# HORNET: Tools to find genes with causal evidence and their regulatory networks using eQTLs

Noah Lorincz-Comi[1], Yihe Yang[1], Jayakrishnan Ajayakumar[1], Makaela Mews[1], ... ...[1], William Bush William Bush[1] and Xiaofeng Zhu[1*]

[1*]Department of Population and Quantitative Health Sciences, Case Western Reserve University.

*Corresponding author(s). E-mail(s): xxz10@case.edu;

**Abstract**

Nearly two decades of genome-wide association studies (GWAS) have identify thousands of disease-associated genetic variants, but very few genes with evidence of causality. Recent methodological advances demonstrate that Mendelian Randomization (MR) using expression quantitative loci (eQTLs) as instrumental variables can detect causal genes. However, existing MR approaches are not well suited to handle the complexity of eQTL GWAS data and so they are subject to bias, inflation, and incorrect inference. We present HORNET, a comprehensive set of statistical and computational tools to perform genome-wide searches for causal genes using summary level GWAS data. HORNET is computationally efficient and robust to a range of real world conditions in which alternative methods and software are not. We present both a command line tool and desktop application to implement HORNET.

**Keywords:** expression quantitative trait loci, multivariable mendelian randomization, causal genes, schizophrenia

# 1 Background

Genetic epidemiologists have spent decades trying to identify genes that cause disease [1]. Significant effort has been given to experimental methods [2, 3], linkage studies [4], and functional annotation [5]. These methods of causal validation can be costly, time-consuming, and have sometimes producing conflicting results [6]. Equally, they generally cannot be scaled to support efficient testing of hundreds or thousands of genes simultaneously. Mendelian Randomization (MR) has been proposed as a cost- and time-efficient alternative to experimental testing that leverages the wealth of publicly available summary data from genome-wide association studies (GWAS) [7–10]. In this context, MR uses instrumental variables that are gene expression quantitative trait loci (eQTLs) to estimate tissue-specific causal effects of gene expression on disease risk [11]. This approach can either consider each gene separately (univariable MR) or jointly with surrounding genes in a regulatory network (multivariable MR). Since it is well known that many genes are members of large regulatory networks [12, 13], multivariable MR is supposedly robust to confounding that univariable MR is not [14–16].

However, there is currently no unified statistical or computational framework for applying multivariable MR to the study of causal genes. Performing multivariable MR with summary data from eQTL and disease GWAS (eQTL-MVMR) has many challenges, including the handling of missing data, linkage disequilibrium (LD) between eQTLs, gene tissue specification, gene prioritization, and causal inference. Without careful attention to each of these challenges, the simple application of traditional multivariable MR methods to these data may produce spurious results which may fail in follow-up experimental testing. We present HORNET, the first comprehensive set of bioinformatic tools that can be used to robustly perform eQTL-MVMR with GWAS summary data. We demonstrate that existing univariable and multivariable implementations of eQTL-MR are vulnerable to biases and/or inflated Type I and II error rates from weak eQTLs, correlated horizontal pleiotropy (CHP), high correlations between genes, missing data, and misspecified LD structure.

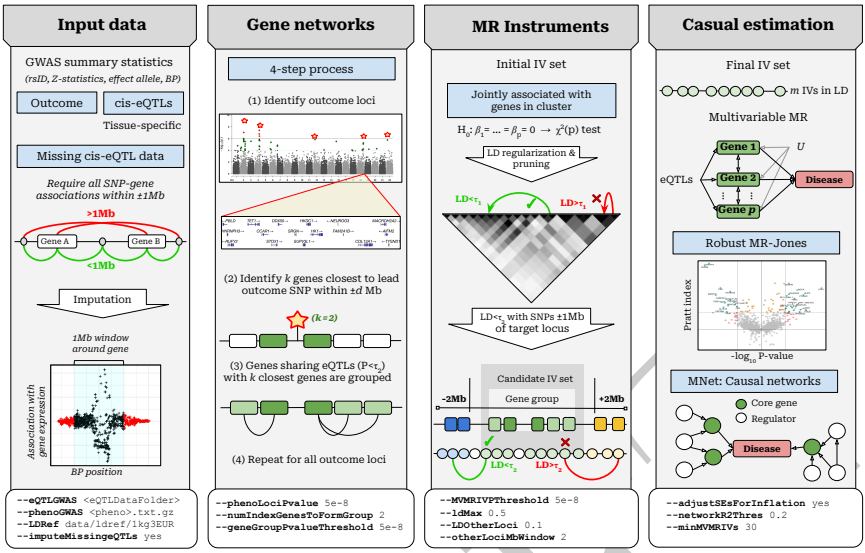


**Fig. 1** Flowchart illustrating genome-wide causal gene searches using HORNET. Example options given to flags that the command line version of HORNET uses are at the bottom of each panel. In the 'Input data' section, ±1Mb is used because it is standard in many publicly available data such as GTEx [17] and eQTLGen [18]. The HORNET software is available from https://github.com/noahlorinczcomi/HORNET

# 2 Methods

## 2.1 Data

HORNET uses summary level data from GWAS of gene expression (eQTL GWAS) and a disease phenotype. eQTL GWAS data should generally provide estimates of association between the expression of each gene and all SNPs within ±1Mb of them. These data are publicly available from consortia such as eQTLGen [19] and the Genotype-Tissue Expression (GTEx) project [17]. Disease GWAS data can typically be downloaded from public repositories such as the GWAS Catalog [20]. HORNET additionally requires an LD reference panel with corresponding `.bim`, `.bed`, and `.fam` files. The 1000 Genomes Phase 3 (1kg) [21] reference panel is automatically included with the HORNET software for African, East Asian, South Asian, European, Hispanic, and trans-ancestry populations, although researches may use their own reference panels such as UK Biobank [22].

## 2.2 IV selection and missing data

Selection of the IV set in eQTL-MVMR using standard methods can either reduce statistical power or make estimation of causal effects impossible. Univariable eQTL-MR for the $k$th gene in a locus of $p$ genes uses the set $\mathcal{S}_k$ of cis-eQTLs as instrumental variables and performs univariable regression [23].

Multivariable eQTL-MR in the same locus uses the superset $\mathcal{S}_\cup = \cup_{k=1}^p \mathcal{S}_k$ and performs multivariable regression [9]. Since most publicly available cis-eQTL data contain estimates of association between SNPs and all genes within ±1Mb of them (e.g., [17, 19]), not all SNPs in $\mathcal{S}_\cup$ may have associated estimates that are present in the data. An alternative approach is to use the set $\mathcal{S}_\cap = \cap_{k=1}^p \mathcal{S}_k$ which contains SNPs with association estimates that are available for all $p$ genes. However, this set may contain very few SNPs, if any, for some relatively large loci which contain many genes that are co-regulated. If the size of $\mathcal{S}_\cap$ is small, there can be limited statistical power for eQTL-MVMR [24]. There is currently no solution to this problem presented in the literature.

We propose a multivariate imputation procedure presented in Algorithm 1 based on matrix completion [25] that assumes a low-rank structure and accounts for GWAS estimation error and LD structure. As mentioned, public cis-eQTL summary data are generally available for SNP-gene pairs within ±1Mb of each other. Using individual-level data from (***), we demonstrate in Figure 7 of the **Supplement** that association estimates outside of the 1Mb window have mean 0 and constant variance with high probability. In simulation and real data, our imputation procedure imputes values with mean zero and increasing precision as distance from the gene center increases (Figure 2 Panel B, right). This procedure requires specification of a soft thresholding parameter $\lambda$, which is chosen across a grid of potential values as that which maximizes the normal likelihood of the fully imputed data.

**Algorithm 1** Pseudo-code of eQTL imputation.

**Require:** The $m \times p$ incomplete matrix of eQTL association estimates between $m$ SNPs and expressions of $p$ genes $\widehat{\mathbf{B}}$, the set of missing values $\mathcal{O}$, the singular values $\eta_1 \geq ..., \geq \eta_p$ of the $p \times p$ weak instrument bias matrix $m\boldsymbol{\Sigma}_{W_\beta W_\beta}$, inverse LD matrix $\boldsymbol{\Theta}$, tuning parameter $\lambda$, tolerance $\epsilon$.

1. Initialize $\widehat{\mathbf{B}}^0 = \boldsymbol{\Theta}^{1/2}\widehat{\mathbf{B}}$ with missing values set to 0
2. Define $d_1^0, ..., d_p^0$ as the singular values of $\widehat{\mathbf{B}}^0 := \mathbf{U}\mathbf{D}\mathbf{V}^\top$
3. Define $\alpha = 1 - \sum_{k=1}^p \eta_k / \sum_{k=1}^p d_k^0$
4. Reconstruct $\widehat{\mathbf{B}}^0 = \mathbf{U}(\alpha\mathbf{D})\mathbf{V}^\top$

**while do** $\|\widehat{\mathbf{B}}^{(t+1)} - \widehat{\mathbf{B}}^{(t)}\|_F > \epsilon$

Find $\mathbf{U}\mathbf{D}\mathbf{V}^\top = \widehat{\mathbf{B}}^{(t)}$ and define the $k$th singular value as $d_k^{(t)}$,

Threshold singular values, $d_k^{(t+1)} = (d_k^{(t)} - \lambda)_+$,

Construct $\widehat{\mathbf{B}}^{(t+1)} = \mathbf{U}\mathbf{D}^+\mathbf{V}^\top$, where $\mathbf{D}^+ = \text{diag}\left[d_k^{(t+1)}\right]_{k=1}^p$,

Set $\widehat{\mathbf{B}}_{/\mathcal{O}}^{(t+1)} = \widehat{\mathbf{B}}_{/\mathcal{O}}^{(0)}$; i.e., only missing values are imputed

**end while**

**Ensure:** Matrix $\boldsymbol{\Theta}^{-1/2}\widehat{\mathbf{B}}$ with no missing values.

After imputation of missing SNP-expression association estimates, the full set of candidate IVs $\mathcal{S}_\cup$ is restricted to those that are significant in a joint test

of association. Let $\widehat{\boldsymbol{\beta}}_j$ be the $p$-length vector of associations between the $j$th eQTL in $\mathcal{S}_\cup$ and the expression of $p$ genes in a tissue, where $\text{Cov}(\widehat{\boldsymbol{\beta}}_j) := \boldsymbol{\Sigma}$ is estimated using the methods in [16] (see **Supplement Section 3**). The initial candidate set $\mathcal{S}_\cup$ is restricted to

$$\mathcal{S} = \left\{ j : \widehat{\boldsymbol{\beta}}_j^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\beta}}_j > F^{-1}_{\chi^2(p)}(\alpha) \right\}, \tag{1}$$

where $\alpha = 5 \times 10^{-8}$ by default in the HORNET software. The set $\mathcal{S}$ is further restricted using LD pruning [26, 27] and CHP bias-correction as described in the next section.

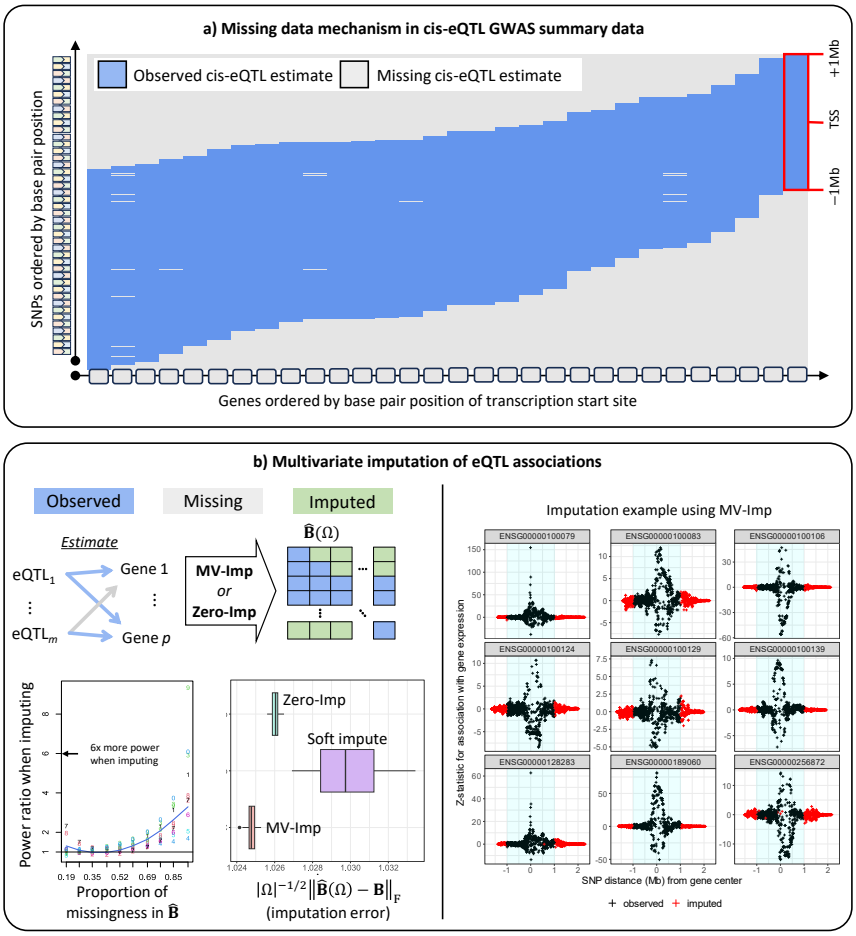


**Fig. 2** This figure illustrates the mechanism in summary cis-eQTL GWAS data that leads to missing data in eQTL-MVMR and how this missing data can be addressed using imputation. a) Only SNP-gene pairs within a defined distance have association estimates present in cis-eQTL summary data. This figure demonstrates this by displaying the available data for SNPs and genes ordered by their chromosomal position using data from the eQTLGen Consortium [19]. b) (left) Visual display of the pattern of missing in the design matrix $\hat{\mathbf{B}}(\Omega)$ used in eQTL-MVMR. Imputation can be performed by setting missing values to be 0 ('Zero-Imp') or by applying the low-rank approximation ('MV-Imp') to $\hat{\mathbf{B}}(\Omega)$ described in Algorithm 1. $|\Omega|$ is the total number of missing values. (right) An example of the MV-Imp method applied to summary data for 9 genes on chromosome 22 using cis-eQTL data from the eQTLGen Consortium [19].

## 2.3 LD

In nearly all applications of MVMR with eQTL data, an estimate of the LD matrix $\mathbf{R}$ for a set of eQTLs used as IVs is required. There are three primary challenges related to the use of eQTLs that are in LD when only individual-level data from a reference panel is available: (i) LD between loci can induce a correlated horizontal pleiotropy (CHP) bias (see **Supplement Section 2.1**),

(ii) imprecise estimates of LD between the eQTLs can inflate the statistics that are used to test the causal null hypothesis (**Supplement Sections 2.4** and **2.6**), (iii) direct application of the estimated LD matrix may be impossible because of non positive definiteness and the choice(s) of regularization [28] may not always be clear. In the next three subsections, we describe these challenges in greater detail and present the solutions that HORNET can implement.

### 2.3.1 CHP from LD between loci

CHP can be introduced in eQTL-MVMR if any eQTLs used as IVs in a target locus are in LD with other eQTLs that are in surrounding loci. This is a form of confounding that can inflate Type I or II error rates when testing the causal null hypothesis [29, 30]. We account for this CHP by removing IVs in the candidate set $\mathcal{S}$ that have LD $r^2 > \kappa$ with other SNPs not in this set but within ±2Mb of the boundaries of the locus. A visual example of this process is presented in Panel b of Figure 3. In practice, estimation of LD between eQTLs in different loci is efficiently made using the available reference panel. This process will reduce the number of eQTLs available for use in MVMR, but may provide partial protection against invalid inference.

