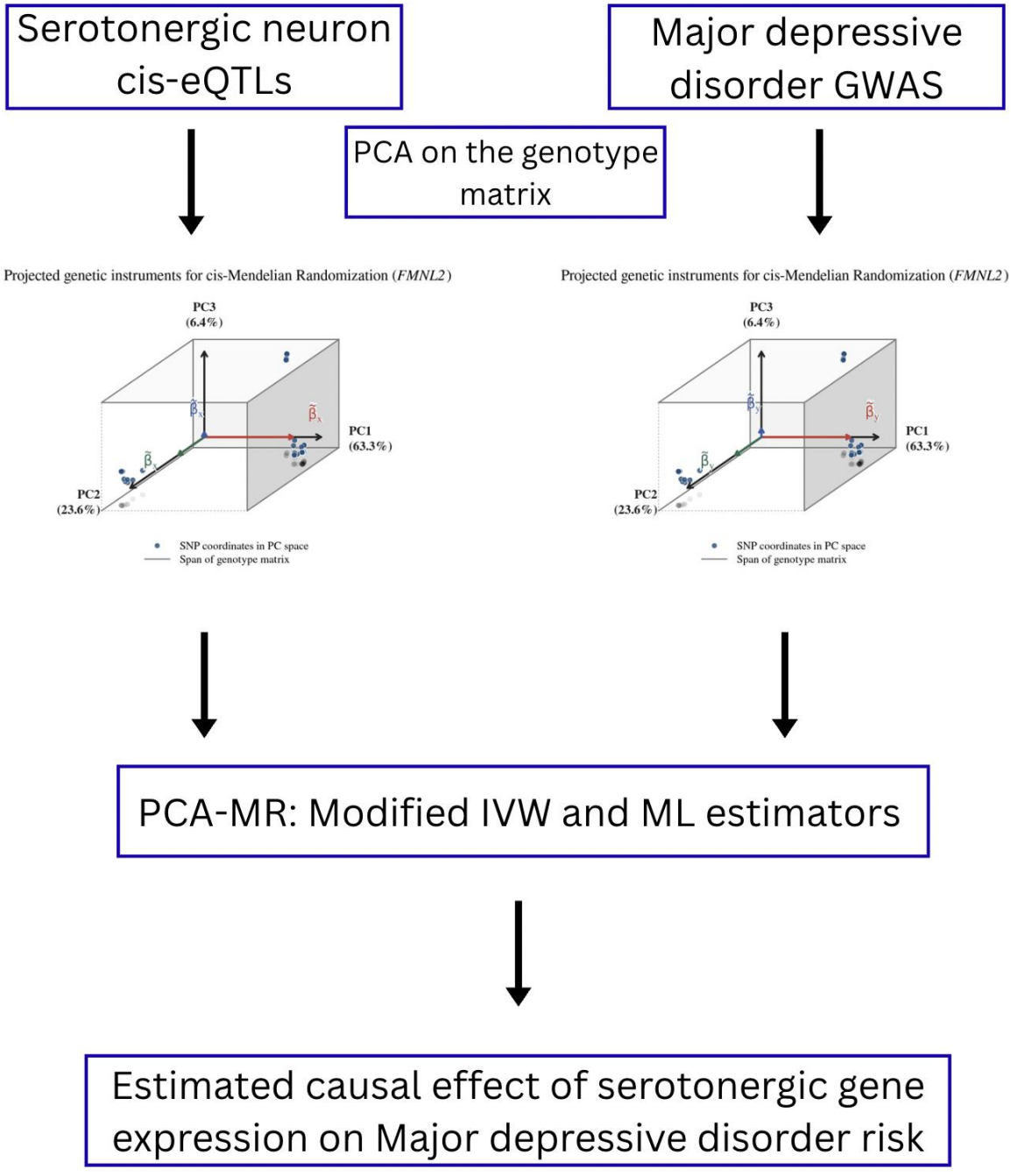
[2]Wake Forest University Information Systems, WFU High Performance Computing Facility, Winston-Salem, NC, USA

Abstract

Inferring the causal genes of complex phenotypes remains a central challenge in human genetics. Unlike traditional Mendelian Randomization (MR), cis-MR is applied to genomic loci using cis-genetic variants. Leveraging the quantitative trait loci for a gene or protein as exposure to disease outcomes can be a powerful and cost-effective method for identifying potential drug targets. However, cis-regions are typically in high linkage-disequilibrium and exposure cohorts are usually sparse in practice. To address these challenges, we propose a novel dimensionality reduction approach called principal component analysis-Mendelian randomization (PCA-MR), with modified inverse variance weighting (IVW) and maximum likelihood (ML) estimators. Simulation studies demonstrated the superiority of PCA-MR-IVW and PCA-MR-ML compared to their LD-pruning counterparts and the existing PCA approach, showing reduced weak-

instrument bias, improved power, controlled type-I error (TIE), and lower root mean square error (RMSE). We apply PCA-MR in two novel contexts, inferring the causal effect of averaged gene expression and time-cumulative gene expression in human iPSC derived serotonin neurons on major depressive disorder risk.

Graphical Abstract



Introduction

Mendelian randomization is a technique that uses single nucleotide polymorphisms (SNPs) as instrumental variables (IVs) to infer the causal effect of biological exposures on complex phenotypes [1,2]. The procedure is analogous to performing a randomized controlled trial (RCT) where genetic variants are assumed to be unconfounded predictors, as they are randomly assorted and fixed at conception. Compared to traditional RCTs that separate control and treatment cohorts, susceptibility to confounding and reverse causation is significantly reduced [3]. Additionally, gathering participants for traditional RCTs can be challenging and the cost of performing experiments can be quite expensive. Conversely, single-cell and bulk tissue RNA-seq datasets are becoming increasingly available to perform cis-MR with expression quantitative trait loci (eQTL) as exposure to disease outcomes. Disease outcomes are typically derived from genome-wide association study (GWAS) summary statistics, which are publicly available and often include tens or hundreds of thousands of participants. When used correctly, cis-MR can be an effective method to elucidate causal biological exposures and prioritize candidate drug targets for therapeutics.

Consider we have an exposure X and we aim to estimate its causal effect $\hat{\theta}$ on the outcome Y using cis-genetic variants as instrumental variables. Regression can be performed to infer the true causal effect θ given the following assumptions: instrumental variables are (1) strongly associated with X; (2) not associated with confounders of the exposure-outcome relationship; (3) associated with Y only through its association to X [4]. Given these assumptions, the causal effect can be modeled as:

$$\beta_{Y,i} = \hat{\theta}\beta_{X,i} \tag{1}$$

where $\beta_{Y,i}$ is the regression coefficient of Y on G_i and $\beta_{X,i}$ is the regression coefficient of X on G_i.

Because recombination events occur less frequently between genetic variants in close proximity, SNPs adjacent to causal variants will often be inherited in tandem [5]. As a result, selected IVs for cis-MR are typically highly correlated [6] due to the propagation of marginal effects from causal (or pleiotropic) variants to neighboring SNPs, combined with conventional IV selection based on exposure association p-value thresholds. This can induce severe degrees of multicollinearity and lead to numerically unstable estimates unless some form of dimensionality reduction is applied [7]. LD-pruning is commonly employed at $r^2 < 0.8$ to reduce instrument collinearity, however in many cases this discards a large proportion of genetic variants, many of which are valid IVs. Although this stabilizes inference, it comes at the cost of instrument retention and statistical power.

To address the challenge of correlated instruments, we propose Principal component analysis-Mendelian randomization (PCA-MR), building on the prior PCA approach (PCA-IVW) [8]. Our framework performs PCA on the genotype matrix of IVs, diagonalizing their correlation. Genetic associations and their standard errors of the exposure and outcome are projected onto these principal components and used as genetic instruments. We additionally propose inverse-variance weighting and maximum likelihood estimators that account for the implied relevance of instruments. Cochran's Q test was implemented to screen for excess heterogeneity, which is assumed to be induced by potential horizontal pleiotropy. Through extensive computer simulations of realistic scenarios, we evaluate type I error, power, and sensitivity analysis in the presence of invalid IVs with uncorrelated and correlated pleiotropic effects. PCA-MR was extended to real datasets, inferring the causal effect of the averaged and

time-cumulative gene expression in human iPSC derived serotonin neurons on major depressive disorder risk. We performed the first genome-wide estimation of time-cumulative eQTLs for cis-MR, capturing the temporal effect of genetic variants on gene expression during cellular development that are missed when using eQTLs mapped from averaged expression of a singular time point.

Results

PCA-MR implementations outperform their generalized counterparts and PCA-IVW in simulations

We carried out three sets of simulation studies to validate PCA-MR extensions and compare them to generalized IVW (GIVW), generalized ML (GML), and the existing PCA approach PCA-IVW. Our first set of simulations aimed to analyze performance across various instrument strengths and low sample size. We simulated 10 valid IVs with an autoregressive correlation structure $(\rho \in \{0.2, 0.4, 0.6, 0.8, 0.9\})$ and target exposure F-statistic $(F \in \{5, 10, 25\})$, across true causal effects $(\theta \in \{0.01, 0.03, 0.05, 0.075, 0.1\})$. The sample size of the exposure and outcome were fixed at $\mathrm{n_X} = 250$ and $\mathrm{n_Y} = 50000$, respectively. PCA-MR outperformed GIVW, GML, and PCA-IVW in all scenarios. PCA-MR-IVW and ML had substantially higher power, coverage, and lower RMSE relative to PCA-IVW, GIVW, and GML (Figure 1 & 2). PCA-MR extensions exhibited controlled TIE under null association, while alternative methods were inflated above $\alpha = 0.05$. Due to the explicit incorporation of exposure association uncertainty into the likelihood, PCA-MR-ML exhibited better coverage than PCA-MR-IVW. PCA-IVW, GIVW, and GML experienced substantial weak instrument bias towards null association, while PCA-MR coverage remained robust.

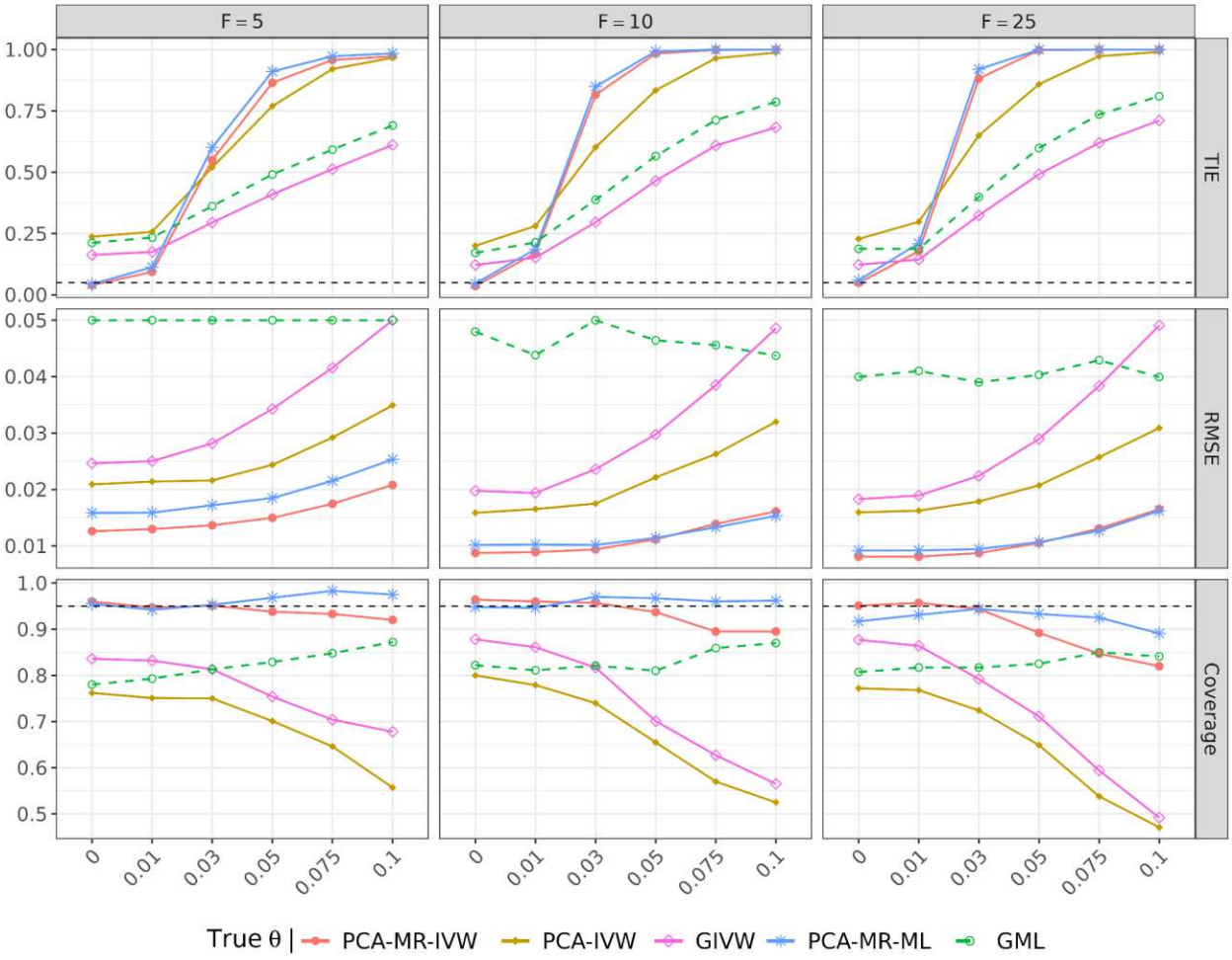


Figure 1. TIE ($\alpha = 0.05$), RMSE, and 95% confidence interval coverage across simulation replicates of PCA-MR-IVW and ML, PCA-IVW, GIVW, and GML for the first scenario regime with all weak-IVs in high LD ($\rho = 0.9$). For GIVW and GML, results are presented using an LD-pruning threshold of $r^2 < 0.8$. Results are shown across exposure instrument strengths $(F \in \{5, 10, 25\})$ as a function of the true θ. Dashed lines represent the nominal TIE rate (0.05) and nominal coverage (0.95), respectively.

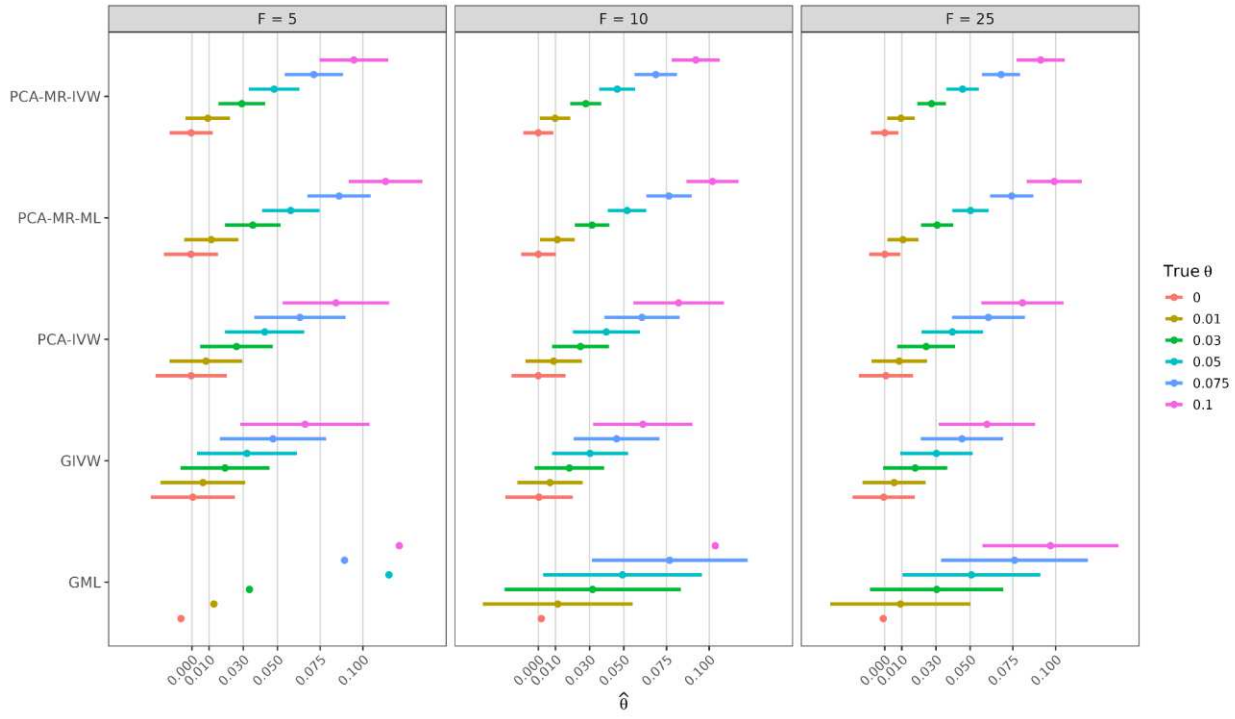


Figure 2. Mean $\hat{\theta}$ estimates and standard deviation intervals across replicates for the first scenario regime with weak-IVs (all valid) in high LD ($\rho = 0.9$). For GIVW and GML, results are presented using an LD-pruning threshold of $r^2 < 0.8$. Extreme standard deviation intervals were omitted, and extreme $\hat{\theta}$ estimates are denoted by "$*$".