### 2.3.2 Inflation from misspecified LD

Mis-specifying the LD matrix corresponding to a set of eQTLs that are used as IVs in eQTL-MR can inflate the statistics used to test the causal null hypothesis [31]. Since individual-level data for the discovery GWAS of the disease phenotype are rarely publicly available, eQTL-MR relies on publicly available reference panels to estimate LD between a set of eQTLs using populations assumed to be similar to the discovery population. This LD matrix can be mis-specified when a reference panel of relatively small size and/or different genetic ancestry is used, making causal inference using standard MR methods such as IVW [32] or principal components adjustment [33] vulnerable to inflated Type I/II error rates [31]. No solution to this problem currently exists for eQTL-MVMR. We propose a novel data-driven approach to correct for this inflation called IFC.

IFC estimates inflation in surrounding null loci to adjust for inflation in a target locus. 'Null' loci have no SNPs that are significantly associated either with the expressions of genes or the disease phenotype (i.e., all $P > 0.01$). These loci are also at least 5Mb away from known disease-associated loci, detected as those with $P < 5 \times 10^{-8}$, an are at least 1Mb away from the transcription start sites of other genes. In these loci, we identify a set of eQTLs to use as IVs for a single gene and perform univariable MR. We repeat this process for all genes meeting the above criteria within chromosome $C$ and calculate the genomic control inflation factor [34], denoted $\lambda_0^C$. We then return to causal estimation using eQTL-MVMR in a target locus of $p$ genes to find the causal estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_k)_{k=1}^p$ and corresponding standard error estimates $\widehat{\text{SE}}(\widehat{\theta}_1), ..., \widehat{\text{SE}}(\widehat{\theta}_p)$. The IFC correction is applied by the inflation-corrected

standard errors

$$\widetilde{\mathrm{SE}}(\widehat{\theta}_k) = \widehat{\mathrm{SE}}(\widehat{\theta}_k) \times \sqrt{\lambda_0} \tag{2}$$

for statistical inference. We demonstrate in **Supplement Sections 2.6.2** using real data from the eQTLGen Consortium [19] that inflation statistics $\lambda_0^C$ are stable across chromosomes and the genome. In these data, the mean inflation value across all chromosomes was 1.29, ranging from 0.89 to 1.91. Panel D of Figure 3 demonstrates that applying IFC and pruning [26, 27] controls the Type I error rate better than pruning alone across a range of scenarios in which the size of the reference panel and its concordance with the discovery data changes.

### 2.3.3 Non-positive definite LD matrix

When using a reference panel to estimate LD between a set of eQTLs that may be used as IVs in eQTL-MVMR, the raw estimate $\widehat{\mathbf{R}}$ may not be positive definite if the size of the reference panel $n_{\mathrm{ref}}$ is of the same order as the size of the IV set $m$ [35]. In this case, we cannot directly use $\widehat{R}$ because eQTL-MVMR requires its inverse, which may not exist. Multiple solutions to this problem exist in the literature, with methods either transforming the IV set [33, 36, 37] or directly applying regularization to $\widehat{\mathbf{R}}$ [38].

We propose a three step procedure to obtain a positive definite estimate of the LD matrix for a set of $m$ eQTLs: (i) prune absolute LD below the threshold $\kappa$ [26, 27], (ii) identify independent LD blocks in $\widehat{\mathbf{R}}$ using our novel data-driven algorithm, (iii) apply adaptive soft thresholding to each LD block [28, 39]. There is a tradeoff between the Type I error rate and power at different values for $\kappa$ (**Supplement Section 2.6.4-2.6.5**) and our software uses $\kappa = 0.3$ by default. Our algorithm for detecting independent LD blocks uses the $m-1$ determinants of the sequential subsets of $\widehat{\mathbf{R}}$, starting with the first 2 eQTLs, then the first 3, and so on. The differences between determinants from subsets of differing size one and ranked from smallest to largest and an adaptive procedure is applied to this list to find the optimal number of cutpoints. To provide some intuition, if the estimated LD matrix corresponding to SNPs in the set $\mathcal{S} = \{j : 1 \leq j \leq \ell\}$ has determinant $\pi$ and the determinant for SNPs in $\mathcal{S}_{+1} = \{j : 1 \leq j \leq \ell + 1\}$ has determinant $\pi_{+1}$, $\pi = \pi_{+1}$ implies that the one additional SNP in $\mathcal{S}_{+1}$ is uncorrelated with all SNPs in $\mathcal{S}$, which would define the position $\ell + 1$ as the index of a cutpoint in $\widehat{\mathbf{R}}$. See **Supplementary Section 2.5** for additional details. Finally, where any independent LD block is still not positive definite, we apply the method of [40] to achieve positive definitness with minimal perturbation.

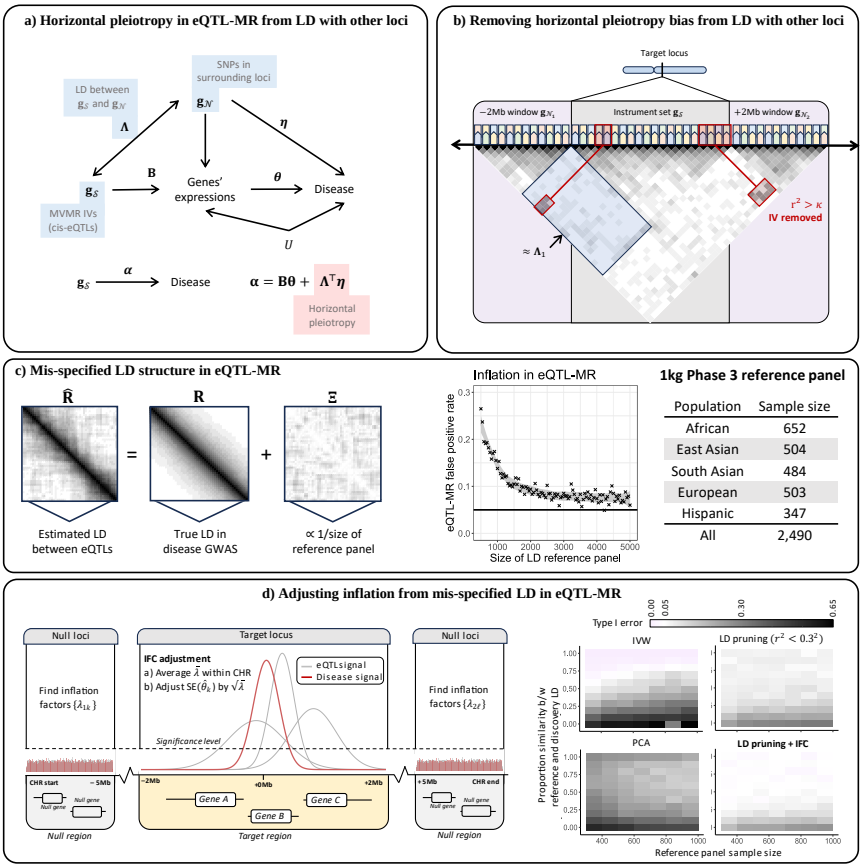


**Fig. 3** This figure illustrates the adjustments for CHP and inflation that are introduced when the eQTLs used in MR are in LD and researchers only have access to relatively small reference panels. a) The goal of eQTL-MVMR is to estimate $\boldsymbol{\theta}$, which may be subject to bias when $\boldsymbol{\Lambda}$ and $\boldsymbol{\eta}$ are each nonzero. b) This is the CHP-adjustment procedure described in Section 2.3.1. c) Results in the panel entitled 'Inflation in eQTL-MR' are from simulation in which the true LD matrix had dimension $500 \times 500$ and an AR1 structure with correlation parameter 0.5. We applied LD pruning at the threshold $r^2 < 0.3^2$. In this simulation, we repeatedly drew an estimate of the LD matrix from a Wishart distribution with degrees of freedom found on the x-axis. The R code used to perform this simulation is available at https://github.com/noahlorinczcomi/HORNET. d) Left is an illustration of our IFC inflation-correction method described in Section 2.3.2. Right presents the results of simulations in which Type I error was compared between different methods when the reference panel sample size and the similarity between the true LD in the reference panel and in the discovery GWAS changed. The full simulation details are described in **Supplementary Section 2.6.3**.

## 2.4 Causal inference

We rely on the convergence of three sources of evidence when prioritizing potentially causal genes: (i) gene selection and hypothesis testing using IFC inflation correction, (ii) Pratt index values [41], (iii) causal network structure [42]. Regarding (i), the HORNET software uses MR with joint selection of causal genes and pleiotropy (MR-Jones; (***)) to prioritize genes in a locus

and MR with unbiased estimating equations (MRBEE) [16] to estimate their causal effects without bias from weak IVs and horizontal pleiotropy. We additionally apply the IFC correction described in section 2.3.2 to make MRBEE more robust to inflation. Regarding (ii), the Pratt index is calculated for each gene in a locus and is used to estimate the portion of explained genetic variance in the outcome that is attributable to each gene conditional on the others in the locus [41]. This index is equal to the product of the multivariable and univariable MRBEE estimates. Genes with the largest Pratt index values are those with the strongest evidence of causality. Regarding (iii), HORNET implements the MNet (***) method to estimate the networks of regulatory relationships between genes in a locus and their pathways of causal effect on the disease phenotype. The MNet method can be used to identify core genes, which are those whose expression directly causes changes in disease risk, and peripheral genes, which are those that only cause changes in disease risk by regulating the expression of core genes [42–44]. Prioritized genes are those with evidence of direct causality on the disease phenotype and regulation by peripheral genes.

## 2.5 Computation

HORNET requires GWAS summary statistics for gene expression and a disease phenotype and an LD reference panel. LD estimation from a reference panel for a set of eQTLs is made using the PLINK software [45], which requires the presence of `.bim`, `.bed`, and `.fam` files. eQTL GWAS data must contain a single file for each chromosome and generally should contain summary statistics for all genotyped SNPs within a cis-region of each available gene. These data are available for blood tissue from the eQTLGen Consortium (n=31k) [19] and the GTEx consortium for 53 other tissues (n<706) [17]. To help researchers identify relevant tissues to select in their analyses, we provide a tissue prioritizing tool based on the heritability of eQTL signals. This tool receives a list of target genes from the researcher and returns a ranked list of tissues in which each target gene has the strongest eQTLs using GTEx v8 summary data [17]. See **Supplement Section 4** for additional details and a demonstration of how to use this tool.

The HORNET suite of tools exists as both a command line tool and a desktop program for Linux, Windows, and Mac machines. Both tools have detailed tutorials located at https://github.com/noahlorinczcomi/HORNET and are introduced briefly in **Supplement Section 5**. By downloading HORNET, users also receive PLINK v1.9 [45] and LD reference panels for European, African, East and South Asian, Hispanic, and trans-ethnic populations from 1000 Genomes Phase 3 (1kg) [21]. By default, our software uses this reference panel from the entire 1kg sample to estimate LD in the discovery population, but users can alternatively specify a specific sub-population in 1kg or even use their own LD reference panels.

# 3 Simulations

We performed three separate simulations to assess the performance of missing data imputation, inflation in eQTL-MR, and inflation-correction methods. The setup of each simulation and a discussion of the results they produced are described in the next three subsections.

## 3.1 Imputing missing data

In the missing data simulation, we used summary statistics from eQTL GWAS for 9 genes on chromosome 1 produced from 236 non-Hispanic White individuals in the (***) cohort (***). We restricted the eQTLs used to only those within ±2Mb of the transcription start site (TSS) of one of the genes, producing 526 fully observed eQTLs. We then set the Z-statistics for eQTL-gene pairs in which the eQTL was >1Mb from the TSS as missing and evaluated three methods of imputation: (i) MV-Imp, which was our multivariate imputation method outlined in Algorithm 1, (ii) imputation of missing values with 0s, (iii) and soft impute [25]. For each simulation, the true LD correlation matrix $\mathbf{R}$ between the 526 eQTLs had a first order autoregressive structure with correlation parameter 0.5. The matrix of measurement error correlations $\mathbf{\Sigma}_{W_\beta W_\beta}$ was estimated from all SNPs in the 1Mb window with squared Z-statistics for all eQTL associations less than the 95th quantile of a chi-square distribution with one degree of freedom. This follows the procedures used in practice [16, 46].

In simulation, our multivariate imputation method outlined in Algorithm 1 has smaller estimation error than imputation with all zero values or the traditional soft impute method [25]. Estimation error in this setting is defined as the difference between true and imputed values. Since there is currently no other way to address missing data in eQTL-MVMR, zero-imputation and soft impute are two straightforward alternatives to our proposed imputation approach. We demonstrate in Section 1.4 of the Supplement that imputing missing data using our algorithm can produce up to 2-4x increases in power vs excluding eQTLs with any missing associations as IVs.

## 3.2 Inflation in eQTL-MR

In the simulation to demonstrate inflation in eQTL-MR, the true LD matrix $\mathbf{R}$ for 500 eQTLs had a first order autoregressive structure with correlation parameter 0.50 and was estimated by sampling from a Wishart distribution with varying degrees of freedom equal to the reference panel sample size. In each simulation, true eQTL and disease standardized effect sizes were drawn from independent multivariate normal distributions with means 0 and covariance matrices $\mathbf{R}$. We then applied LD pruning [26, 27] at the threshold $r^2 < 0.3^2$ to restrict the IV set used in univariable MR. We performed MR using IVW [32] and the Type I error rate was recorded.

Panel C in Figure 3 demonstrates that LD reference panels that contain genotype information for less than 5,000 individuals can inflate the false

positive rate in eQTL-MVMR. When the reference panel contained 500 individuals, the false positive rate approached 0.30. As a comparison, the largest population-stratified sample of individuals in the 1000 Genomes Phase 3 reference sample [21] is 652 and the smallest is 347.

### 3.3 Correcting inflation from misspecified LD

In the simulation evaluating the performance of inflation-correction methods, we used an LD matrix denoted $\mathbf{R}$ that was estimated for 168 SNPs using 413k non-related European individuals in the UK Biobank [22] and LD pruned at the threshold $r^2 < 0.85^2$. We let $\mathbf{R}$ be be the true LD matrix in the disease GWAS, applied a perturbation to $\mathbf{R}$ denoted $\widetilde{\mathbf{R}}$, then drew the working LD matrix, denoted $\widehat{\mathbf{R}}$, from a Wishart distribution with $n_{\text{ref}}$ degrees of freedom and parameter $\widetilde{\mathbf{R}}$. The parameter $n_{\text{ref}}$ represented the size of the LD reference panel and ranged from 300 to 1000; linear perturbations to $\mathbf{R}$ were applied in the following way: $\widetilde{\mathbf{R}} = \xi\mathbf{R} + (1-\xi)\mathbf{I}$ where $\xi \in \{0.0, 0.1, ..., 0.9\}$. We additionally used eQTL estimates for these SNPs and 7 genes from the eQTLGen Consortium [19] to estimate the genetic correlation matrix to use in simulations, denoted $\mathbf{S}$. At each iteration of the simulation, we drew eQTL Z-statistics from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{S} \otimes \mathbf{R}$ and outcome Z-statistics from a mean-zero normal distribution with covariance matrix $\mathbf{R}$. The working LD covariance matrix used in MR was $\widehat{\mathbf{R}} \sim \text{Wishart}(n_{\text{ref}}, \widetilde{\mathbf{R}})$ and we evaluated the Type I error of the following MR methods: IVW [32], LD pruning at $r^2 < 0.3^2$ [26, 37], principal components at the variance explained threshold of 0.99 [33], and our IFC method (see Section 2.3.2) with LD pruning at $r^2 < 0.3^2$

Simulation results presented in Panel D of Figure 3 demonstrate that Type I error is inflated for all corrective methods when the true LD matrix in the reference panel is sparser than the true LD matrix in the discovery population. When these two true LD matrices are the same, only our LD pruning plus IFC correction preserves the Type I error rate at its nominal 0.05 level when the size of the LD reference panel is not exceedingly large. The IVW [32] method has deflated Type I error (i.e., $< 0.05$) that only disappears as the size of the LD reference panel approaches infinity. PCA [33] and LD pruning [37] methods each have inflated Type I error. For example, when the reference and discovery populations have the same true LD structure and the reference panel contains 1000 individuals, the PCA method has a Type I error rate of 0.21. In the same scenario, pruning and IFC correction has Type I error rate of 0.05, whereas pruning alone had a Type I error rate of 0.08.