Sensitivity analysis identifies heterogeneity due to uncorrelated and correlated horizontal pleiotropic IVs

In our second set of simulations we simulated 10 valid IVs and 4 invalid IVs ($\gamma^U \neq 0$), with the following pleiotropic effects due to uncorrelated horizontal pleiotropy $(\gamma^U \in \{0.01, 0.03, 0.05, 0.075, 0.1\})$ and target exposure F-statistic $(F \in \{5, 10, 25\})$. For all replicates, $\theta = 0$, the outcome sample size was fixed at $n_Y = 50000$, and the autoregressive parameter was fixed at $\rho = 0.9$. Cochran's Q test was able to identify heterogeneity due to uncorrelated horizontal pleiotropy across all simulations with $\gamma^U \neq 0$. Power increased with instrument strength $(F \in \{5, 10, 25\})$ and effectively plateaued from the lowest UHP effect

setting, $\gamma^U = 0.01$ onwards (Figure 3). The third set of scenarios simulated 10 valid IVs and 4 invalid IVs ($\gamma^C \neq 0$), with the following pleiotropic effects due to correlated horizontal pleiotropy $(\gamma^C \in \{0.01, 0.03, 0.05, 0.075, 0.1\})$ and instrument strengths $(F \in \{5, 10, 25\})$. For all replicates, $\theta = 0$, the outcome cohort sample size was fixed at $n_Y = 50000$, and the autoregressive parameter was fixed at $\rho = 0.9$. Cochran's Q test was also able to identify heterogeneity due to correlated horizontal pleiotropy ($\gamma^C \neq 0$), however the power to consistently do so required a much higher γ^C in comparison to γ^U. In these scenarios, heterogeneity was consistently identified from $\gamma^C = 0.03$ onwards, and power increased with instrument strength (Figure 3). Cochran's Q test for PCA-MR IVW and ML TIE under no UHP or CHP effect was controlled across all instrument strengths.

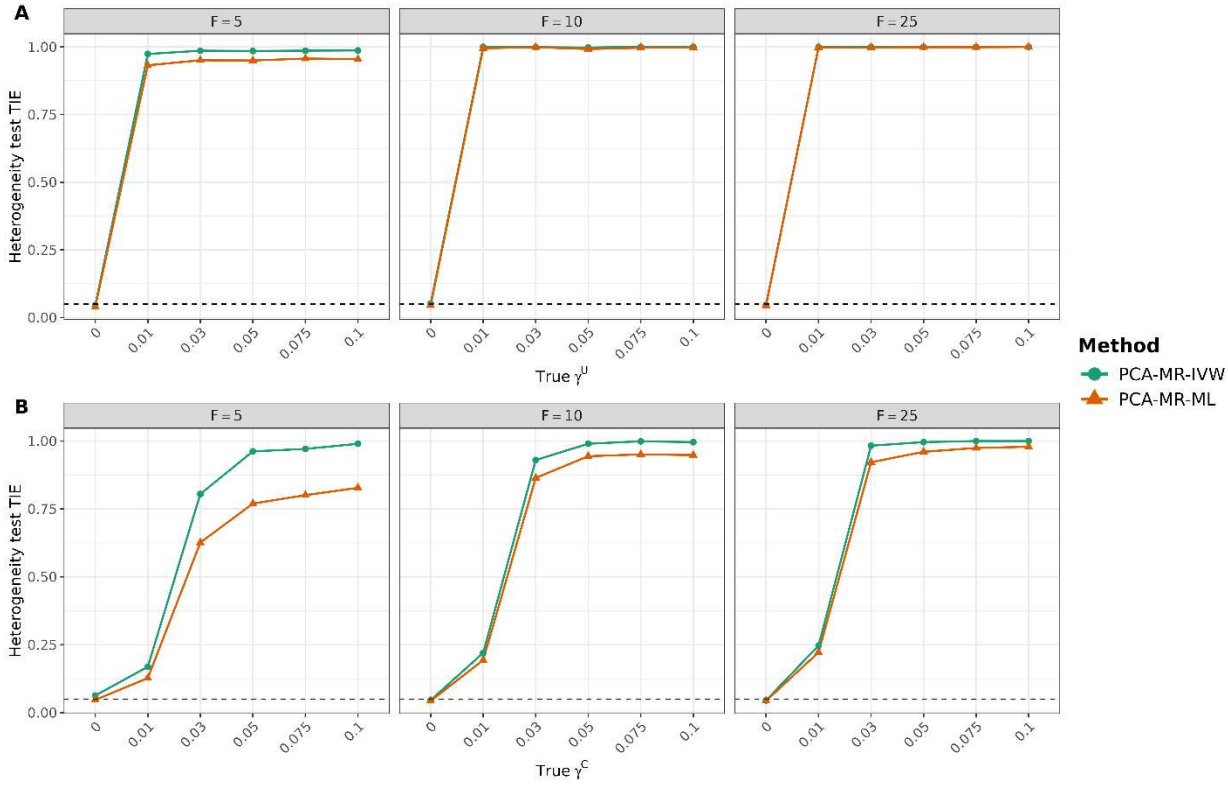


Figure 3. Sensitivity analysis simulations of PCA-MR-IVW and ML Cochran's Q test TIE ($\alpha = 0.05$). (A) Simulations with 10 valid IVs and 4 invalid IVs with an effect of uncorrelated horizontal pleiotropy (γ^U) as denoted on the x-axis. (B) Simulations with 10 valid IVs and 4 invalid IVs with an effect of correlated horizontal pleiotropy (γ^C) as denoted on the x-axis.

Real dataset application with sparse eQTL summary statistics

We applied all frameworks on two real datasets with low exposure sample sizes and sparse eQTL summary statistics. We used the averaged ($n_X = 161$) and time-cumulative gene expression ($n_X = 117$) of human iPSC derived serotonin neurons as exposure to major depressive disorder risk (See methods). PCA-MR extensions and PCA-IVW were applied without LD-pruning, while GIVW and GML were applied with an LD-pruning threshold of $r^2 < 0.8$. PCA-MR-IVW and ML truncated PCs to explain 99.9% of variance subject to $\kappa(\mathbf{R}) < 100$, and PCA-IVW truncated PCs to explain 99.9% of Ψ's variance. We selected candidate genes that had at least two eQTLs with p-value < 1e-4 for cis-MR. In the averaged expression dataset, there were 2879 genes that met selection criteria, with 2246 genes remaining after LD pruning. For the time-cumulative dataset, 1414 genes were selected and 1061 genes remained after LD-pruning. PCA-MR-IVW and ML increased genetic discovery (Bonferroni p-value correction) in both averaged and time-cumulative gene expression applications by over 500% in comparison to GIVW and GML, and by over 700% in comparison to PCA-IVW (Figure 4). Additionally, Cochran's Q identified excess heterogeneity amongst a subset of significant PCA-MR-IVW and ML hits in both dataset applications (Figure 4).