## 4 Real data analysis with schizophrenia

We applied the HORNET methods and software to the analysis of genes whose expression in basal ganglia, cerebellum, cortex, hippocampus, amygdala, and blood tissues cause schizophrenia risk. Schizophrenia GWAS data were from [47], which included 130k European individuals and were primarily from the

Psychiatric Genomics Consortium (PGC) core data set. eQTL GWAS data in brain tissue were from [48], which contained GWAS data from European samples of sizes 208 for basal ganglia, 492 for cerebellum, 2,683 for cortex, 168 for hippocampus, and 86 for amygdala tissue. eQTL GWAS data in blood were from the eQTLGen Consortium [19] for 31k predominantly European individuals. We performed analyses with HORNET in all schizophrenia loci with at least one P-value less than 0.005, grouped genes sharing eQTLs with P-values less than 0.001, applied LD pruning at the threshold $r^2 < 0.7^2$, and removed SNPs in LD with any IVs in the target locus beyond $r^2 > 0.5^2$ in a 1Mb window. Finally, all IVs had a P-value for joint association less than 0.005 in the joint test of Equation 1. We performed HORNET in each tissue separately and present the results in Figure 4.

Figure 4 uses the data described above to provide examples of the primary results produced by genome-wide analysis with HORNET, including causal estimates for prioritized genes, genome-wide genetic variance explained and Pratt index values for each tissue, and an estimated regulatory and causal network. These results suggest that for many loci, the genetic variance in schizophrenia is almost entirely explained by the causal effects of gene expression in select tissues, but that large differences across tissues exist (Panel c). For example, only 17.2% of genetic variation in schizophrenia in the *KCTD13* locus is explained by the expression of genes in blood tissue, compared to 75.2% in the cerebellum and 59.4% in the cortex. In this locus, we observed that expression of the *INO80E* gene in the cortex increased schizophrenia risk ($P = 2.1 \times 10^{-9}$), but that the specific schizophrenia variation attributable to this effect was small (Pratt index=0.09). Alternatively, expression of the *DOC2A* gene in the cortex was strongly associated with increased schizophrenia risk ($P < 10^{-50}$) and also had a relatively large Pratt index value of 0.67 (Panels b and d), suggesting that *DOC2A* is potentially a better gene target than *INO80E* in the cortex.

We attempted to better understand the complex regulatory network that exists in the human leukocyte antigen (HLA) complex of 6p21.33 [49]. Genetic variants in this region are highly associated with risk of schizophrenia [50, 50–52] and many other traits such as brain morphology [53], autism spectrum disorder [54], and Type II diabetes [55]. The HORNET software applied MNet (***) to uncover regulatory relationships between 18 genes in this locus and their pathways of causal effect on schizophrenia risk when expressed in cerebellum tissue. These results suggest a densely connected gene regulatory network in which the *HLA-C* gene is a so-called 'regulatory hub' [56, 57]. The *HLA-C* gene is directly associated with the regulation of 8 other genes and is indirectly associated with the regulation of all genes in the locus except *OR2J3*. Only *HLA-C* and *FLOT1* have direct causal effects on schizophrenia risk, and all other 15 peripheral genes (*OR2J3* excluded) have causal effects on schizophrenia that only are mediated by *FLOT1* and/or *HLA-C* expression.

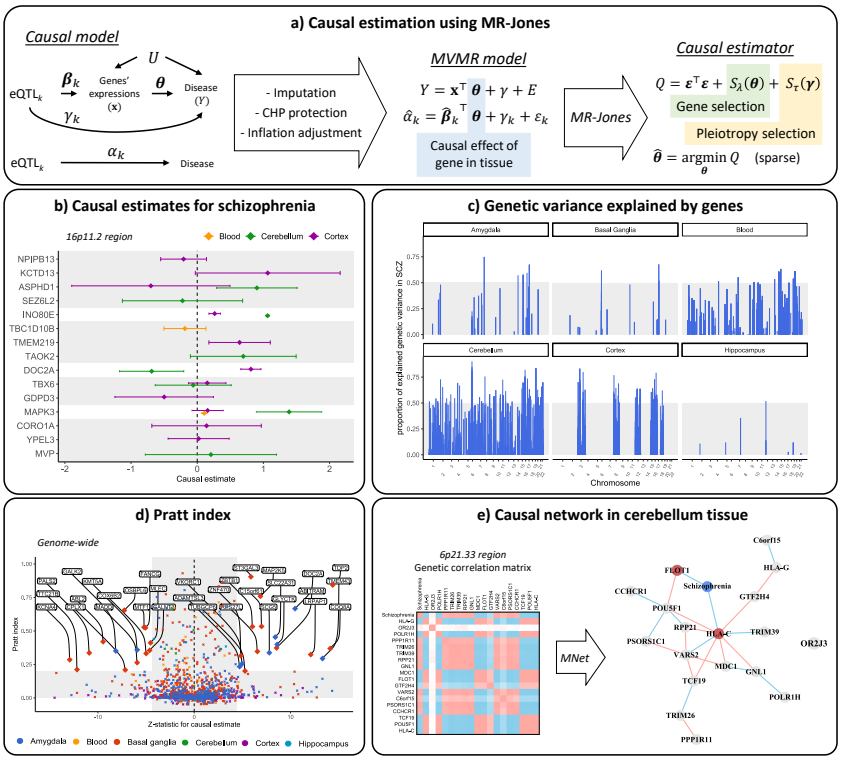


**Fig. 4** This figure presents the results of using HORNET to search for genes modifying schizophrenia risk when expressed in different tissues. a) Description of the MR-Jones causal model, MVMR model, and estimator. b) Causal estimates for multiple genes in blood, cerebellum, and cortex tissues in the schizophrenia-associated *KCTD13* locus. c) Genetic variance in schizophrenia explained by MVMR models fitted using MR-Jones across the genome and six tissues. Areas in which no variance explained values exist either had no genes prioritized by MR-Jones or had insufficient eQTL signals to perform MVMR. d) Pratt index values for all causal estimates made for all tissues. Pratt index values outside the range of (-0.1,1) are not shown. This may happen because of large variability in univariable MR estimates for some loci. e) Estimated gene regulatory and schizophrenia causal network for 18 genes in the schizophrenia-associated *FLOT1* locus of the HLA complex.



# 5 Conclusion

Existing methods for finding causal genes using multivariable Mendelian Randomization (MR) with GWAS summary statistics are generally vulnerable to bias and inflation from missing data, misspecified LD structure, and confounding by other genes. Equally, no flexible and comprehensive set of computational tools to robustly perform this task current exists. We introduced a suite of statistical and computational tools in the HORNET software that addresses these common challenges in multivariable MR using eQTL GWAS data. HORNET can generally provide unbiased causal estimation and robust inference across a range of real-world conditions in which existing methods in alternative software packages may not. HORNET can be downloaded as a command line

tool and/or desktop application from https://github.com/noahlorinczcomi/HORNET, where users will also find detailed tutorials demonstrating how to use HORNET.

# References

[1] F. Hormozdiari, G. Kichaev, W.-Y. Yang, B. Pasaniuc, and E. Eskin, "Identification of causal genes for complex traits," *Bioinformatics*, vol. 31, no. 12, pp. i206–i213, 2015.

[2] D. Rees and J. Alcolado, "Animal models of diabetes mellitus," *Diabetic medicine*, vol. 22, no. 4, pp. 359–370, 2005.

[3] L. M. Tai, K. L. Youmans, L. Jungbauer, C. Yu, M. J. LaDu, *et al.*, "Introducing human apoe into a$\beta$ transgenic mouse models," *International journal of Alzheimer's disease*, vol. 2011, 2011.

[4] J. Ott, J. Wang, and S. M. Leal, "Genetic linkage analysis in the age of whole-genome sequencing," *Nature Reviews Genetics*, vol. 16, no. 5, pp. 275–284, 2015.

[5] K. Wang, M. Li, and H. Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic acids research*, vol. 38, no. 16, pp. e164–e164, 2010.

[6] C. T. Lewandowski, J. M. Weng, and M. J. LaDu, "Alzheimer's disease pathology in apoe transgenic mouse models: the who, what, when, where, why, and how," *Neurobiology of disease*, vol. 139, p. 104811, 2020.

[7] K. J. Gleason, F. Yang, and L. S. Chen, "A robust two-sample transcriptome-wide mendelian randomization method integrating gwas with multi-tissue eqtl summary statistics," *Genetic epidemiology*, vol. 45, no. 4, pp. 353–371, 2021.

[8] A. Zhu, N. Matoba, E. P. Wilson, A. L. Tapia, Y. Li, J. G. Ibrahim, J. L. Stein, and M. I. Love, "Mrlocus: Identifying causal genes mediating a trait through bayesian estimation of allelic heterogeneity," *PLoS genetics*, vol. 17, no. 4, p. e1009455, 2021.

[9] E. Porcu, S. Rüeger, K. Lepik, F. A. Santoni, A. Reymond, and Z. Kutalik, "Mendelian randomization integrating gwas and eqtl data reveals genetic determinants of complex and clinical traits," *Nature communications*, vol. 10, no. 1, p. 3300, 2019.

[10] A. van Der Graaf, A. Claringbould, A. Rimbert, B. C. H. B. T. . H. P. A. . van Meurs Joyce BJ 10 Jansen Rick 11 Franke Lude 1 2, H.-J. Westra, Y. Li, C. Wijmenga, and S. Sanna, "Mendelian randomization while

jointly modeling cis genetics identifies causal relationships between gene expression and lipids," *Nature communications*, vol. 11, no. 1, p. 4930, 2020.

[11] D. Gill, M. K. Georgakis, V. M. Walker, A. F. Schmidt, A. Gkatzionis, D. F. Freitag, C. Finan, A. D. Hingorani, J. M. Howson, S. Burgess, *et al.*, "Mendelian randomization for studying the effects of perturbing drug targets," *Wellcome open research*, vol. 6, 2021.

[12] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks," *Frontiers in cell and developmental biology*, vol. 2, p. 38, 2014.

[13] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature reviews Molecular cell biology*, vol. 9, no. 10, pp. 770–780, 2008.

[14] E. Sanderson, "Multivariable mendelian randomization and mediation," *Cold Spring Harbor perspectives in medicine*, p. a038984, 2020.

[15] Z. Lin, H. Xue, and W. Pan, "Robust multivariable mendelian randomization based on constrained maximum likelihood," *The American Journal of Human Genetics*, vol. 110, no. 4, pp. 592–605, 2023.

[16] N. Lorincz-Comi, Y. Yang, G. Li, and X. Zhu, "Mrbee: A novel bias-corrected multivariable mendelian randomization method," *bioRxiv*, pp. 2023–01, 2023.

[17] G. Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, *et al.*, "The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, 2015.

[18] U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, *et al.*, "Large-scale cis- and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression," *Nature genetics*, vol. 53, no. 9, pp. 1300–1310, 2021.

[19] U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Kasela, *et al.*, "Unraveling the polygenic architecture of complex traits using blood eqtl metaanalysis," *BioRxiv*, p. 447367, 2018.

[20] E. Sollis, A. Mosaku, A. Abid, A. Buniello, M. Cerezo, L. Gil, T. Groza, O. Güneş, P. Hall, J. Hayhurst, *et al.*, "The nhgri-ebi gwas catalog: knowledgebase and deposition resource," *Nucleic acids research*, vol. 51, no. D1, pp. D977–D985, 2023.

[21] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[22] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, p. e1001779, 2015.

[23] A. Gkatzionis, S. Burgess, and P. J. Newcombe, "Statistical methods for cis-mendelian randomization with two-sample summary-level data," *Genetic epidemiology*, vol. 47, no. 1, pp. 3–25, 2023.

[24] N. Lorincz-Comi, Y. Yang, G. Li, and X. Zhu, "Mrbee: A novel bias-corrected multivariable mendelian randomization method," *biorxiv, 523480*, 2023.

[25] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[26] F. Dudbridge and P. J. Newcombe, "Accuracy of gene scores when pruning markers by linkage disequilibrium," *Human heredity*, vol. 80, no. 4, pp. 178–186, 2016.

[27] A. F. Schmidt, C. Finan, M. Gordillo-Marañón, F. W. Asselbergs, D. F. Freitag, R. S. Patel, B. Tyl, S. Chopade, R. Faraway, M. Zwierzyna, *et al.*, "Genetic drug target validation using mendelian randomisation," *Nature communications*, vol. 11, no. 1, p. 3255, 2020.

[28] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Stat.*, vol. 36, no. 1, pp. 199–227, 2008.

[29] J. Morrison, N. Knoblauch, J. H. Marcus, M. Stephens, and X. He, "Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics," *Nature genetics*, vol. 52, no. 7, pp. 740–747, 2020.

[30] M. Verbanck, C.-Y. Chen, B. Neale, and R. Do, "Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases," *Nature genetics*, vol. 50, no. 5, pp. 693–698, 2018.

[31] L. Jiang, L. Miao, G. Yi, X. Li, C. Xue, M. J. Li, H. Huang, and M. Li, "Powerful and robust inference of complex phenotypes' causal genes with dependent expression quantitative loci by a median-based mendelian randomization," *The American Journal of Human Genetics*, vol. 109, no. 5, pp. 838–856, 2022.

[32] S. Burgess and J. Bowden, "Integrating summarized data from multiple genetic variants in mendelian randomization: bias and coverage properties of inverse-variance weighted methods," *arXiv preprint arXiv:1512.04486*, 2015.

[33] S. Burgess, V. Zuber, E. Valdes-Marquez, B. B. Sun, and J. C. Hopewell, "Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables," *Genetic epidemiology*, vol. 41, no. 8, pp. 714–725, 2017.

[34] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

[35] A. Gkatzionis, S. Burgess, and P. J. Newcombe, "Statistical methods for cis-mendelian randomization," *arXiv e-prints*, pp. arXiv–2101, 2021.

[36] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, G. I. of ANthropometric Traits (GIANT) Consortium, D. G. Replication, M. analysis (DIAGRAM) Consortium, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, "Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits," *Nature genetics*, vol. 44, no. 4, pp. 369–375, 2012.

[37] P. J. Newcombe, D. V. Conti, and S. Richardson, "Jam: a scalable bayesian framework for joint analysis of marginal snp effects," *Genetic epidemiology*, vol. 40, no. 3, pp. 188–201, 2016.

[38] Q. Cheng, X. Zhang, L. S. Chen, and J. Liu, "Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology," *Nat. Commun.*, vol. 13, no. 1, pp. 1–13, 2022.

[39] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 672–684, 2011.

[40] Y.-G. Choi, J. Lim, A. Roy, and J. Park, "Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage," *Journal of Multivariate Analysis*, vol. 171, pp. 234–249, 2019.

[41] H. Aschard, "A perspective on interaction effects in genetic association studies," *Genetic epidemiology*, vol. 40, no. 8, pp. 678–688, 2016.

[42] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: from polygenic to omnigenic," *Cell*, vol. 169, no. 7, pp. 1177–1186, 2017.

[43] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi, "Structure and dynamics of core/periphery networks," *Journal of Complex Networks*, vol. 1, no. 2, pp. 93–123, 2013.

[44] I. Mathieson, "The omnigenic model and polygenic prediction of complex traits," *The American Journal of Human Genetics*, vol. 108, no. 9, pp. 1558–1563, 2021.

[45] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "Plink: a tool set for whole-genome association and population-based linkage analyses," *The American journal of human genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[46] X. Zhu, T. Feng, B. O. Tayo, J. Liang, J. H. Young, N. Franceschini, J. A. Smith, L. R. Yanek, Y. V. Sun, T. L. Edwards, *et al.*, "Meta-analysis of correlated traits via summary statistics from gwass with an application in hypertension," *Am. J. Hum. Genet.*, vol. 96, no. 1, pp. 21–36, 2015.

[47] V. Trubetskoy, A. F. Pardiñas, T. Qi, G. Panagiotaropoulou, S. Awasthi, T. B. Bigdeli, J. Bryois, C.-Y. Chen, C. A. Dennison, L. S. Hall, *et al.*, "Mapping genomic loci implicates genes and synaptic biology in schizophrenia," *Nature*, vol. 604, no. 7906, pp. 502–508, 2022.

[48] N. de Klein, E. A. Tsai, M. Vochteloo, D. Baird, Y. Huang, C.-Y. Chen, S. van Dam, R. Oelen, P. Deelen, O. B. Bakker, *et al.*, "Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases," *Nature genetics*, vol. 55, no. 3, pp. 377–388, 2023.

[49] J. Klein and A. Sato, "The hla system," *New England journal of medicine*, vol. 343, no. 10, pp. 702–709, 2000.

[50] M. Ikeda, A. Takahashi, Y. Kamatani, Y. Momozawa, T. Saito, K. Kondo, A. Shimasaki, K. Kawase, T. Sakusabe, Y. Iwayama, *et al.*, "Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect," *Schizophrenia bulletin*, vol. 45, no. 4, pp. 824–834, 2019.

[51] F. S. Goes, J. McGrath, D. Avramopoulos, P. Wolyniec, M. Pirooznia, I. Ruczinski, G. Nestadt, E. E. Kenny, V. Vacic, I. Peters, *et al.*, "Genome-wide association study of schizophrenia in ashkenazi jews," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 168, no. 8, pp. 649–659, 2015.

[52] S. Consortium, "Genome-wide association study identifies five new schizophrenia loci," *Nat Genet*, vol. 43, no. 10, pp. 969–976, 2011.

[53] M.-H. Chen, L. M. Raffield, A. Mousas, S. Sakaue, J. E. Huffman, A. Moscati, B. Trivedi, T. Jiang, P. Akbari, D. Vuckovic, *et al.*, "Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations," *Cell*, vol. 182, no. 5, pp. 1198–1213, 2020.

[54] "Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia," *Molecular autism*, vol. 8, pp. 1–17, 2017.

[55] M. Vujkovic, J. M. Keaton, J. A. Lynch, D. R. Miller, J. Zhou, C. Tcheandjieu, J. E. Huffman, T. L. Assimes, K. Lorenz, X. Zhu, *et al.*, "Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis," *Nature genetics*, vol. 52, no. 7, pp. 680–691, 2020.

[56] W. Deng, K. Zhang, S. Liu, P. X. Zhao, S. Xu, and H. Wei, "Jrmgrn: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions," *Bioinformatics*, vol. 34, no. 20, pp. 3470–3478, 2018.

[57] D. Yu, J. Lim, X. Wang, F. Liang, and G. Xiao, "Enhanced construction of gene regulatory networks using hub gene information," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–20, 2017.

# Supplement to 'HORNET: Tools to find genes with causal evidence and their regulatory networks using eQTLs'

Noah Lorincz-Comi[1], Yihe Yang[1], Jayakrishnan Ajayakumar[1], Makaela Mews[1], ... ...[1], William Bush[1] and Xiaofeng Zhu[1*]

[1*]Department of Population and Quantitative Health Sciences, Case Western Reserve University.

*Corresponding author(s). E-mail(s): xxz10@case.edu;



## Contents

# 1 Missing data

## 1.1 Demonstration of missingness

As mentioned in the main text, the set of instrumental variables (IVs) used in multiple exposure Mendelian Randomization (MVMR) is the union set of exposure-specific IV sets. In summary data from cis-eQTL GWAS in which each exposure is the expression of gene, not all SNPs are tested for an association with each gene. Generally, only SNPs within ±1Mb of a gene are tested for an association with the expression of that gene. This means that the union set may contain at least some SNPs for which there is no estimate of association between them and each gene in a locus under study. Visual representations of this are displayed in Figures 1 and 2. To avoid introducing missing data by using the union set of gene-specific IV sets in MVMR, one may consider using the intersection set of gene-specific IV sets, guaranteeing no missing data. However, for a locus containing a moderately large number of genes (e.g., ~10 or more), the intersection set may actually be of very small size or even empty. This could respectively introduce a $p > n$ scenario or even prevent MVMR from performed.

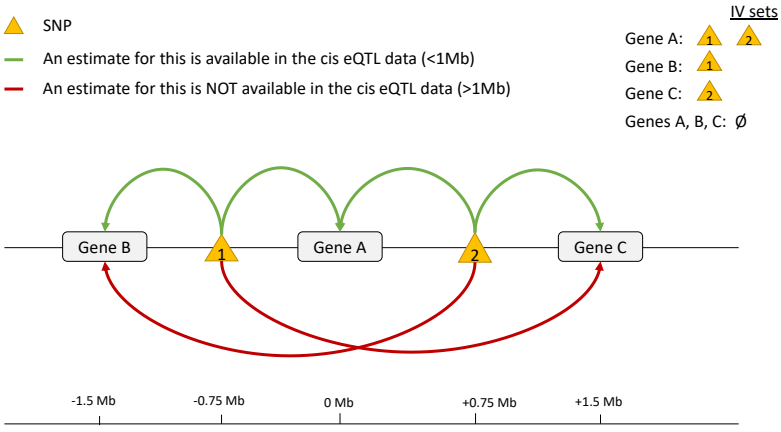


**Fig. 1** This is an example representation of the data that is available in the summary-level eQTLGen [1] and GTEx [2] cis-eQTL public data sets. Only associations between SNPs and the expression of genes within ±1Mb of those SNPs have estimates included in the available data. Since in multivariable MR, we select as the IV set the union of gene-specific IV sets, this union set may contain no SNPS with association estimates observed for all genes in a group. Put another way, the intersection of all gene-specific IV sets that is restricted only to SNPs with no non-missing values may be empty.

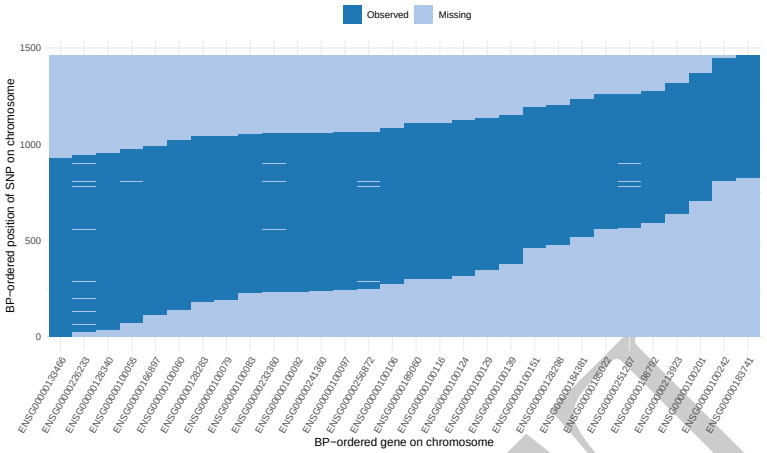


**Fig. 2** Determinations of missing values for each SNP (in order by BP position on y-axis) that was used in multivariable MR with the 30 genes ordered by BP position on CHR 22 (x-axis), as an example. These genes were grouped using the procedure described in **Methods**. It was stated in **Methods** that the nature of the *cis*-eQTL data is such that only SNPs within +-1Mb of a gene center have estimates of association with the gene expression available in the data. In our analyses, we included multiple genes in causal estimation. Denote the set of SNPs used as IVs in multivariable MR for a group of genes of size $p_k$ as $\mathcal{M}_k$, which is the union set $\cup_{\ell=1}^{p_k}\mathcal{M}_k^{\ell}$ of the gene-specific IV sets $\mathcal{M}_k^1 ... \mathcal{M}_k^{p_k}$. This union set is the set of SNPs whose ordered positions are on the y-axis. As Figure 1 demonstrated, restricting this union set to only non-missing association estimates between each SNP and gene expressions may make the set empty. In the figure above, this scenario would correspond to being unable to draw any horizontal line through the plot such that the line never touches a 'Missing' area.

| Tissue | Minimum % missing eQTL associations across genes in a locus | | Maximum % missing eQTL associations across genes in a locus | | Locus size (Mb) | | |
|---|---|---|---|---|---|---|---|
| | Minimum across all loci | Mean across all loci | Mean across all loci | Maximum across all loci | Minimum | Maximum | Mean |
| Basal ganglia | 0.00 | 0.11 | 0.16 | 0.60 | 1.04 | 3.32 | 1.81 |
| Blood | 0.00 | 0.20 | 0.31 | 0.72 | 0.26 | 3.89 | 1.95 |
| Cerebellum | 0.00 | 0.06 | 0.10 | 0.58 | 0.35 | 3.25 | 1.58 |
| Coronary artery | 0.00 | 0.09 | 0.14 | 0.58 | 0.64 | 3.05 | 1.82 |
| Cortex | 0.00 | 0.05 | 0.08 | 0.59 | 0.33 | 3.87 | 1.57 |
| Hippocampus | 0.00 | 0.12 | 0.17 | 0.45 | 0.82 | 2.91 | 1.71 |
| Lung | 0.00 | 0.08 | 0.12 | 0.55 | 0.51 | 3.27 | 1.59 |
| Spinal cord | 0.00 | 0.06 | 0.09 | 0.49 | 1.23 | 2.61 | 1.80 |

**Fig. 3** This figure presents a high-level summary of the rates of missing eQTL associations in gene groups formed while applying HORNET to the study of schizophrenia (see main text). Values in the first four columns after tissue type correspond to missingness rates; values in the final three columns correspond to the sizes, from the smallest base pair position of an eQTL used as an IV, to the largest, of loci analyzed by HORNET. Missingness rates are first aggregated from the gene level to the locus level, then again from the locus level to the genome level. For example, the '0.00' value in the first row and second column indicates that the smallest rate of missingness observed for any gene that was analyzed by HORNET in basal ganglia tissue was 0.00, the next column indicates the mean rate of missingness across all loci analyzed by HORNET in the same tissue, and so on. eQTL GWAS data for basal ganglia, cerebellum, cortex, hippocampus, and spinal cord tissues were from [3]; coronary artery and lung tissue data were from [4]; blood tissue data were from [1]. The complete set of commands given to HORNET to perform these analyses is available at https://github.com/noahlorinczcomi/HORNET/real_data.

## 1.2 Support from cis-eQTLs in a larger window

Since most publicly available summary data from cis-eQTL GWAS contain association estimates between SNPs and the expression of genes within ±1Mb of each them, we wanted to better understand the pattern of association between gene expression and SNPs >1Mb away. For this, we used individual-level data from (***). We estimated associations between gene expression and all available SNPs within ±5Mb using the TensorQTL pipeline [5]. Displayed in Figure 4 are these association estimates for 25 genes on chromsome 1 that had eQTLs with corresponding P-values less than $5 \times 10^{-8}$. These results demonstrate that, on average, the most significant eQTL signals are near the transcription start site and that significant eQTLs are unlikely to be observed outside of a 1Mb window but within 5Mb.

## 1.3 Multivariate Imputation

### 1.3.1 Procedure

In this subsection, we describe the procedure that we used to impute missing data in the set $\cup_{\ell=1}^{p}\mathcal{M}^{\ell}$ that is the union of $p$ gene-specific IV sets each denoted as $\mathcal{M}^{\ell}$. Our imputation method is similar to the soft imputation method using matrix completion [6] but corrects for measurement error in the eQTL GWAS and accounts for LD structure. 'Measurement error in the eQTL GWAS' here

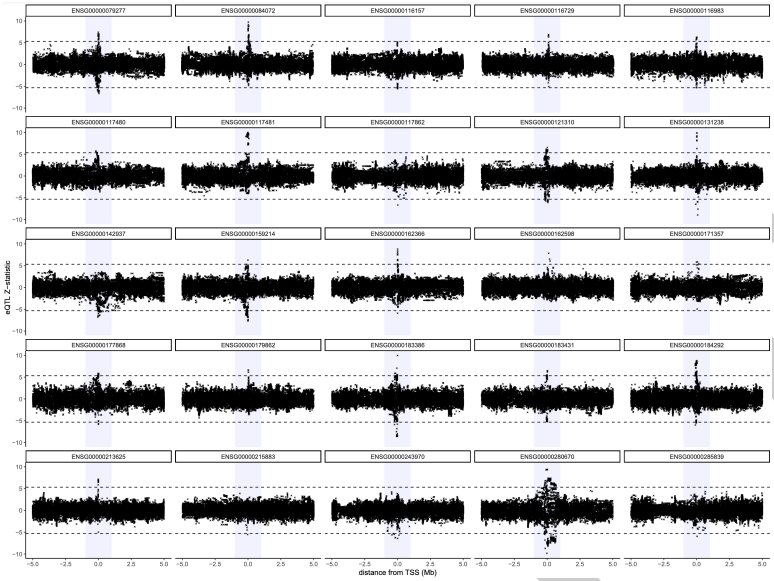


**Fig. 4** Displayed are Z-statistics for eQTLs that are within ±5Mb around 25 genes on chromosome 1 in blood tissue and non-Hispanic White individuals from the *** cohort. The shaded blue regions represent the ±1Mb region around the transcription start site (TSS) for each gene. Horizontal dashed lines represent $\pm F^{-1}(1 - 5 \times 10^{-8})$, where $F^{-1}(\alpha)$ is the quantile function of the standard normal distribution evaluate at $\alpha$. These results indicate that eQTL Z-statistics are highly likely to be considered not genome-wide significant, i.e., to have a corresponding P-value greater than $5 \times 10^{-8}$ outside of the ±1Mb window from the TSS.

refers to the nonzero variance of $\hat{\beta}_{j\ell}$, the *estimated* association of the $j$th SNP with the $\ell$th gene in a select tissue. Only when $\hat{\beta}_{j\ell} = \beta_{j\ell}$, the *true* association, is there no measurement error in $\hat{\beta}_{j\ell}$. Let $\mathbf{\Sigma}_{W_\beta W_\beta}$ denote the $p \times p$ variance-covariance matrix of $\hat{\boldsymbol{\beta}}_j$, the $p$-length vector of associations between the $j$th SNP and $p$ genes in a locus. Let $\widehat{\mathbf{B}}$ be the $m \times p$ matrix of estimated associations between $m$ SNPs and the expression of $p$ genes, $\mathbf{B}$ denote the corresponding matrix of true associations, and $\mathcal{O}$ be the set of non-missing values in $\widehat{\mathbf{B}}$, of which there are $|\mathcal{O}|$.