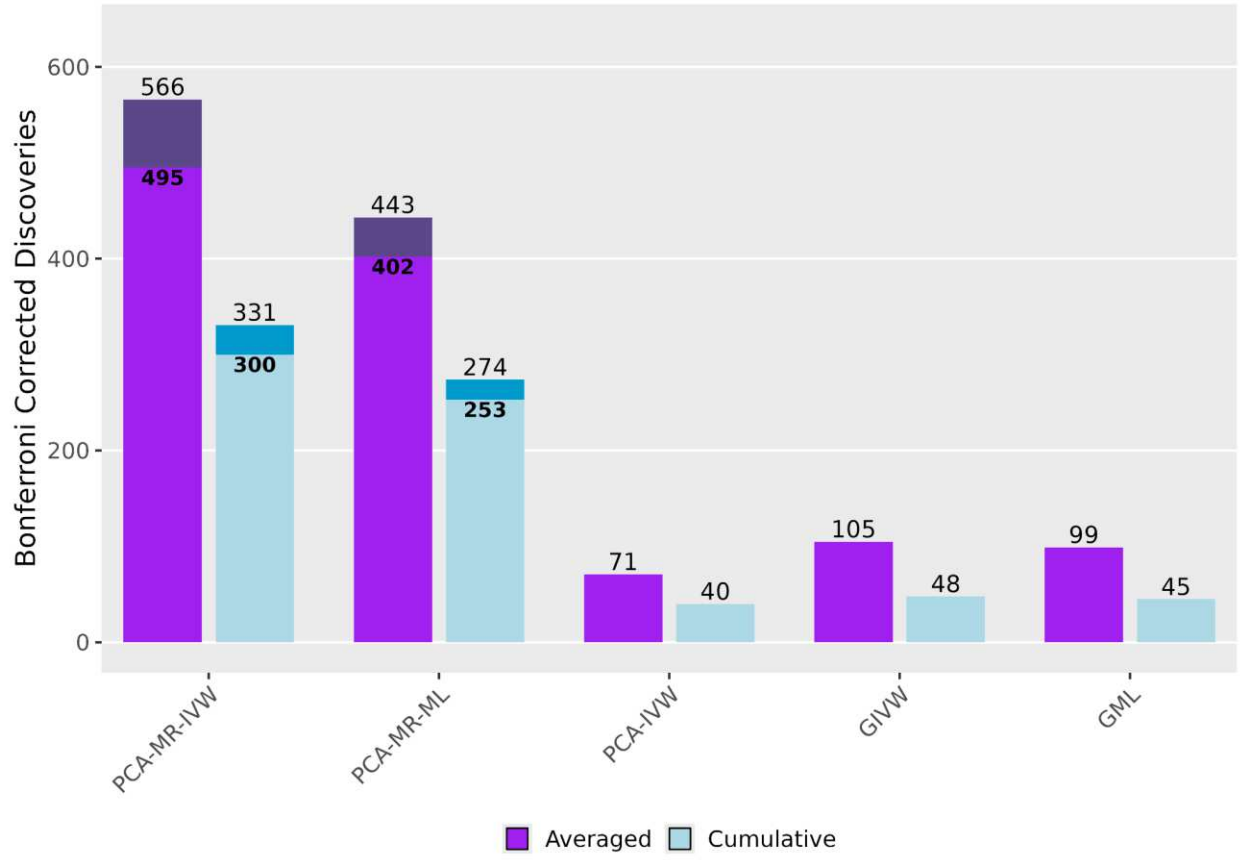


Figure 4. Count of significant genes identified after Bonferroni correction across frameworks in the MR with averaged and time-cumulative serotonergic gene expression and MDD. For GIVW and GML, results are presented using an LD-pruning threshold of $r^2 < 0.8$. The darker shade on PCA-MR IVW and ML bars denote the count of significant genes that were also significant for Cochran's Q test.

It is important to note that the true causal genes in serotonergic neurons on MDD are unknown. The applied dataset analysis aimed to assess genetic discovery and potential drug target prioritization using sparse exposures rather than provide ground truth validation. TIE validation of PCA-MR is provided by the first and second simulation studies. Across the frameworks, we evaluated a heuristic benchmark of MDD (outcome) enrichment amongst significant genes. This benchmark does not assess false positive rate or causality. Using the Open

Targets platform [9] we queried depressive disorder related association scores for significant genes identified across frameworks. The association score is a metric ranging from 0 to 1 that quantifies the strength of aggregated evidence (i.e. genetic associations, somatic mutations, pathway analyses, literature, animal models, and RNA expression), linking a specific target (i.e. gene or protein) to a disease. We calculated a standardized metric,

$OT_{MDD} = \frac{\mathrm{OT}}{N_{identifiable}} \times 1000$ that represents the summed MDD association score per 1000 identifiable genes with a non-null causal effect estimate. Intuitively, a larger $\mathrm{OT}_{\mathrm{MDD}}$ indicates that the framework captured more biological relevance amongst significant genes. PCA-MR IVW and ML substantially increased $\mathrm{OT}_{\mathrm{MDD}}$ and $\mathrm{OT}_{\mathrm{MDD}}\ \mathrm{Hits}$ when compared to PCA-IVW, GIVW, and GML (Table 1).

Averaged Expression ($n_X = 161$)				**Cumulative Expression ($n_X = 117$)**		
Method	$\mathrm{OT}_{\mathrm{MDD}}$	$\mathrm{OT}_{\mathrm{MDD}}\ \mathrm{Hits}$	$\mathrm{N}_{\mathrm{identifiable}}$	$\mathrm{OT}_{\mathrm{MDD}}$	$\mathrm{OT}_{\mathrm{MDD}}\ \mathrm{Hits}$	$\mathrm{N}_{\mathrm{identifiable}}$
PCA-MR-IVW	2.70	170	2477	4.13	111	1194
PCA-MR-ML	2.65	142	2159	3.74	88	1193
PCA-IVW	0.81	29	2739	0.40	16	1388
GIVW	0.65	33	2233	0.41	12	1058
GML	0.86	33	2233	0.40	12	1059

Table 1. Summary of external MDD relevance amongst significant genes identified across frameworks. $\mathrm{OT}_{\mathrm{MDD}}$ denotes the summed Open Targets association MDD association score per 1000 genes. $\mathrm{OT}_{\mathrm{MDD}}\ \mathrm{Hits}$ denotes the number of significant genes with non-zero MDD

association scores. $N_{identifiable}$ denotes the number of genes for which the respective framework returned a non-null causal estimate.

Empirical motivations for the alternative weight matrix

IVW and ML weight observations by the inverse covariance matrix of outcome and/or exposure associations, thus giving more influence to instruments with higher confidence. However, we empirically observed that the magnitude of genetic variance explained by each PC was approximately proportional to the magnitude of its projected exposure association (Figure 5). Theorem 1 provides the statistical theory that may underlie these observations (See Theorem 1). We were motivated by these factors to utilize an alternative precision matrix for IVW and ML that additionally weights instruments by the amount of genetic variance they explain, with Λ serving as a heuristic proxy for instrument relevance. In simulations, the alternative weight matrix utilized in PCA-MR-IVW and ML substantially improved inference relative to precision-based weighting alone (See supplementary results).

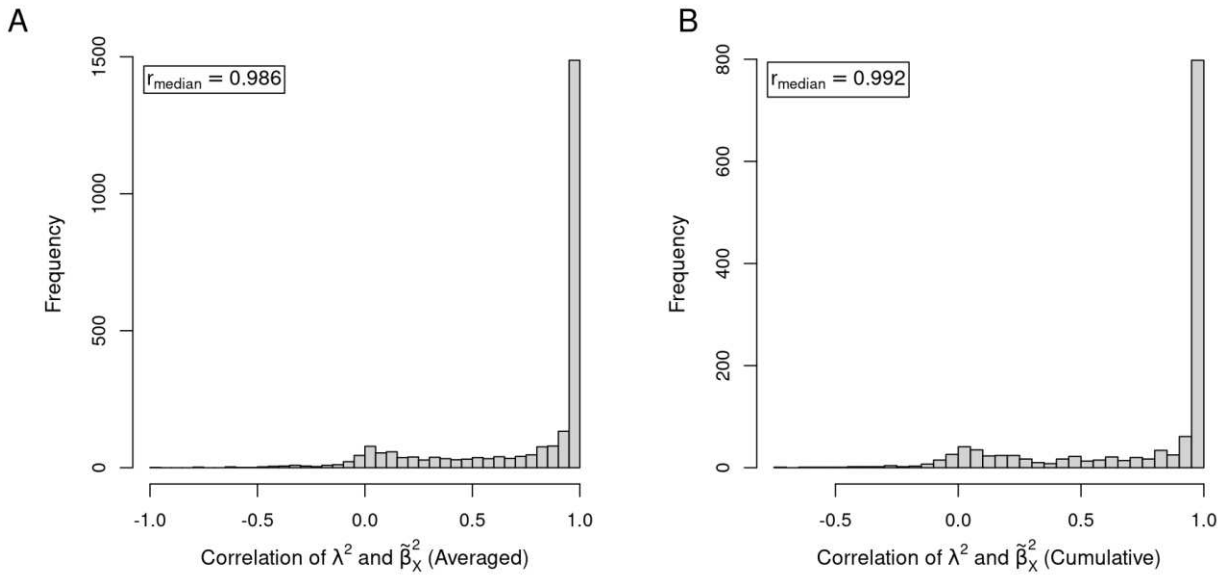


Figure 5. Histogram of the per-locus correlation of λ^2 and $\tilde{\beta}_X^2$ across loci with at least 3 PCs in the averaged (A) and time-cumulative (B) expression datasets.