The main principle of soft imputation is to iteratively apply soft thresholding to the singular values of $\widehat{\mathbf{B}}$ until convergence. Since $\widehat{\mathbf{B}}$ contains missing values, we first impute the missing values in $\widehat{\mathbf{B}}$ with 0, a reasonable estimate of their true value given the results from individual-level data presented in Figure 4. Our matrix completion algorithm is presented in Algorithm 1 in the main text. This algorithm modifies the traditional soft impute method [6] by subtracting the singular values of $\mathbf{\Sigma}_{W_\beta W_\beta}$ from the singular values of an initialized $\widehat{\mathbf{B}}$. This method requires the tuning parameter $\lambda$ to be used in soft thresholding and will only accept solutions in which the rank of the imputed matrix is less than a user-defined value. Below, we evaluate the performance of this imputation method in simulation in Figure 6 and provide some examples using real data in Figure 7.

### 1.3.2 Simulations

First, we simulated true associations between $m = 150$ SNPs and $p = 10$ genes, which formed the matrix $\mathbf{B}$. Next, we randomly drew association estimates $\hat{\mathbf{B}}$ from the matrix normal distribution $\mathcal{N}(\mathbf{B}, \mathbf{R}, \mathbf{\Sigma}_{W_\beta W_\beta})$, where $\mathbf{R}$ is the matrix of LD correlations between the 100 SNPs. In our simulations $\mathbf{R}$ had a first-order autoregressive structure with correlation parameter $\rho$ which was in the set $\{0.0, 0.1, ..., 0.8, 0.9\}$. The matrix $\mathbf{\Sigma}_{W_\beta W_\beta}$ representing measurement error covariance between the rows of $\hat{\mathbf{B}}$ had a Toeplitz structure and was multiplied by the factor $\sqrt{m}\log p \approx 5.3$. We then applied our matrix completion algorithm to these data, searching over a grid of $\lambda$ parameter values and fixing the maximum acceptable rank of the solution at 5.

The simulation results in Figure 6 demonstrate that our imputation method well-approximates the true underlying distribution of the observed association values when the true mean of the missing association values is 0, and that LD structure does not appear to affect these results. An example of the imputation for a single case is presented in Figure 7. Results from individual level data presented in Figure 4 demonstrate that the true mean is likely to be 0 for almost all areas outside of the ±1Mb window of a gene's center. Results from real data presented in Figure 8 demonstrate that this imputation method can capture the variance in association estimates at the boundaries of the observed windows well, and that association estimates further from the gene center approach 0 with decreasing variance. The data used in these results are described in the caption of Figure 8.

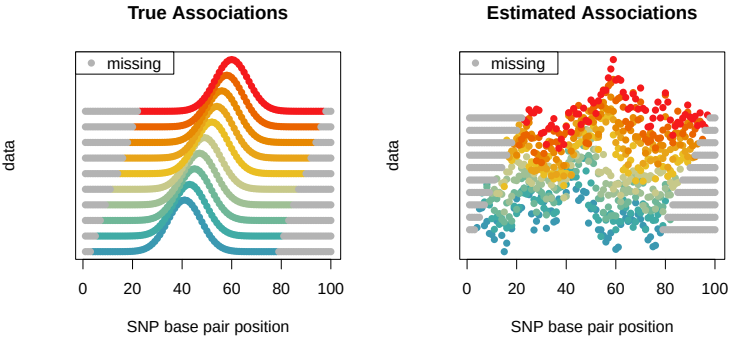


**Fig. 5** This figure displays the data we generated to perform simulation using our eQTL imputation method described in Algorithm 1 in the main text. Positions on the x-axis correspond to unique SNPs. Each gene is represented by a different color, where gray always represents missing values. For each gene, y-axis values are arbitrary but the relative magnitude corresponds to the magnitude of association with the SNP at that base pair position with the expression of the specific gene. The base pair locations of missingness for each gene depend on the base pair position of the gene center, which is located at the peak of its distribution. Gene centers/distribution peaks are staggered for each gene to replicate the real data. The left panel displays true association values and the right panel displays an example of estimated association values.

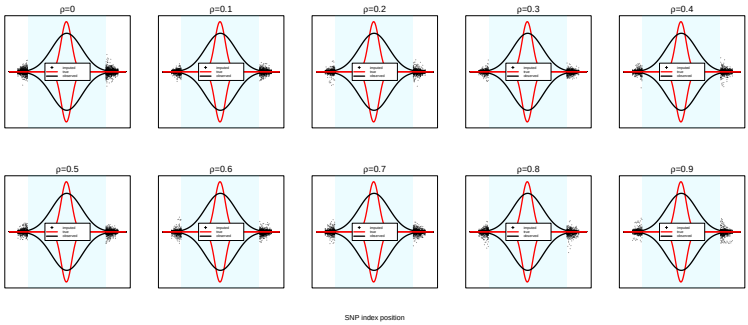


**Fig. 6** These are the results of simulations in which LD structure among the 150 SNPs varied from an AR1 structure with correlation paramater $\rho = 0$ to $\rho = 0.9$. Y-axis display the relative strength of association between a SNP indexed on the x-axis and the expression of the first gene of 10 in simulation. The true distributions of the true and observed associations are respectively represented by red and black lines. All imputed values across all 100 simulations are represented by black points.

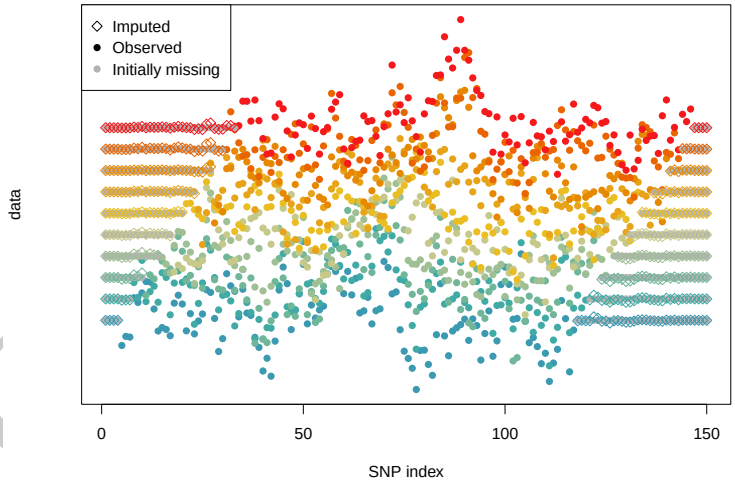


**Fig. 7** These are the results of simulations which are described in the text above this figure when $\mathbf{R}$ has an AR1 structure with correlation parameter $\rho = 0.5$. These results demonstrate that our imputation method well approximates the true underlying association values when the data is missing, which in this simulation are each 0. Observed values are equal to the true values plus measurement error from finite GWAS sample size.

## 1.4 Power after imputing missing values

As mentioned above and in the main text, current eQTL-MVMR approaches are restricted to using IVs for which associations between all SNPs and target genes have been observed in the summary eQTL GWAS data. In this

section, we present the results of a simulation in which we compare the power of our multivariate imputation method and current methods that use only completely observed data for testing the causal null hypothesis. We simulated summary-level data using the same procedure described in Section 1.3.2 but varying proportions of total missingness in the true $100 \times 10$ design matrix $\mathbf{B}$, ranging from 19% to 85%. This was accomplished by varing the proportion of missingness that was present for each gene. We compared the power of the IVW method [7] with correlated IVs when we excluded IVs with any missing to power when we imputed missing using our IFC approach. The full R code used to perform these simulations and generate its results are present at https://github.com/noahlorinczcomi/HORNET/simulations.

These results indicate that using imputed vs fully observed data can result in tests of the causal null hypothesis that are up to approximately 4 times as powerful when the proportion of missingness is large. When the proportion of missingness is moderate around 52%, which is consistent with the results observed in real data analyses (see Figure 3 in Section 1.1), still applying imputation to the observed eQTL summary statistics can result in approximately 1.27x more power. The gains in power continue to become larger after approximately 35% or more of the eQTL associations are imputed.

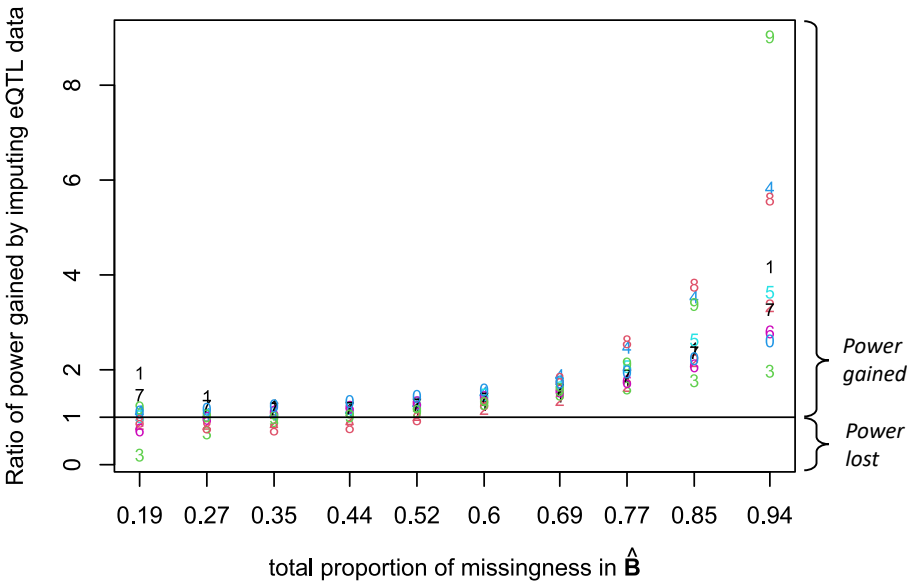


**Fig. 8** This figure displays differences in the power of eQTL-MVMR using IVW with completely observed vs imputed eQTL summary statistics. The x-axis is the proportion of missingness in the working design matrix $\hat{\mathbf{B}}$ and the y-axis the the power when using imputed data divided by the power when using the fully observed data (a 'complete SNP' analysis). 'Power' in this setting refers to the power of rejecting the causal null hypothesis. Each point is a number which indicates the gene. For example, '1' corresponds to the first gene and '0' to the 10th. The horizontal line is placed at 1, below which the complete SNP analysis is more powerful than the analysis using imputed data and above which the converse is true by the factor displayed on the y-axis. Displayed are the power ratios after smoothing power estimates within each analysis type using quadratic regression.



# 2 Accounting for LD in eQTL-MVMR

## 2.1 CHP bias from LD

### 2.1.1 Notation

A $m \times p$ matrix $\mathbf{X}$ with normally-distributed elements will be denoted as $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{R})$, where $\boldsymbol{\Sigma} : p \times p$ represents covariance between columns of $\mathbf{X}$ and $\mathbf{R} : m \times m$ covariance between rows. $\mathbf{X}$ can also be written in vectorized form as $\text{vec}(\mathbf{X}) \sim \mathcal{N}(\text{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma} \otimes \mathbf{R})$.

Consider two loci (denoted as locus 1 and 2), where locus 1 contains $p_1$ genes that use $m_1$ SNPs as instruments (IVs) and locus two contains $m_2$ SNPs, where all of the $m_1$ SNPs are *cis*-eQTLS for at least one gene in their locus. Denote $\widehat{\mathbf{B}}_1 = (\beta_{ij})_{i,j=1}^{m_1,p_1} : m_1 \times p_1$ and $\widehat{\mathbf{B}}_2 : m_2 \times p_2$ as the GWAS estimates for association with the expressions of the $p_1$ genes in locus 1 and the $p_2$ genes in locus 2. Denote $\widehat{\boldsymbol{\alpha}}_1$ as outcome GWAS estimates for $m_1$ SNPs in locus 1. Assume all GWAS estimates are standardized to have variance 1 and let $\mathbf{R}_1 : m_1 \times m_1$ and $\mathbf{R}_2 : m_2 \times m_2$ denote the true LD correlation matrices

for SNPs in loci 1 and 2, where $\mathbf{R}_{12} : m_2 \times m_1$ is the LD correlation matrix between the $m_1$ and $m_2$ SNPs. Define $\mathbf{x}_1 : p_1 \times 1$ as the vector of total gene expression in a tissue for $p_1$ genes in locus 1 and $\mathbf{g}_1 : m_1 \times 1$ as the genotype vector for the $m_1$ SNPs in locus 1.

### 2.1.2 Models

Consider that the causal effects $\boldsymbol{\theta}$ of the $p_1$ gene expressions in locus 1 on the outcome trait $y$ are of interest and we want to use MR to estimate them. First, we can specify a model for the relationship between $\mathbf{g}_1$ and $\mathbf{g}_2$. Assume that the elements of $\mathbf{g}_1$ and $\mathbf{g}_2$ are approximately normally distributed, or there is some underlying normal distribution from which their realizations are drawn. It follows that $\mathbf{g}_2 = \boldsymbol{\lambda}_{12}^\top \mathbf{g}_1 + \boldsymbol{\epsilon}_2$ where $\boldsymbol{\lambda}_{12} \approx \mathbf{R}_{12}\mathbf{R}_1^{-1}$. Now we specify the following models for the expression values for the $p_1$ genes in locus 1 and their causal effects on the outcome trait:

$$\begin{aligned}
\mathbf{x}_1 &= \boldsymbol{\gamma}_1^\top \mathbf{g}_1 + \boldsymbol{\gamma}_2^\top \mathbf{g}_2 + \boldsymbol{\epsilon}_1 && (1)\\
&= (\boldsymbol{\gamma}_1^\top + \boldsymbol{\gamma}_2^\top \boldsymbol{\lambda}_{12}^\top)\mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_1, && (2)\\
&= \mathbf{B}_1^\top \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_1 && (3)\\
y &= \boldsymbol{\theta}^\top \mathbf{x}_1 + \boldsymbol{\pi}^\top \mathbf{g}_2 + \boldsymbol{\epsilon}_y && (4)\\
&= \boldsymbol{\theta}^\top \mathbf{B}_1^\top \mathbf{g}_1 + \boldsymbol{\pi}^\top \boldsymbol{\lambda}_{12}^\top \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_y && (5)\\
&= \boldsymbol{\alpha}_1^\top \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_y, && (6)
\end{aligned}$$

where $\tilde{\boldsymbol{\epsilon}}_y$ represents uncorrelated error in a simplified notation. The above results imply that

$$\boldsymbol{\alpha}_1 = \mathbf{B}_1\boldsymbol{\theta} + \boldsymbol{\lambda}_{12}\boldsymbol{\pi}, \quad (7)$$

where we want to use MR to estimate $\boldsymbol{\theta}$. Figure 9 shows these models in a directed acyclic graph (DAG).

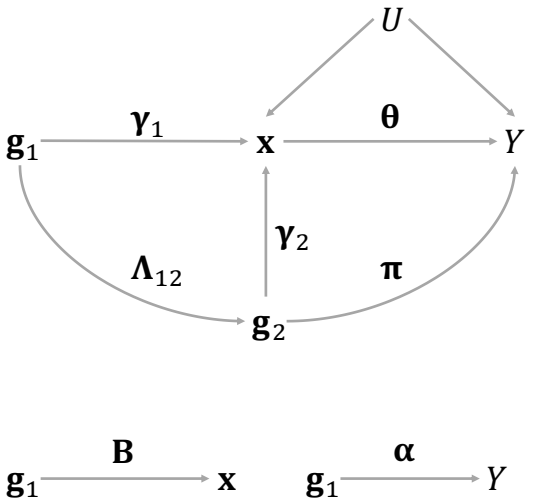


**Fig. 9** DAG representing the models specified in section 3.1. $\mathbf{g}_1 : m_1 \times 1$ is a vector of SNP genotypes used as IVs in MR to estimate $\boldsymbol{\theta}$, $\mathbf{g}_2 : m_2 \times 1$ is a generic vector of genotypes for other SNPs no in $\mathbf{g}_1$, $\mathbf{x} : p \times 1$ is a vector of expressions for $p$ genes in a tissue, $Y$ is the outcome trait, and $U$ is a generic confounding. If $\mathbf{g}_1$ and $\mathbf{g}_2$ are in LD, $\boldsymbol{\Lambda}_{12} \neq \mathbf{0}$. If $\mathbf{g}_2$ is associated with $Y$ conditional on $\mathbf{x}$, $\boldsymbol{\pi} \neq \mathbf{0}$.

In practice, we only have access to GWAS estimates of $(\boldsymbol{\alpha}, \mathbf{B})$, which we denote as $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{B}})$. Therefore, we use the following model to estimate $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbf{B}}\boldsymbol{\theta} + \boldsymbol{\Lambda}_{12}^{\top}\boldsymbol{\pi} + \tilde{\boldsymbol{\epsilon}}, \tag{8}$$

where $\tilde{\boldsymbol{\epsilon}}$ contains the measurement errors $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$ and $\hat{\mathbf{B}} - \mathbf{B}$ (see Section **??** below). When estimating $\boldsymbol{\theta}$ in Equation 7 using only $\hat{\mathbf{B}}$, there will be horizontal pleiotropy bias if $\boldsymbol{\Lambda}_{12}^{\top}\boldsymbol{\pi} \neq \mathbf{0}$, which may be considered unbalanced if $m^{-1}\mathbf{1}_m^{\top}\boldsymbol{\Lambda}_{12}^{\top}\boldsymbol{\pi} \neq 0$ and correlated horizontal pleiotropy (CHP) if the correlation between $\boldsymbol{\Lambda}_{12}^{\top}\boldsymbol{\pi}$ and $\text{vec}(\hat{\mathbf{B}})$ is not $\mathbf{0}$. It was shown above that $\boldsymbol{\pi}$ is the association of $\mathbf{g}_2$ with $Y$ conditional on $\mathbf{g}_1$. Next we aim to provide an expression for the joint distribution of $(\hat{\mathbf{B}}, \boldsymbol{\Lambda}_{12}^{\top}\boldsymbol{\pi})$ to identify the potential sources of CHP bias in Equation 7.

First, we state the marginal distribution of $\hat{\mathbf{B}} = \mathbf{B} + \mathbf{W}_{\beta}$ where $\mathbf{B} = (\boldsymbol{\beta}_k)_{k=1}^{p}$ and $\mathbf{W}_{\beta}$ are random. As in [8, 9], let

$$\boldsymbol{\beta}_k \sim \epsilon_k \text{N}\left(\mathbf{0}, \frac{h_k^2}{\tilde{m}_k}\mathbf{I}_{m_1}\right) + (1 - \epsilon_k)\text{N}(\mathbf{0}, \mathbf{0}) \tag{9}$$

be a mixture of $\tilde{m}_k$ SNPs that are associated with the expression of gene $k$ and $m_1 - \tilde{m}_k$ that are not. We specify this mixture explicitly because in the data it

is true since the total set of $m$ SNPs used in MR is not a set of SNPs associated with the expression levels of *all* genes in a group and tissue, but typically only a subset of genes. The estimation error $\mathbf{W}_\beta = (\mathbf{w}_{\beta_k})_{k=1}^p$ is uncorrelated with $\mathbf{B}$ and has the distribution

$$\mathbf{w}_{\beta_k} \sim \mathrm{N}\left(\mathbf{0}, \frac{1}{n_k}\mathbf{R}_1\right) \tag{10}$$

for all $m_1$ SNPs where $n_k$ is the sample size in the GWAS for the expression of the $k$th gene. In MR, we have $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_k)_{k=1}^p$, whose columns will have the distribution

$$\hat{\boldsymbol{\beta}}_k \sim \epsilon_k \mathrm{N}\left(\mathbf{0}, \left[\frac{h_k^2}{\tilde{m}_k}\mathbf{I} + \frac{1}{n_k}\mathbf{R}_1\right]\right) + (1-\epsilon_k)\mathrm{N}\left(\mathbf{0}, \frac{1}{n_k}\mathbf{R}_1\right). \tag{11}$$

Recall that CHP bias can arise when $\hat{\mathbf{B}}$ is correlated with $\boldsymbol{\Lambda}_{12}^\top\boldsymbol{\pi}$. Let $\boldsymbol{\Lambda}_{12}^\top(\boldsymbol{\tau}^\top \otimes \mathbf{R}_{12}^\top) := \mathrm{Cov}(\mathrm{vec}[\mathbf{B}], \boldsymbol{\Lambda}_{12}^\top\boldsymbol{\pi})$ where $\boldsymbol{\tau}^\top : p \times 1$ represents genetic covariance between $\boldsymbol{\pi}$ and the columns of $\mathbf{B}$. For example, consider $\mathrm{Cov}(\boldsymbol{\beta}_k, \boldsymbol{\pi}) := \tau_k\mathbf{R}_{12}^\top = [\mathrm{E}(\beta_{jk}\pi_s r_{js})]_{j,s=1}^{m_1,m_2}$. Since this covariance has a kronecker product structure, it can be zero if either one of two conditions are met, namely if (i) $\boldsymbol{\tau} = \mathbf{0}$ or (ii) $\mathbf{R}_{12}^\top = \mathbf{0}$. In principle, these conditions are met if either (i) the association between $\mathbf{g}_2$ and $Y$ *conditional on* $\mathbf{g}_1$ is independent of the *total* association between $\mathbf{g}_1$ and $Y$ or (ii) $\mathbf{g}_1$ and $\mathbf{g}_2$ are not in LD.

### 2.1.3 CHP bias in traditional MR methods

Summary-based MR (SMR) [10] and MR-Robin [11], MR methods incorporating eQTLs that can only include a single gene in causal estimation, may suffer from UHP and/or CHP bias because of nonzero $\mathbf{R}_{12}$ and $\boldsymbol{\pi}$ for neighboring genes. In this section, we aim to better understand the extent to which SMR (a simmpler version of MR-Robin more popularly used) is vulnerable to UHP and CHP bias when considering Alzheimer's disease (AD) as the outcome trait and the expressions in blood of genes on chromosome (CHR) 19 using the real data that we used in the main text. First, we identified mutually exclusive groups of genes using the procedure described in **Methods**. First we define some notation within a group of genes. Let $\mathcal{M}$ denote the set of $M$ SNPs used as IVs for the entire group of $p$ genes, $\mathcal{M}_k$ denote the set of $m_k$ SNPs that are IVs for the $k$-th gene, $\mathbf{R}_k$ be the LD matrix for this gene, $\mathcal{M}_k^\perp$ be the set of SNPs in $\mathcal{M}$ but not in $\mathcal{M}_k$ that are in LD with SNPs in $\mathcal{M}_k$ via $\mathbf{R}_{k,-k}$, $\boldsymbol{\Lambda}_{k,-k} \approx \mathbf{R}_{k,-k}^\top\mathbf{R}_k^{-1}$, and $\boldsymbol{\pi}_{-k}$ be the association between SNPs in $\mathcal{M}_k^\perp$ and AD risk conditional on SNPs in $\mathcal{M}_k$. We estimated the following quantities:

$$I_1 = \|\mathbf{R}_k^{-1/2}\boldsymbol{\Lambda}_{k,-k}^\top\boldsymbol{\pi}_{-k}\|_2^2, \tag{12}$$

$$I_2 = \frac{1}{m_k}\mathbf{1}_{m_k}^\top\mathbf{R}_k^{-1/2}\boldsymbol{\Lambda}_{k,-k}^\top\boldsymbol{\pi}_{-k}, \tag{13}$$

$$I_3 = \theta_k - (\boldsymbol{\theta})_k, \tag{14}$$

where $\theta_k$ is estimated in univariable (single-gene) MR and $(\boldsymbol{\theta})_k$ is the multivariable (multiple-gene) MR estimate for the corresponding $k$th gene. Let $\hat{\boldsymbol{\delta}}_k := \hat{\boldsymbol{\alpha}}_k - \hat{\mathbf{B}}_k\hat{\theta}_k$. Below we list estimands for each of these quantities and their corresponding distributions under specified null hypotheses:

$$\hat{I}_1 = \|\boldsymbol{\Sigma}_{\Delta k}^{-1/2}\hat{\boldsymbol{\delta}}_k\|_2^2 \sim \chi^2(m_k) \quad \text{under} \quad H_0 : \boldsymbol{\Lambda}_{k,-k}^\top \boldsymbol{\pi}_{-k} = \mathbf{0} \tag{15}$$

$$\hat{I}_2 = \frac{1}{m_k}\mathbf{1}_{m_k}^\top \boldsymbol{\Sigma}_{\Delta k}^{-1/2}\hat{\boldsymbol{\delta}}_k \sim \mathrm{N}(\mathbf{0}, \eta) \quad \text{under} \quad H_0 : \frac{1}{m_k}\mathbf{1}_{m_k}^\top \boldsymbol{\Lambda}_{k,-k}^\top \boldsymbol{\pi}_{-k} = \mathbf{0} \tag{16}$$

$$\hat{I}_3 = \|\hat{\sigma}_\Theta^{-1/2}[\hat{\theta}_k - (\hat{\boldsymbol{\theta}})_k]\|_2^2 \sim \chi^2(1) \quad \text{under} \quad H_0 : \theta_k = (\boldsymbol{\theta})_k, \tag{17}$$

where

$$\eta := \frac{1}{m_k^2}\mathbf{1}_{m_k}^\top \boldsymbol{\Sigma}_\Delta \mathbf{1}_{m_k} \approx \frac{1}{m_k}, \tag{18}$$

$$\boldsymbol{\Sigma}_\Delta = \mathrm{Cov}(\hat{\boldsymbol{\delta}}_k) \tag{19}$$

$$\hat{\boldsymbol{\Sigma}}_\Delta = \mathbf{R}_k + \hat{\theta}_k^2 \sigma^{2(k)}_{\mathbf{W}_\beta \mathbf{W}_\beta}\mathbf{R}_k - 2\hat{\theta}_k \sigma^{(k)}_{\mathbf{W}_\beta \mathbf{w}_\alpha}\mathbf{R}_k \tag{20}$$

and

$$\hat{\sigma}_\Theta^2 = \widehat{\mathrm{Var}}(\hat{\theta}_k) + \widehat{\mathrm{Var}}[(\hat{\boldsymbol{\theta}})]_{k,k} - 2\widehat{\mathrm{Cov}}[\hat{\theta}_k|\hat{\boldsymbol{\beta}}_k, (\hat{\boldsymbol{\theta}})_k|\hat{\mathbf{B}}]. \tag{21}$$

Where $\hat{\sigma}_k$ and $(\hat{\sigma})_k$ represent the estimated standard deviations of the residuals during estimation of $\hat{\theta}_k$ and $\hat{\boldsymbol{\theta}}_k$ and $\mathbf{R}_{k(k)}$ is the LD matrix between valid IVs (see below for criteria) used in their respective estimators, $\hat{\sigma}_\Theta^2 = \mathbf{A}_k^{-1}\hat{\sigma}_k(\hat{\sigma})_k\mathbf{R}_{k(k)}\mathbf{A}_{(k)}^{-\top}$ for constant matrices $\mathbf{A}_k$ and $\mathbf{A}_{(k)}$. Regarding $I_2$, since $\hat{\boldsymbol{\delta}}_k$ is the estimated residual from linear regression by MR using the expression of gene $k$ as the exposure, $\hat{\boldsymbol{\delta}}$ is guaranteed to have a sample mean of $\mathbf{0}$. However, for each of $p$ genes in a group, we used the MRBEE estimator [12] with IMRP adjustment [13]. This method can estimate $\theta_k$ without bias from horizontal pleiotropy using a subset of the $m_k$ IVs, after which $\hat{\boldsymbol{\delta}}_k$ will become a reliable estimator for $\boldsymbol{\delta}_k$ (see [9]) and $\hat{\boldsymbol{\delta}}_k$ is not guaranteed to have a sample mean of $\mathbf{0}$.

To obtain an unbiased estimate for $\theta_k$, we also applied the following restrictions on the IV set: (i) P-value for association with gene expression less than $5\text{x}10^{-5}$, (ii) absolute LD between SNPs used to instrument expression of the $k$th gene less than 0.9, (iii) $\geq$10 candidate IVs evaluated by MRBEE (some of which may have been further excluded due to evidence of nonzero horizontal pleiotropy at $P < 0.05$ using the tests in [9, 12, 13]), and (iv) LD between the $j$th of $m_k$ SNPs and the $M - m_k$ other SNPs in the group less than 0.2. The latter worked to reduce bias from CHP via nonzero $\mathbf{R}_{k,-k}$ while still retaining enough SNPs for efficient estimation of $\hat{\theta}_k$. Regarding $I_3$, rejection of the corresponding null hypothesis is evidence of omitted-variable bias (OVB) (see [9, 12]), which can be due to mediation or confounding (CHP). Both may be

considered biased causal estimates, although where this bias is due to CHP or mediation cannot be determined by the test for $I_3 \neq 0$.

Under these restrictions to obtain a valid IV set, univariable MR using bias-corrected SMR (i.e., single-exposure MRBEE [9, 12]) could only be performed for 194 of the 752 ( 25%) total genes with *cis*-eQTLs on CHR 19. This is another major limitation of univariable MR - the valid IV set can be reduced so small that causal estimation becomes unreliable and therefore should not even be performed. Figures 10, 11, 12, and 13 provide some inference for $I_1$, $I_2$, and $I_3$, respectively. These results indicate substantial nonzhero unbalanced horizontal pleiotropy across CHR 19. Subsequently, there is widespread evidence of differences in causal estimates made using univariable vs multivariable causal estimates, suggesting the presence off OV bias that may be due to CHP. We found that 37.6% of genes tested using univariable MR on CHR 19 (73/194) had evidence ($P < 5 \times 10^{-5}$) of nonzero horizontal pleiotropy ($I_1 \neq 0$), 13.7% of which had evidence of imbalance ($I_2 \neq 0$), and 48.2% of genes had multivariable causal estimates that differed from univariable causal estimates ($I_3 \neq 0$ where the test was available [see Figure 13 caption]). See the corresponding figure captions for more details.

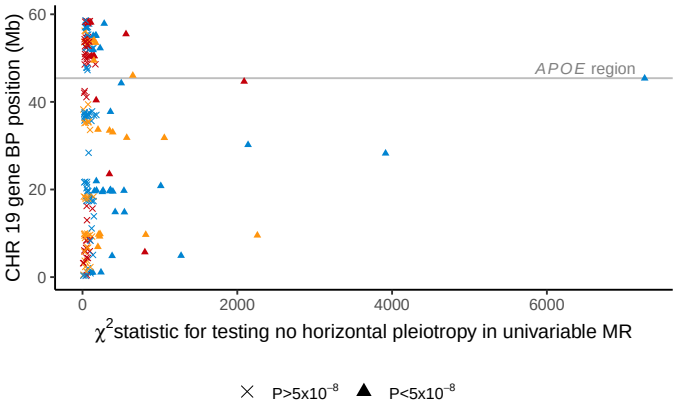


**Fig. 10** These results display $\chi^2$ test statistics for testing $H_0 : I_1 = 0$ as stated above. Each point represents a gene. This is a test for nonzero horizontal pleiotropy in the IV set in univariable MR using MRBEE [9, 12], which can be considered a version of SMR [10, 14] corrected for bias from horizontal pleiotropy, weak instruments, sample overlap, and measurement/estimation error. This test was performed for each gene in CHR 19 that was put into a gene group in the main text. Different point colors represent distinct gene groups (see main text for how these groups were formed), with colors alternating from bottom to top on the y-axis from blue to red to yellow. Triangles represent genes for which $H_0$ is rejected at the level of genome-wide significance (i.e., $P < 5 \times 10^{-8}$); crosses represent genes for which $H_0$ is not rejected. The genomic region surrounding the *APOE* gene (known to be highly relevant for Alzheimer's disease risk) is labelled with a horizontal grey line. These results indicate substantial horizontal pleiotropy for many genes on this chromosome, where the strongest evidence of horizontal pleiotropy is observed in the *APOE* region. Only results for which univariable MR could be reliably performed are displayed (see text above figure).