Discussion

We propose PCA-MR-IVW and ML to infer the causal effect of biological exposures (i.e. serotonergic gene expression), on disease outcomes (i.e. major depressive disorder) using open sourced eQTL and GWAS summary statistics. The framework improves upon the previously proposed PCA approach (PCA-IVW), encompassing three distinctions that stand to increase power, precision, and numerical stability. First, we maximize the amount of genetic variance captured after dimensionality reduction by performing PCA on the genotype matrix. This is an advantageous distinction relative to performing PCA on the weighted correlation matrix of genetic variants, which encompasses an additional covariance transformation that incurs the loss of one degree of freedom from $\mathbf{R}$. Second, PCA-MR extensions utilize an alternative weight matrix that accounts for instrument relevance, incorporating the amount of genetic variance explained by PCs in addition to their exposure and/or outcome association variance. Theorem 1 provides the statistical theory rationale for this modeling choice, offering a more principled approach to encode the expected relevance of instruments relative to weighting the correlation matrix by exposure associations (See equation 2). Third, we enforce a $\kappa(\mathbf{R}) < 100$ truncation to promote a well-conditioned inversion of the covariance matrix of genetic associations, effectively preventing numerical instabilities in high LD scenarios. We implemented sensitivity analysis through Cochran's Q test to identify excess heterogeneity due to potential correlated and uncorrelated horizontal pleiotropy. In simulations, PCA-MR-IVW and ML exhibited controlled TIE, increased power and precision, and substantially reduced weak-instrument bias when compared to PCA-IVW, GIVW, and GML. We provide a user-friendly R package that includes

PCA-MR and its extensions (See data and code availability statement). The frameworks are light-weight and can generally perform cis-MR on 1000 loci in under 10 minutes using 1 core and 2GB of memory.

Our real-data set analysis performed cis-MR using the averaged and time-cumulative expression of human iPSC derived serotonergic gene expression as exposure to MDD risk. From the averaged dataset, we highlight 5 significant genes identified by PCA-MR-IVW with promising external relevance to MDD in Open Targets queries including *NPY1R* ($\hat{\theta}$ = 0.042, p-value = 1.2e-229), *SDK1* ($\hat{\theta}$ = -0.053, p-value = 8.6e-27), *GPX4* ($\hat{\theta}$ = -0.077, p-value = 2e-19), *MIF* ($\hat{\theta}$ = -0.01, p-value = 4.3e-6), and *RND2* ($\hat{\theta}$ = -0.032, p-value = 2.4e-157).

A significant proportion of the true effect of an allele on gene expression is temporal [10]. We modeled human iPSC derived serotonin neuron gene expression as a function of pseudotime to evaluate the causal effect of time-cumulative gene expression on major depressive disorder risk. The time-cumulative eQTL exposures represented dynamic changes in gene expression during cellular development [11]. We highlight 5 significant genes using PCA-MR-IVW with external relevance to MDD in Open Targets queries including *ESR2* ($\hat{\theta}$ = -0.0076, p-value = 4.3e-5), *CHGA* ($\hat{\theta}$ = -0.01, p-value = 7.8e-22), *PWWP2A* ($\hat{\theta}$ = -0.029, p-value = 2.8e-150), *KCNH1* ($\hat{\theta}$ = -0.01, p-value = 1.1e-10), *ERBB2* ($\hat{\theta}$ = 0.013, p-value = 2.3e-7).

There are several limitations of this study that are worth considering. First, the set of genes in serotonergic neurons that causally influence MDD risk are unknown, significant genes identified across frameworks cannot be validated against ground truth. Second, because individual-level data is typically not accessible for the exposure and outcome cohorts, we utilized the 1000G reference panel (N=503) to reference European genetic variation. External reference panels introduce additional variation in SNP correlation estimates, which PCA-MR models as

fixed. Although increasing the MAF quality control threshold to 5% (from the 1% threshold commonly employed) is expected to alleviate the problem of LD uncertainty, it does not necessarily eliminate it. Third, PCA-MR IVW and ML are high-powered alternatives to PCA-IVW, GIVW, and GML for genetic discovery and drug target prioritization in large scale cis-MR applications, however these methods assume all IVs are valid. Modeling both the uncorrelated and correlated pleiotropic effect of genetic variants when expressed as a linear combination is a challenging task and would likely require new statistical theory, we leave this for future work. We have proposed sensitivity analysis via Cochran's Q test, however it explicitly identifies excess heterogeneity amongst causal estimates and is not a direct test for horizontal pleiotropy.

Methods

Burgess et al. 2017 [8] proposed the utilization of principal components as instruments for MR. This concept assumes that any linear combination of valid instruments is itself a valid instrument. The method applies PCA to a weighted version of the SNP correlation matrix consisting of the following elements:

$$\Psi_{i,j} = \hat{\beta}_{X,i}\hat{\beta}_{X,j}\hat{\sigma}_{Y,i}^{-1}\hat{\sigma}_{Y,j}^{-1}\mathbf{R}_{i,j} \tag{2}$$

where $\mathbf{R}_{i,j} = \frac{1}{n-1}\mathbf{G}^T\mathbf{G}$ is the correlation of genetic variants *i* and *j*, and $\mathbf{G}$ is the standardized genotype matrix. After centering Ψ, PCA performs eigendecomposition such that $\Sigma_\Psi = W^T\Lambda W$ where $\Sigma_\Psi = \frac{1}{n-1}\Psi^T\Psi$ is the sampling covariance matrix of Ψ, W is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues. Variants with larger exposure association and precise outcome standard errors dominate the decomposition,

prioritizing dimensions with apparently stronger instrument relevance and outcome precision. Exposure and outcome summary statistics are projected onto these linear combinations of genetic variants and used as instrumental variables for IVW MR.

A robust dimensionality reduction approach for cis-Mendelian Randomization

We propose a novel dimensionality reduction approach called PCA-MR. PCA-MR performs PCA on the standardized genotype matrix, $\mathbf{G} \in \mathbb{R}^{n \times p}$ consisting of *n* individuals and *p* genetic variants. Eigendecomposition is performed such that $\mathbf{R} = W \Lambda W^T$ where $\mathbf{R} = \frac{1}{n-1}\mathbf{G}^T\mathbf{G}$ is the correlation matrix of genetic variants, W is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues. Performing PCA on $\mathbf{G}$, as opposed to Ψ, can be advantageous because it avoids the loss of one degree of freedom incurred by the additional covariance transformation. Decomposing $\mathbf{R}$ directly will typically identify a subspace with higher effective dimensionality, yielding more PCs (instruments) to perform inference with after quality control truncations. Because $\mathbf{R}$ is real and symmetric, its eigenvectors are orthogonal and the respective eigenvalues represent the amount of variation explained of $\mathbf{R}$. $\mathbf{R}$ is standardized, with each SNP having a variance of 1. Intuitively, an eigenvector with an eigenvalue of 10 can be interpreted as an orthogonal dimension that captures the variance of 10 independent SNPs. A sampling covariance matrix of exposure and outcome effects can be estimated by multiplying the outer product of standard errors of genetic variants *i* and *j* by their respective correlation.

$$\Sigma_X = SE_{X,i} \cdot SE_{X,j} \cdot \mathbf{R}_{i,j} \quad (3)$$
$$\Sigma_Y = SE_{Y,i} \cdot SE_{Y,j} \cdot \mathbf{R}_{i,j}$$

The transformed vectors of genetic associations are estimated by multiplying the vector of marginal genetic associations by the transposed matrix of eigenvectors .

$$\tilde{\beta_X} = W^T \beta_X \quad (4)$$
$$\tilde{\beta_Y} = W^T \beta_Y$$

The transformed sampling covariance matrices are estimated by pre- and post-multiplying the marginal sampling covariance matrices (Equation 5), by the matrix of eigenvectors:

$$\tilde{\Sigma}_X = W \Sigma_X W^T \quad (5)$$
$$\tilde{\Sigma}_Y = W \Sigma_Y W^T$$

It is important to note that principal components can still be valid even if their respective eigenvalues are below 1. Although these dimensions do not capture large axes of genetic variation, their validity remains determined by IV analysis assumptions. However, low variance components may increase susceptibility to numerical instabilities. To mitigate this, we sequentially enforce the following criteria to truncate PCs:

1. PCs explain (≤) 99.9% of the variance of $\mathbf{R}$
2. $\kappa(\mathbf{R}) < 100$

These quality controls are intended to ensure that the diagonalization of $\mathbf{R}$ is well-conditioned and effectively ensures that the covariance matrix of genetic association can reliably be inverted. The cumulative variance explained of $\mathbf{R}$ after the $\kappa < 100$ truncation will typically fall anywhere between 95-99%, capturing a substantial amount of genetic variance while preventing numerical instabilities in high LD regions.