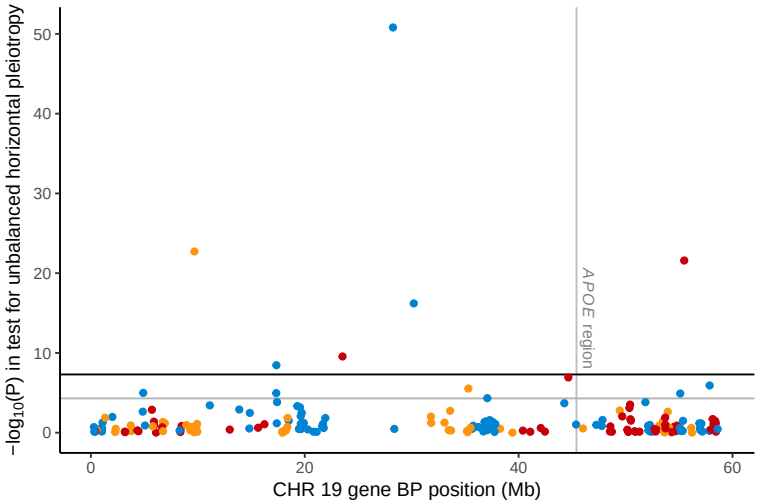


**Fig. 11** This is a test for unbalanced (i.e., nonzero mean) horizontal pleiotropy in univariable MR with MRBEE [9, 12], which can be considered a version of SMR [10, 14] corrected for bias from horizontal pleiotropy, weak instruments, sample overlap, and measurement/estimation error. Each point represents a gene. This test was performed for each gene in CHR 19 that was present in a gene group in the main text. Genome-wide ($P < 5 \times 10^{-8}$) and marginal ($P < 5 \times 10^{-5}$) significance thresholds are displayed by black and gray horizontal lines, respectively. Different point colors represent distinct gene groups (see main text for how these groups were formed), with colors alternating from left to right on the x-axis from blue to red to yellow. The genomic region surrounding the *APOE* gene (known to be highly relevant for Alzheimer's disease risk) is labelled with a vertical grey line. These results indicate that many genes have evidence of unbalanced horizontal pleiotropy in univariable MR, including genes in the *APOE* region. Only results for which univariable MR could be reliably performed are displayed (see text above Figure 10).

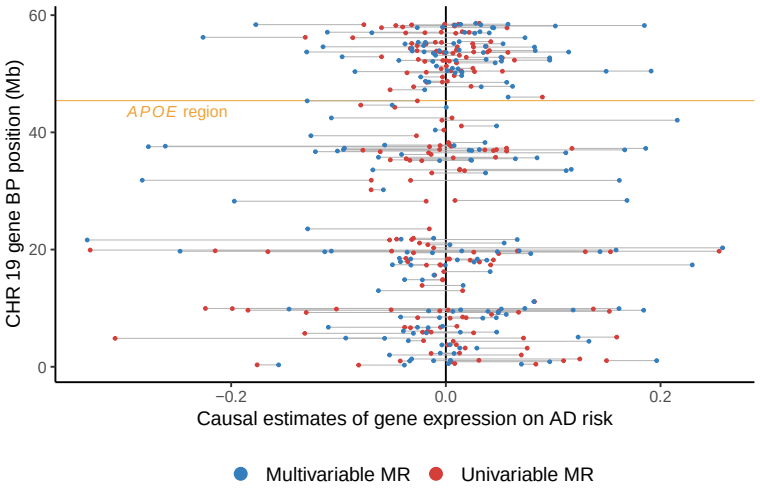


**Fig. 12** These results display differences between causal estimates made for each gene on chromosome (CHR) 19 using univariable MR vs multivariable MR. Each pair of points (paired by horizontal grey lines) corresponds to a single gene. Causal estimates were made using MRBEE [9, 12], which can be considered a version of SMR [10, 14] corrected for bias from horizontal pleiotropy, weak instruments, sample overlap, and measurement/estimation error. Blue points represent multivariable MR estimates and red points represent univariable MR estimates. An absence of ommitted variable (OV) bias across CHR 19 would be observed if all red and blue points overlapped. Differences between these points for each gene, represented by horizontal grey lines, indicates OV bias, which is observed for many genes across CHR 19. The *APOE* gene region is highlighted by the yellow horizontal line, in which OV bias is observed. Only results for which univariable MR could be reliably performed are displayed (see text above Figure 10).

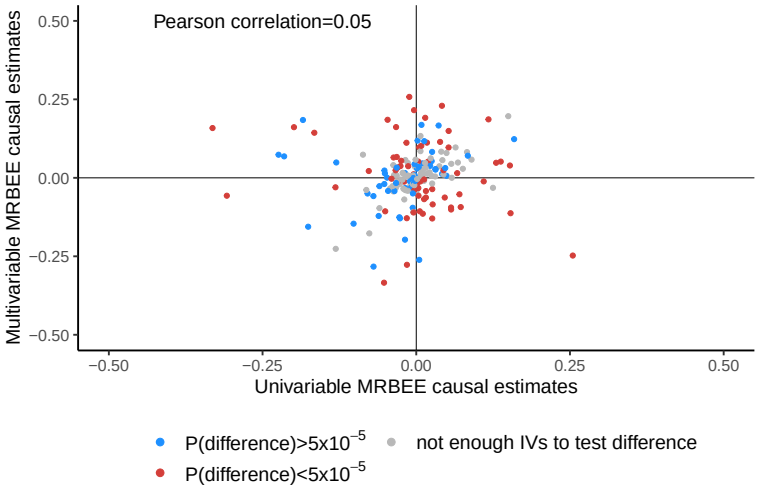


**Fig. 13** These results display the bivariate association between univariable MR and multivariable MR causal estimates and indications of the significance in testing $H_0 : I_3 = 0$ from the text above. Red and blue points respectively represent genes for which this null hypothesis was rejected and not rejected at $P< 5 \times 10^{-5}$. Grey points correspond to genes for which the test could not be reliably performed because of imprecise variance estimation in $\hat{I}_3$. For these genes, we could not estimate a positive $\hat{\sigma}^2_{\Theta}$ (see Equation 21) because of very small valid IV counts in univariable MR. 'Pearson correlation' corresponds to the linear correlation between univariable and multivariable MR causal estimates. This value will be 1 if there is no omitted variable (OV) bias (due either to CHP or mediation effects) and will approach 0 as OV bias becomes stronger.

### 2.1.4 HORNET CHP correction

The method of protecting against CHP bias from eQTLs from other loci that are in LD with eQTLs in a target locus is outlined in Figure 14.

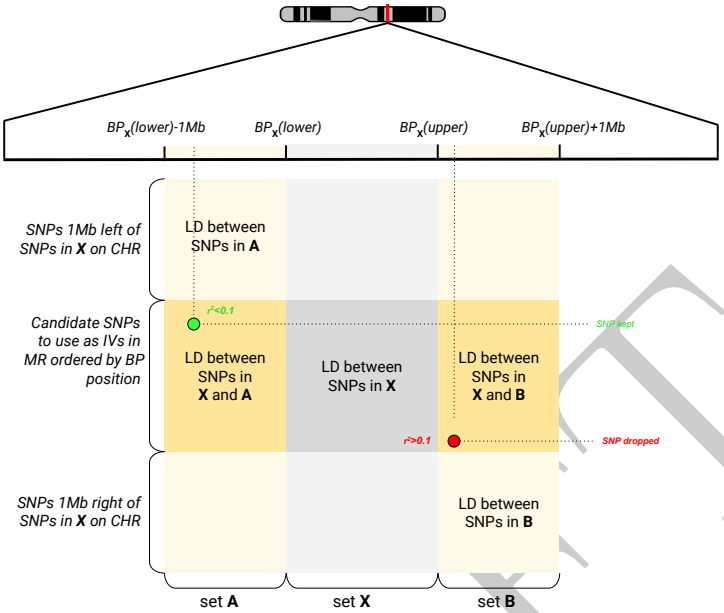


**Fig. 14** Visual depiction of how the final IV sets were pruned based on their LD with non-IVs within surrounding Mb windows of defined size (1Mb as the example in this figure). It is shown in section **??** that horizontal pleiotropy bias may be present in MR if the IVs used are in LD with SNPs that are conditionally associated with the outcome trait given the exposures. To reduce the potential for this bias, we subsetted the original IV set to only those SNPs that were not in LD $r^2 > 0.1$ with any SNPs within a 1Mb window outside of the BP range of the original IV set. This is displayed graphically as the green and red points in the figure, where **A** and **B** are sets of SNPs within 1Mb of the minimum and maximum BP positions of the original IV set, respectively.

## 2.2 MRBEE bias-correction under LD

We now aim to demonstrate the validity of the MRBEE bias-correction in the case of correlated IVs (i.e., SNPs in LD used as IVs in MR). The original MRBEE theory [9] was based on independent IVs, but we demonstrate here that the bias-correction in that case is the same as in our case of correlated IVs under a fixed-effects model for $\mathbf{B}$. Let $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_j)_{j=1}^{m} = (\hat{\beta}_{jk})_{j=1,k=1}^{m,p}$ be the $m \times p$ matrix of GWAS estimates of the $m$ IVs on the expressions of $p$ genes in a tissue (MR exposures), $\hat{\boldsymbol{\alpha}}_j = (\hat{\alpha}_j)_{j=1}^{m}$ be the $m$-length vector of IV estimates of association with the outcome, and the IVs have the $m \times m$ positive definite LD correlation matrix $\mathbf{R} = (r_{js})_{js=1}^{m}$, where $\mathbf{P} = \mathbf{R}^{-1}$. We now assume a fixed effects model for $(\boldsymbol{\alpha}, \mathbf{B})$ for the purposes of causal estimation. This may be considered equivalent to causal estimation using MR conditional on the true causal SNPs used to instrument the exposures. We defined measurement error

models for $(\hat{\boldsymbol{\beta}}_j, \hat{\alpha}_j)$ as in [9, 12] in the following way:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_j \\ \hat{\alpha}_j \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_j + \mathbf{w}_{\beta_j} \\ \alpha_j + w_{\alpha_j} \end{pmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{\beta}_j \\ \boldsymbol{\alpha}_j \end{bmatrix}, \boldsymbol{\Lambda} := \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{w}_\beta \mathbf{w}_\beta} & \boldsymbol{\sigma}_{\mathbf{w}_\beta w_\alpha} \\ \boldsymbol{\sigma}^\top_{\mathbf{w}_\beta w_\alpha} & \sigma^2_{w_\alpha} \end{bmatrix} \right), \tag{22}$$

where the errors in our measurements of $(\hat{\boldsymbol{\beta}}_j, \hat{\alpha}_j)$ were due only to sampling error introduced by finite GWAS sample sizes, and $(\boldsymbol{\beta}_j, \alpha_j)$ are fixed. MR-Jones (***) makes a bias-correction to the IVW [7] estimating equation, which we denote as $S_{IVW}(\boldsymbol{\theta})$. Let $\mathbf{W}_\beta = (\mathbf{w}_{\beta_j})_{j=1}^m$ and $\mathbf{w}_\alpha = (w_{\alpha_j})_{j=1}^m$. It is shown in [9, 12] that

$$\mathrm{E}[S_{IVW}(\boldsymbol{\theta})] = -\mathrm{E}(\mathbf{W}_\beta^\top \mathbf{P} \mathbf{W}_\beta) + \mathrm{E}(\mathbf{W}_\beta^\top \mathbf{P} \mathbf{w}_\alpha). \tag{23}$$

MRBEE and MR-Jones subtract from $S_{IVW}(\boldsymbol{\theta})$ the quantities in Equation 23 to produce an unbiased estimate of the causal parameter $\boldsymbol{\theta}$. Our goal now is to show that the quantity in Equation 23 is equal to $-\mathrm{E}(\mathbf{W}_\beta^\top \mathbf{W}_\beta) + \mathrm{E}(\mathbf{W}_\beta^\top \mathbf{w}_\alpha)$. Under the normality assumption in Equation 22,

$$\bar{\mathbf{W}} := (\mathbf{W}_\beta, \mathbf{w}_\alpha) \sim \mathrm{MatrixNormal}(\mathbf{0}_{m \times (p+1)}, \mathbf{R}, \boldsymbol{\Lambda}), \tag{24}$$

where $\mathbf{R}$ represents covariance between rows and $\boldsymbol{\Lambda}$ covariance between columns. By the positive-definiteness of $\mathbf{P} := \mathbf{R}^{-1}$,

$$\mathbf{P}^{1/2} \bar{\mathbf{W}} \sim \mathrm{MatrixNormal}(\mathbf{0}, \mathbf{I}_m, \boldsymbol{\Lambda}) \tag{25}$$

and

$$\bar{\mathbf{W}}^\top \mathbf{P} \bar{\mathbf{W}} \sim \mathrm{Wishart}(m, \boldsymbol{\Lambda}). \tag{26}$$

The proof for $\mathrm{E}(\mathbf{W}_\beta^\top \mathbf{P} \mathbf{W}_\beta)$ follows immediately from the properties of the Wishart distribution (see [**?** ]). Define the permutation matrix $\mathbf{C}_1 := (\mathbf{I}_{p \times p}, \mathbf{0}_{p \times 1})_{p \times (p+1)}$ such that

$$\mathbf{C}_1 \bar{\mathbf{W}}^\top \mathbf{P} \bar{\mathbf{W}} \mathbf{C}_1^\top = \mathbf{W}_\beta^\top \mathbf{P} \mathbf{W}_\beta. \tag{27}$$

It follows that

$$\mathrm{E}(\mathbf{C}_1 \bar{\mathbf{W}}^\top \mathbf{P} \bar{\mathbf{W}} \mathbf{C}_1^\top) = m \mathbf{C}_1 \boldsymbol{\Lambda} \mathbf{C}_1^\top = m \boldsymbol{\Sigma}_{\mathbf{W}_\beta \mathbf{W}_\beta}, \tag{28}$$

which is the desired result. For $\mathrm{E}(\mathbf{B}_\beta^\top \mathbf{P} \mathbf{w}_\alpha)$, we will show the proof element-wise also following the properties of the Wishart distribution. Consider the following:

$$\left[ (\mathbf{P}^{1/2} \mathbf{W}_\beta)^\top (\mathbf{P}^{1/2} \mathbf{w}_\alpha) \right]_k \sim \sigma^{[k]}_{\beta\alpha} \chi^2(m), \tag{29}$$

for $k = 1, ..., p$, which has expectation $m\sigma^{[k]}_{\beta\alpha}$ and since $m\boldsymbol{\sigma}_{\beta\alpha} = m(\sigma^{[k]}_{\beta\alpha})^{p}_{k=1}$ the result is proven.