Generalized Least Squares reparameterization

The principal components identified by PCA-MR can be used to carry out IVW with multiplicative random-effects. We considered the following linear model

$$\tilde{\beta}_{Y,i} = \hat{\theta}\tilde{\beta}_{X,i} + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0, \phi\tilde{\Sigma}_Y) \tag{6}$$

where $\hat{\theta}$ is the causal estimate of X on Y, and ϕ is the overdispersion factor (i.e., the ratio of weighted residual sum of squares to the degrees of freedom). MR with overdispersion assumes that the probability of pleiotropic effects being negative or positive is equal [7]. In the case of balanced horizontal pleiotropy, ϕ increases and corrects causal effect uncertainty. GLS weights observations by the precision matrix $\tilde{\Omega}^{-1}$, where $\tilde{\Omega}$ is the covariance matrix of genetic associations for the outcome. We employ an alternative weight matrix that additionally weights PCs by the amount of genetic variance explained. The choice was motivated by Theorem 1 and the empirical alignment of genetic variation and exposure association across PCs. We define the weight matrix as follows:

$$\tilde{W}_\Lambda = \Lambda\tilde{\Sigma}_Y^{-1} \tag{7}$$

where Λ is the diagonal matrix of truncated eigenvalues. The incorporation of Λ models the expectation of instrument relevance (See Theorem 1), weighting PCs relative to the amount of genetic variance they explain. When using correlated instruments, the IVW estimators can be characterized by the following equations [8].

$$\hat{\theta} = (\tilde{\beta}_X^T\tilde{W}_\Lambda\tilde{\beta}_X)^{-1}(\tilde{\beta}_X^T\tilde{W}_\Lambda\tilde{\beta}_Y) \tag{8}$$

$$\mathrm{Var}(\hat{\theta}) = \phi(\tilde{\beta}_\mathrm{X}^\mathrm{T}\tilde{\mathrm{W}}_\Lambda\tilde{\beta}_\mathrm{X})^{-1}$$

where $\phi = \dfrac{\mathrm{r}^\mathrm{T}\mathrm{W}_\Lambda\mathrm{r}}{\mathrm{df}}$.

Maximum–likelihood reparameterization

The principal components identified by PCA-MR can be extended to the maximum likelihood approach with random-effects [12,13]. We considered the following linear model:

$$\tilde{\beta}_{Y,i} = \hat{\theta}\tilde{\beta}_{X,i} + \epsilon_i,\ \epsilon_{\mathrm{i}} \sim \mathcal{MVN}(0, \phi \begin{pmatrix} \tilde{\Sigma}_X & 0 \\ 0 & \tilde{\Sigma}_Y \end{pmatrix}) \tag{9}$$

where $\phi = \sqrt{\frac{2\ell(\hat{\theta}, \mathrm{b_X})}{df}}$. Assuming there is no sample overlap (Which is typical of two-sample cis-MR cohorts), we have the likelihood up to some constants:

$$\ell(\hat{\theta}, \tilde{\mathrm{b}}_{\mathrm{X}} | \tilde{\beta}_Y, \tilde{\beta}_X, \tilde{\Sigma}_Y, \tilde{\Sigma}_X) = -\frac{1}{2}[(\tilde{\beta}_X - \tilde{\mathrm{b}}_X)^{\mathrm{T}} \mathrm{W_X} (\tilde{\beta}_X - \tilde{\mathrm{b}}_X) \tag{10}$$

$$+ (\tilde{\beta}_Y - \hat{\theta}\tilde{\mathrm{b}}_X)^{\mathrm{T}} \mathrm{W_Y} (\tilde{\beta}_Y - \hat{\theta}\tilde{\mathrm{b}}_{\mathrm{X}})]$$

where $\mathrm{W_X} = \Lambda\tilde{\Sigma}_{\mathrm{X}}^{-1}$ and $\mathrm{W_Y} = \Lambda\tilde{\Sigma}_{\mathrm{Y}}^{-1}$. The maximum likelihood estimator models the uncertainty of exposure genetic associations by jointly estimating $\tilde{\mathrm{b}}_{\mathrm{X}}$ and $\hat{\theta}$ [12]. Causal effect and precision estimates can be more robust in cases where the variance of exposure effects is larger when compared to IVW, which treat exposure effects as fixed. However, in instances of extreme multicollinearity the likelihood surface may flatten in certain directions and converge at a numerically unstable estimate. The $\kappa(\mathbf{R}) < 100$ truncation is expected to alleviate this problem, although not entirely. Conversely, IVW estimates a closed-form solution and will not experience such optimization problems. The variance of $\hat{\theta}$ can be obtained from the Hessian matrix:

$$\mathrm{Var}(\hat{\theta}) = \phi \mathrm{H}(\hat{\theta})^{-1} \tag{11}$$

where $\phi = \frac{2\ell(\hat{\theta}, \tilde{\mathrm{b}}_{\mathrm{X}})}{df}$ and $\mathrm{H}(\hat{\theta})^{-1}$ is the diagonal entry of the inverse Hessian matrix corresponding to $\hat{\theta}$.

Sensitivity Analysis

We implemented sensitivity analysis with Cochran's Q test [14], which identifies heterogeneity amongst selected IVs and their respective causal effect estimates, for PCA-MR IVW and ML extensions. The statistical test for heterogeneity estimates a Q-statistic as the weighted sum of squared residuals. The Q-statistic is assumed to follow a chi-squared distribution with $\mathrm{v}-1$ degrees of freedom under the null hypothesis of homogeneity.

$$\mathrm{Q} = (\tilde{\beta}_Y - \hat{\theta}\tilde{\beta}_X)^T \mathrm{W}_\Lambda (\tilde{\beta}_Y - \hat{\theta}\tilde{\beta}_X) \tag{12}$$

$$\mathrm{Q} \sim \chi^2_{\mathrm{v}-1}$$

Cochran's Q test is primarily designed to identify heterogeneity induced by uncorrelated horizontal pleiotropy under the InSIDE assumption, wherein the direct pleiotropic effect is independent of instrument strength. However, Cochran's Q can identify heterogeneity due to correlated horizontal pleiotropy at larger γ^C's (See results).

Simulations with autoregressive LD structure

We performed extensive simulations evaluating type I error, power, and sensitivity analysis in the presence of invalid IVs with uncorrelated and correlated pleiotropic effects. Two-sample summary statistics were simulated using the *simmrd* R package curated by Lorincz-Comi et al., 2024 [15]. *Simmrd* simulates two-sample summary statistics under the following model:

$$U = G^T\gamma^C + \epsilon_U \tag{13}$$

$$X = B^T G + \pi_X U + \epsilon_X$$

$$Y = X^T\theta + \pi_Y U + G^T\gamma^U + \epsilon_Y = \alpha^T G + \tilde{\epsilon}_Y$$

where U is a confounder exerting an effect on both the exposure X (π_X) and the outcome Y (π_Y), θ is the true causal effect of exposure on outcome, B is the true association of G and X, α is the true association of G and Y, γ^C is the effect of correlated horizontal pleiotropy, and γ^U is the

effect of uncorrelated horizontal pleiotropy. The simulations with the R package proceeded as follows:

1. Genotypes were generated from $g_j \sim \text{Binomial}(2, 0.3)$ and standardized to mean = 0 and unit variance.
2. The block-diagonal LD matrix $\mathbf{R}$ was generated where the correlations within blocks follow an autoregressive structure $\Sigma_{i,j} = \rho^{]i-j]}$, no correlation between blocks.
3. Genetic variant associations with the exposure were drawn from $\beta \sim \mathcal{N}(0, \mathbf{R})$ and rescaled such that the diagonals of $\beta^T\beta = \text{h}^2_{\text{cis}}$, where h^2_{cis} is the target exposure SNP heritability.
4. CHP effects for m^C genetic variants were drawn from $\gamma_{\text{j}}^{\text{C}} \sim \text{Uniform}(-1, 1)$ and scaled to explain the target amount of U's variance.
5. UHP effects for m^U genetic variants were drawn from $\gamma_{\text{j}}^{\text{U}} \sim \text{Uniform}(-1, 1)$ and scaled to explain the target amount of Y's variance.
6. The sign of π^{Y} was set to -1 and the sign of θ was set to sign of $\text{Cov}(\text{X}, \text{Y})$ and rescaled such that the $\text{Var}(\text{X}|\text{Y}|\text{U}) = 1$.
7. Perform OLS on the phenotypes Y and X (GWAS).
8. Rank genetic variants by increasing marginal exposure strength and minimize $\hat{i} = \arg\min_i |F_i - F_0|$ to obtain the cumulative subset of *i* SNPs with an aggregate F-statistic closest to the prespecified F_0.
9. Generate the reference LD matrix of selected IVs from $\text{Wishart}(\text{n}_{\text{ref}}, \mathbf{R})$ and standardize to obtain the correlation matrix $\hat{\mathbf{R}}$

Simulation scenarios

Our simulations aimed to mimic a cis-region of a molecular-omics study with low exposure and reference panel sample sizes, as is typical in cis-MR applications. The major parameters include the exposure sample size ($\mathrm{n_X}$), outcome sample size ($\mathrm{n_Y}$), reference panel sample size ($\mathrm{n_{ref}}$), number of causal genetic variants ($\mathrm{m_{causal}}$), number of genetic variants with UHP effects ($\mathrm{m_U}$), number of genetic variants with CHP effects ($\mathrm{m_C}$), causal effect (θ), effect of UHP (γ^{U}), effect of CHP (γ^{C}), exposure SNP heritability ($\mathrm{h^2_{cis}}$), and the number of LD blocks in the region. We simulated three scenario regimes: (1) all valid IVs; (2) invalid IVs with UHP effects; (3) invalid IVs with CHP effects. For all scenarios we fixed the following parameters: $\mathrm{n_Y} = 50000$, $\mathrm{n_{ref}} = 500$, $\mathrm{m_{causal}} = 10$, $\mathrm{h^2_{cis}} = 0.3$, $n_X = 250$, and the number of LD blocks = 2. For the first scenario with all valid IVs, we varied the causal effect $(\theta \in \{0.01, 0.03, 0.05, 0.075, 0.1\})$, autoregressive correlation parameter $(\rho \in \{0.2, 0.4, 0.6, 0.8, 0.9\})$, and the exposure F-statistics $(F \in \{5, 10, 25\})$. For the second scenario we included 10 valid IVs ($m_{causal} = 10$) with null causal effect ($\theta = 0$) and 4 invalid IVs ($\mathrm{m^U} = 4$) with UHP effects $(\gamma^{\mathrm{U}} \in \{0.01, 0.03, 0.05, 0.075, 0.1\})$. We fixed the correlation parameter of IVs at $\rho = 0.9$ and the exposure sample size at $n_X = 250$. For the third scenario we included 10 valid IVs ($m_{causal} = 10$) with null causal effect ($\theta = 0$) and 4 invalid IVs ($\mathrm{m^U} = 4$) with CHP effects $(\gamma^{\mathrm{C}} \in \{0.01, 0.03, 0.05, 0.075, 0.1\})$. We fixed the correlation parameter of IVs at $\rho = 0.9$ and the exposure sample size at $n_X = 250$. In total, we simulated 126 different scenarios with 1000 replicates each. Across all simulation studies, PCA-MR-IVW and PCA-MR-ML used PCs that explained 99.9% of the variance in the correlation matrix of genetic variants, subjected to $\kappa(\mathbf{R}) < 100$. PCA-IVW utilized PCs that explained 99.9% of variance, and generalized IVW and ML were applied with various LD-pruning thresholds $(\{r^2 \in 0.1, 0.4, 0.6, 0.8\})$.

Theorem 1

Suppose we have an exposure X and are performing a regression to estimate the association of J genetic variants within a cis-region, and we have the correlation matrix of genetic variants $\mathbf{R}$ calculated from the respective cohort. Let $\beta \in \mathbb{R}^J$ be the sparse vector of true joint effects with m causal SNPs having non-zero entries. After regression, the vector of marginal associations for J genetic variants is $b = \mathbf{R}\beta$. We can perform eigendecomposition such that $\mathbf{R} = W\Lambda W^T$. After projecting the marginal associations onto the transposed eigenvectors of $\mathbf{R}$, we have $\tilde{b} = W^T b = W^T\mathbf{R}\beta = W^T(W\Lambda W^T)\beta$ after substituting for b and $\mathbf{R}$. This can be rewritten as $\tilde{b} = (W^T W)\Lambda W^T\beta$ through the associative property. Because W is orthogonal, $W^T W = I$, thus we have $\tilde{b} = \Lambda W^T\beta$. We can define the vector of projected joint effects as $\tilde{\beta} = W^T\beta$. Thus, we have $\tilde{b} = \Lambda\tilde{\beta}$, and $\tilde{\beta} = \Lambda^{-1}\tilde{b}$ after rearranging. We can calculate the cis-heritability for the exposure as $h^2_{cis} = \tilde{\beta}^T\Lambda\tilde{\beta}$, and the cis-heritability explained by the i-th projected joint effect as $h^2_{cis,i} = \lambda_i\tilde{\beta}^2_i$. Although $\beta \in \mathbb{R}^p$ is sparse in practice, strong correlation structure that is typical of cis-regions often results in the distribution of joint effects across PCs. For presumptive purposes, we assume a reasonable exchangeable random-effects prior of the projected joint effects $\tilde{\beta} \sim \mathcal{N}(0, \sigma^2_\beta I)$. Then, $E[\tilde{\beta}^2_i] = \sigma^2_\beta$ and $E[h^2_{cis,i}] = \lambda_i\sigma^2_\beta$. Given these assumptions, we can assume that $E[h^2_{cis,i}] \propto \lambda_i$.

Application of PCA-MR in real datasets, constructing exposure eQTL summary statistics

We applied PCA-MR extensions in real dataset applications using the directional gene expression of human iPSC derived serotonin neurons as exposure to major depressive disorder risk. We evaluate discovery (Bonferroni p-value correction) using day 52 serotonin neuron (Averaged) eQTLs downloaded at https://doi.org/10.5281/zenodo.4333872 (N=161), and time-

cumulative serotonin neuron eQTLs (N=117) as exposure to major depressive disorder risk (Cases = 412,305, controls = 1,588,397). Major depressive disorder statistics were downloaded at https://pgc.unc.edu/for-researchers/download-results/ [16]. Both gene expression and GWAS cohorts are of European descent. We utilized 1000 Genomes Phase 3 to reference European genetic variation, which can be downloaded at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ [17].

Estimation of time-cumulative gene expression with FPCA

MR studies traditionally use averaged gene expression from a singular RNA-seq snapshot. The resulting eQTL associations are static, however a significant proportion of the true effect of an allele on gene expression is temporal [10]. To capture time-cumulative gene expression, we modeled the gene expression trajectory $X_i(t)$ of each individual i as a continuous function of pseudotime t:

$$X_i^K(t) = u(t) + \sum_{k=1}^{K} \xi_{i,k}\phi_k(t) \tag{14}$$

where $\mu(t)$ is the mean function, ϕ_k is the k-th eigenfunction identified by the spectral decomposition of the covariance surface $\mathrm{G(s,t)} = \mathrm{cov(X(s), X(t))}$, and $\xi_{i,k}$ is the k-th functional principal component score quantifying individual-specific variation relative to $\mu(t)$. In practice, longitudinal RNA-seq data is inherently sparse, irregular, and overdispersed. We used PACE, an algorithm specifically designed for such cases to estimate ϕ_k and $\xi_{i,k}$ [18]. We utilized the first K eigenfunctions that explained 95% of the variance in gene expression. Estimation of time-cumulative expression was obtained by integrating $X_i^K(t)$.

$$\tilde{X}_i = \int_t (\hat{X}_i(t) - \hat{\mu}(t))dt \tag{15}$$

for 18,416 protein coding genes.

The respective genotypes of all analyzed iPSC cell lines were obtained from the HipSci consortium [19]. We utilized LIMIX to map time-cumulative expression quantitative trait loci (eQTL) [20,21]. Variants (Minor allele frequency > 5%, missingness > 5%, and Hardy-Weinberg equilibrium p-value < 1e-06), within a cis-region spanning +/- 250kb upstream and downstream of hg19 gene coordinates were included. Gaussianization of gene expression was achieved by applying a rank-based inverse normal transformation. Association tests were carried out using a linear mixed model with genotype kinship as random effects and the first 15 principal components of time-cumulative expression as fixed covariates. eQTLs with p-value < 1e-4 were considered valid instruments for cis-MR.

Time-cumulative gene expression estimation

Jerber et al. 2021 [22] curated a protocol to differentiate human iPSCs lines (iPSCs) from the HipSci project (N=215) [23], towards a midbrain cell fate. From their study, RNA expression snapshots harvested at days 11, 30, and 52 were used to estimate cumulative gene expression. Preprocessed AnnData files containing normalized cell by gene count matrices were downloaded at https://doi.org/10.5281/zenodo.4333872. Utilizing all available snapshots, we consolidated non-proliferating floor plate progenitors, neuroblasts, and rotenone-untreated proliferating and non-proliferating serotonergic neurons. Quality control encompassed selecting cells with genes expressed > 200, mitochondrial reads > 10%, scaling, and batch correction using harmony [24,25]. A total of 117 iPSC cell lines with open source imputed and phased genotypes were present across all snapshots (N=117). We adapted trajectory inference and cumulative gene

expression pipelines from Sun et al. 2025 for our analysis [11]. Slingshot [26] was used to estimate per-cell pseudotime trajectories, and Principal analysis by conditional estimation (PACE) was used to estimate expression trajectories for each gene.

Availability of Data and Materials

Time-cumulative eQTL summary statistics can be accessed at https://doi.org/10.5281/zenodo.19672485. Summary statistics of simulation scenarios 1-4 can be accessed at https://doi.org/10.5281/zenodo.19873928. All code generated in study, and the *PCAMR* R package can be accessed at https://github.com/chrislaw133/PCA-MR. Limix eQTL source code can be accessed at https://github.com/single-cell-genetics/limix_qtl. Major depressive disorder GWAS summary statistics can be accessed at https://pgc.unc.edu/for-researchers/download-results/. D52 rotenone-untreated serotonin neuron eQTL summary statistics can be accessed at https://doi.org/10.5281/zenodo.4333872. 1000 Genomes reference panels can be accessed at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

Supplementary Methods and Information

Supplementary information and methods can be accessed at https://doi.org/10.5281/zenodo.19747473. Supplementary tables 1-7 can be accessed at https://doi.org/10.5281/zenodo.19686850.

Acknowledgements

Dr. Mike Rizzo[1]

Funding Declaration

The author of this manuscript received no funding for this work.

Competing Interests

The author declares no competing interests.

Consent for Publication

The author declares consent for publication.

Ethics approval and consent to participate

Not applicable

References

1. Zhang H. Pros and cons of Mendelian randomization. Fertil Steril [Internet]. 2023 [cited 2025 Nov 16];119:913–6. https://doi.org/10.1016/j.fertnstert.2023.03.029

2. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet [Internet]. 2014 [cited 2025 Nov 16];23:R89–98. https://doi.org/10.1093/hmg/ddu328

3. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27:1133–63. https://doi.org/10.1002/sim.3034

4. Wootton RE, Jones HJ, Sallis HM. Mendelian randomisation for psychiatry: how does it work, and what can it tell us? Mol Psychiatry [Internet]. Nature Publishing Group; 2022 [cited 2026 Mar 20];27:53–7. https://doi.org/10.1038/s41380-021-01173-3

5. Kivikoski M, Rastas P, Löytynoja A, Merilä J. Predicting recombination frequency from map distance. Heredity [Internet]. Nature Publishing Group; 2023 [cited 2026 Apr 19];130:114–21. https://doi.org/10.1038/s41437-022-00585-3

6. Zhou A, Karhunen V, Tian H, Pott J, Patel A, Slob EAW, et al. Variant selection to maximize variance explained in cis-Mendelian randomization. HGG Adv [Internet]. 2026 [cited 2026 Feb 27];7:100573. https://doi.org/10.1016/j.xhgg.2026.100573

7. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, et al. Guidelines for performing Mendelian randomization investigations: update for summer 2023. Wellcome Open Res [Internet]. 2023 [cited 2025 Nov 16];4:186. https://doi.org/10.12688/wellcomeopenres.15555.3

8. Burgess S, Zuber V, Valdes-Marquez E, Sun BB, Hopewell JC. Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. Genet Epidemiol [Internet]. 2017 [cited 2025 Oct 31];41:714–25. https://doi.org/10.1002/gepi.22077

9. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res [Internet]. 2023 [cited 2026 Apr 3];51:D1353–9. https://doi.org/10.1093/nar/gkac1046

10. Popp JM, Rhodes K, Jangi R, Li M, Barr K, Tayeb K, et al. Cell type and dynamic state govern genetic regulation of gene expression in heterogeneous differentiating cultures. Cell Genomics [Internet]. Elsevier; 2024 [cited 2025 Nov 14];4. https://doi.org/10.1016/j.xgen.2024.100701

11. Sun J, Dong Q, Wei J, Gao Y, Yu Z, Hu X, et al. ti-scMR: trajectory-inference-based dynamic single-cell Mendelian randomization identifies causal genes underlying phenotypic differences. NAR Genom Bioinform [Internet]. 2025 [cited 2025 Sept 23];7:lqaf082. https://doi.org/10.1093/nargab/lqaf082

12. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. Genet Epidemiol [Internet]. 2013 [cited 2026 Mar 25];37:658–65. https://doi.org/10.1002/gepi.21758

13. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol [Internet]. 2017 [cited 2025 Nov 6];46:1734–9. https://doi.org/10.1093/ije/dyx034

14. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. Epidemiology [Internet]. 2017 [cited 2025 Nov 9];28:30. https://doi.org/10.1097/EDE.0000000000000559

15. Lorincz-Comi N, Yang Y, Zhu X. simmrd: An open-source tool to perform simulations in Mendelian randomization. Genetic Epidemiology [Internet]. 2024 [cited 2025 Nov 6];48:59–73. https://doi.org/10.1002/gepi.22544

16. Adams MJ, Streit F, Meng X, Awasthi S, Adey BN, Choi KW, et al. Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. Cell [Internet]. Elsevier; 2025 [cited 2025 Nov 7];188:640-652.e9. https://doi.org/10.1016/j.cell.2024.12.002

17. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature [Internet]. Nature Publishing Group; 2015 [cited 2025 Sept 24];526:68–74. https://doi.org/10.1038/nature15393

18. Yao F, Müller H-G, Wang J-L. Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association [Internet]. ASA Website; 2005 [cited 2025 Nov 18];100:577–90. https://doi.org/10.1198/016214504000001745

19. Streeter I, Harrison PW, Faulconbridge A, Flicek P, Parkinson H, Clarke L. The human-induced pluripotent stem cell initiative—data resources for cellular genetics. Nucleic Acids Res [Internet]. 2017 [cited 2025 Sept 23];45:D691–7. https://doi.org/10.1093/nar/gkw928

20. Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nat Methods [Internet]. Nature Publishing Group; 2015 [cited 2025 Sept 23];12:755–8. https://doi.org/10.1038/nmeth.3439

21. Lippert C, Casale FP, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits [Internet]. bioRxiv; 2014 [cited 2025 Sept 23]. p. 003905. https://doi.org/10.1101/003905

22. Jerber J, Seaton DD, Cuomo ASE, Kumasaka N, Haldane J, Steer J, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nat Genet [Internet]. 2021 [cited 2025 Sept 23];53:304–12. https://doi.org/10.1038/s41588-021-00801-6

23. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature [Internet]. Nature Publishing Group; 2017 [cited 2025 Sept 23];546:370–5. https://doi.org/10.1038/nature22403

24. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods [Internet]. Nature Publishing Group; 2019 [cited 2025 Sept 23];16:1289–96. https://doi.org/10.1038/s41592-019-0619-0

25. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology [Internet]. 2020 [cited 2025 Sept 23];21:12. https://doi.org/10.1186/s13059-019-1850-9

26. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics [Internet]. 2018 [cited 2025 Sept 23];19:477. https://doi.org/10.1186/s12864-018-4772-0

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- PCAMRSupplementaryInformation.docx