## 2.3 Heritability estimation

Assume the random model for $\boldsymbol{\beta}_k : M \times 1$ as in Equation 9 holds for associations between $M$ SNPs and the expression of gene $k$ in a tissue (here $M$ is the total number of SNPs tested for association with gene expression). Let $\tilde{m}_k \leq M$ denote the total number of SNPs causally related to the expression of this gene in tissue which has SNP heritability $h^2_k$. In our analyses, all association estimates were standardized by estimated standard error in GWAS such that $\hat{\boldsymbol{\zeta}}_k \approx \sqrt{n_k}\hat{\boldsymbol{\beta}}_k$ for GWAS sample size $n_k$ and $\hat{\boldsymbol{\zeta}}_k$ was the unit of analysis. Under the assumptions in model 9,

$$h^2_k = \left[\frac{\mathrm{E}(\hat{\boldsymbol{\zeta}}^\top_k \hat{\boldsymbol{\zeta}}_k)}{M} - 1\right]\frac{\tilde{m}_k}{n_k}. \tag{30}$$

This is seen immediately from the following result

$$\mathrm{E}(\hat{\boldsymbol{\zeta}}^\top_k \hat{\boldsymbol{\zeta}}_k) = \mathrm{trace}\left(n_k\left[\frac{h^2_k}{\tilde{m}_k}\mathbf{I} + \frac{1}{n_k}\mathbf{R}\right]\right) \tag{31}$$

from the original model in Equation 11. A natural estimate of $h^2_k$ is

$$\hat{h}^2_k = \left[\frac{\hat{\boldsymbol{\zeta}}^\top_k \hat{\boldsymbol{\zeta}}_k - \frac{1}{n_k}}{M} - 1\right]\frac{\hat{\tilde{m}}_k}{n_k} \tag{32}$$

where the $-1/n_k$ term is introduced as a measurement/estimation error bias-correction [9, 12]. In practice, $\tilde{m}_k$ is rarely known and so must be estimated from the data. We estimated this quantity using a procedure similar to that used by PLINK [15] where we let $2\hat{\tilde{m}}_k$ be the number of SNPs of the total $M$ with association $P < 5 \times 10^{-5}$ for $\hat{\zeta}_{jk}$ that are independent ($\mathrm{r}^2 < 0.05$) of all other $M - 1$ SNPs for the gene group. We assume the factor 2 on $\tilde{m}_k$ consistent with results in [16**? ?** ] that *cis*-eQTLs explain approximately 1/3rd of the SNP heritability and *trans*-eQTLS explain the rest. Note that whether you assume a random or fixed effects model for $\boldsymbol{\beta}_k$, the result in Equation 30 is the same. To see this, let $\hat{\boldsymbol{\zeta}}_k \sim \mathrm{N}(\boldsymbol{\zeta}_k, n_k^{-1}\mathbf{R})$ and use the same technique as in Equation 31 then rearrange to arrive at the result in Equation 32.

## 2.4 Source of bias in MRBEE from a misspecified LD matrix

In MR with gene expression as the exposure(s) of interest, we use eQTLs as instrumental variables (IVs). Standard methods of performing MR assume that these eQTLs will be independent of each other. However, there may only be very few (e.g., less than 5) IVs in a cis-region that are significant eQTLs

and also independent of each other. If we only have, for example, 5 IVs to perform MR, there may be little power to detect causal effects. A more powerful approach would include IVs that are in LD with each other, assuming that a larger set of correlated IVs can explain more variance in the expression of a gene than a smaller set of independent IVs. Performing MR with $m$ correlated IVs requires estimating their LD matrix $\mathbf{R}$, which is usually accomplished in practice by using an external LD reference panel from approximately the same population, such as the 1000 Genomes reference panels [17]. It is well-known that the IVW estimator, equivalent to a generalized least squares estimator, will not be biased by misspecification of $\mathbf{R}$. That is, if in practice we use $\hat{\mathbf{R}} \neq \mathbf{R}$, the IVW estimator will not be biased because of it. The IVW estimator is generally biased from other sources as described above and in [9, 12].

MRBEE makes a bias-correction to IVW for these other sources of bias which include measurement error/weak IVs and sample overlap. The MRBEE estimator with a set of $m$ IVs with no evidence of horizontal pleiotropy is

$$\hat{\boldsymbol{\theta}}_{\texttt{MRBEE}} = \left(\hat{\mathbf{B}}^\top \hat{\mathbf{R}}^{-1} \hat{\mathbf{B}} - m\boldsymbol{\Sigma}_{W_\beta W_\beta}\right)^{-1} \hat{\mathbf{B}}^\top \hat{\mathbf{R}}^{-1} \hat{\boldsymbol{\alpha}}. \tag{33}$$

In contrast to IVW, if each $\hat{\alpha}_j$ is standardized such that it represents the Z-statistic for association between the $j$th IV and the outcome trait, then $\hat{\boldsymbol{\alpha}} \sim \mathcal{N}(\boldsymbol{\alpha}, \mathbf{R})$. If $\hat{\mathbf{R}} = \mathbf{R}$, then MRBEE is not biased by $\hat{\mathbf{R}}$. This follows from Equation 24 under the assumption that $\texttt{Var}(\hat{\boldsymbol{\alpha}}) = \mathbf{R}$. However, if $\hat{\mathbf{R}} \neq \mathbf{R}$ then the bias-correction to IVW that MRBEE makes is not correct and therefore $\hat{\boldsymbol{\theta}}_{\texttt{MRBEE}}$ may be biased. This can be seen by the following. In Equation 23, it was stated that the bias in the IVW estimating equation is

$$-\texttt{E}(\mathbf{W}_\beta^\top \hat{\mathbf{R}}^{-1} \mathbf{W}_\beta)\boldsymbol{\theta} + \texttt{E}(\mathbf{W}_\beta^\top \hat{\mathbf{R}}^{-1} \mathbf{w}_\alpha) \tag{34}$$

which MRBEE assumes to be $-m(\boldsymbol{\Sigma}_{W_\beta W_\beta}\boldsymbol{\theta} - \boldsymbol{\sigma}_{W_\beta w_\alpha})$. If $\hat{\mathbf{R}} \neq \mathbf{R}$, then the bias in Equation 34 is more complex and MRBEE does not correctly adjust for it.

We now aim to investigate the extent to which MRBEE, and by extension MR-Jones, will be biased by a misspecified value of $\hat{\mathbf{R}}$. In this section, we consider a simple case in which $\hat{\mathbf{R}} = \xi\mathbf{R} + (1-\xi)\mathbf{I}$ for some constant $0 \leq \xi \leq 1$. In section 2.6, we consider more complex cases in which the size of the LD reference panel also varies. We performed simulations with generated GWAS summary statistics for 100 IVs using the following models

$$(\boldsymbol{\alpha}, \mathbf{B}) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}, \frac{1}{4}\mathbf{R}\right) \tag{35}$$

$$\boldsymbol{\Sigma} = \mathbf{D} \begin{pmatrix} 1.0 & 0.2 & 0.2 \\ 0.2 & 1.0 & 0.5 \\ 0.2 & 0.5 & 1.0 \end{pmatrix} \mathbf{D} \tag{36}$$

$$\mathbf{D} = 0.1\mathbf{I}_3 \tag{37}$$

$$\mathbf{R} = \text{AR1}(0.5) \tag{38}$$

$$\widehat{\mathbf{B}} \sim \mathcal{N}\left(\mathbf{B}, \frac{1}{50}\boldsymbol{\Sigma}, \mathbf{R}\right) \tag{39}$$

$$\widehat{\mathbf{R}} = \text{AR1}(\xi), \qquad \xi \in \{0.0, 0.1, ..., 0.8, 0.9\} \tag{40}$$

$$\hat{\boldsymbol{\theta}}_{\text{IVW}} = \arg\min_{\boldsymbol{\theta}} -\frac{1}{2}\left(\boldsymbol{\alpha} - \hat{\mathbf{B}}^\top\boldsymbol{\theta}\right)^\top \widehat{\mathbf{R}}^{-1}\left(\boldsymbol{\alpha} - \hat{\mathbf{B}}^\top\boldsymbol{\theta}\right) \tag{41}$$

$$\hat{\boldsymbol{\theta}}_{\text{MRBEE}} = \arg\min_{\boldsymbol{\theta}} -\frac{1}{2}\left(\boldsymbol{\alpha} - \hat{\mathbf{B}}^\top\boldsymbol{\theta}\right)^\top \widehat{\mathbf{R}}^{-1}\left(\boldsymbol{\alpha} - \hat{\mathbf{B}}^\top\boldsymbol{\theta}\right) - \boldsymbol{\theta}^\top\boldsymbol{\Sigma}_{W_\beta W_\beta}\boldsymbol{\theta}, \tag{42}$$

$$\tag{43}$$

where the constants 1/4 and 1/50 respectively represent minor allele frequency and the proportion of measurement error variance to the total signal variance. These simulation models implicitly assume no measurement error in the outcome associations $\boldsymbol{\alpha}$, which is irrelevant for our purpose here because neither IVW nor MRBEE will have any more or less bias as measurement error is added or removed from $\boldsymbol{\alpha}$. We performed 10,000 simulations for each scenario in which $\xi$ varied and the results are presented in Figure 16.

These results indicate that IVW is consistently biased irrespective of how close the working LD matrix $\hat{\mathbf{R}}$ is the the true LD matrix $\mathbf{R}$. On the other hand, MRBEE is unbiased when $\hat{\mathbf{R}} = \mathbf{R}$, but incurs a small upward bias when $\xi < rho = 0.5$ and a small downward bias when $\xi > \rho = 0.5$. Each of these biases are smaller than the bias incurred by IVW, expect when an extremely dense AR1(0.9) structure is assumed to fit data that were generated from AR1(0.5), which is unlikely to ever occur in practice. Interestingly, MRBEE is unbiased when the LD matrix is assumed to be equal to the identity matrix, although its variance in this setting is greater than in other settings when a denser LD structure is assumed.

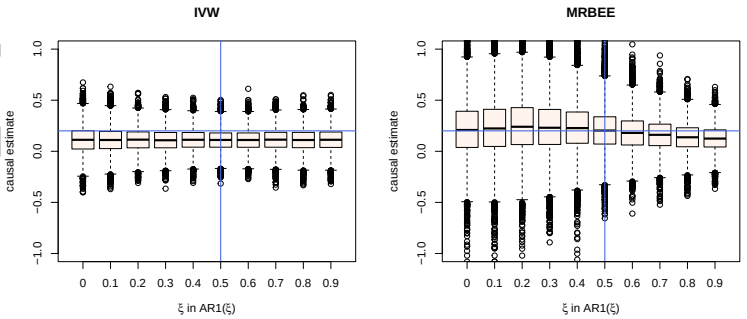


**Fig. 15** These are the results of 10,000 simulations, the settings for which are fully described in the text above. Displayed are boxplots of causal estimates made by IVW (left) and MRBEE (right) in different scenarios of assumed LD structure. The horizontal blue lines are positioned at the true causal effect, 0.2. The vertical blue lines are positioned at the value of $\rho$ which was used to generate the data.

## 2.5 Finding LD blocks

We present a simple yet flexible approach for automatically detecting LD blocks in an LD matrix. We start by assuming that in a $n \times n$ positive definite matrix $\mathbf{R}$ of LD covariances between $n$ SNPS in the set $\mathcal{G}$, there exists $k < \tau$ approximately independent blocks of minimum size $b$, where $(\tau, b)$ are hyper-parameters either set by the user or determined by our algorithm using the data. We also assume that the rows of $\mathbf{R}$ correspond to SNPs that are ordered by the base pair position, which is the standard output format of the PLINK [15] software that HORNET uses.

Our method for finding LD blocks follows the algorithm described below. The basic idea of this algorithm is to calculate the determinant for every submatrix of $\mathbf{R}$ starting from the first 2 SNPs, then the first 3, then the first 4, and so on. As each new SNP is added, if the subvector that is added with the SNP is the $\mathbf{0}$ vector, then the submatrix will have a new eigenvalue equal to 1 and the determinant will not change. Points at which a change in an LD block may occur are therefore points at which the determinant of the submatrix does not change when a new SNP is added. Our algorithm ranks the $\tau$ smallest changes in the sequence of determinants and finds the combination of cutpoints which minimize a global penalty function. The global penalty function is equal to the sum of local penalty functions. Let $\mathbf{R}_1 = \text{Cov}(\mathbf{g}_1)$ be the first $p_1 \times p_1$ submatrix of $\mathbf{R} : n \times n, n > p_1$, where the remaining $(n - p_1)^2$ elements of $\mathbf{R}$ are comprised of LD covariances between $\mathbf{g}_{-1}$ and $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_{-1})^\top$. Each local penalty function $P_{(\tau,b)}$ is equal to

$$P_{(\tau,b,k)} = -\log\|\mathbf{\Omega}_1\|_\text{F} \log p_1 + \log\|\mathbf{\Omega}_{-1}\|_\text{F} \log(n - p_1) + \log k, \qquad (44)$$

where $k$ is the number of blocks to use and $\mathbf{\Omega}_1 = (r_{ij}^2)_{i,j=1}^{p_1}$. In practice, $(\tau, b)$ can be easily set by the researcher, but $k$ must be determined from the data by searching a grid of $k$ values and finding the value which minimizes the global penalty. The R code used to implement this algorithm is available in the `LD_block_finder.r` file located at the `noahlorinczcomi/HORNET` Github repository. After applying the algorithm, elements of $\mathbf{R}$ that are not within an LD block are set to be 0. An example of this algorithm applied to 484 SNPs in the 2q37.1 region, where the LD matrix between these SNPs was estimated using 438k non-related European individuals in the UK Biobank [18] and was converted to a positive definite matrix using the method of [19]. These results indicate that our block-finding algorithm can detect LD blocks that, visually, appear to be approximately independent. An alternative approach would be to apply some regularization to the full LD matrix, such as soft or hard thresholding, tapering, or banding [20].

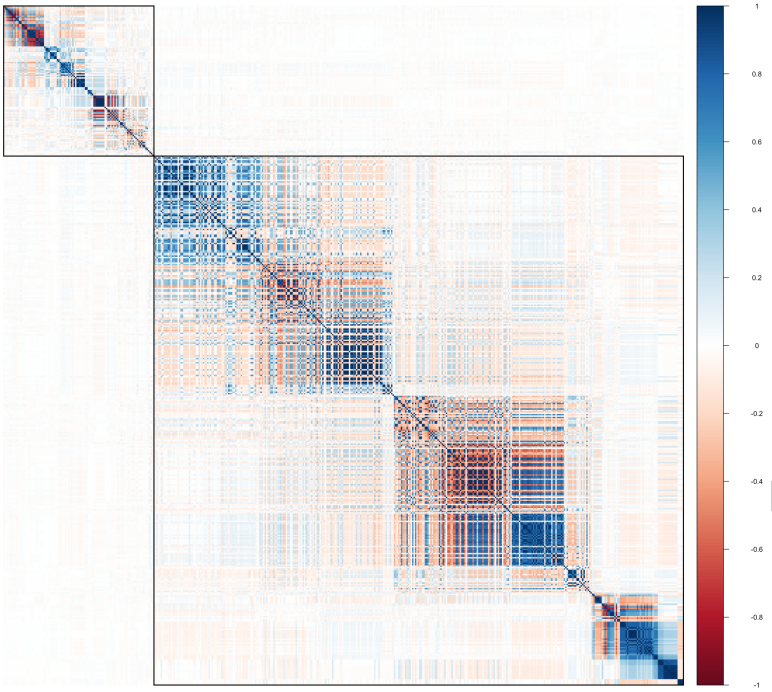


**Fig. 16** The LD matrix for the 484 SNPs displayed here was estimated from 438k non-related European individuals in the UK Biobank. The black lines are generated by our block-finding algorithm outlined in the text.

## 2.6 Misspecified LD

### 2.6.1 Background

We mentioned in Section 2.4 that using a LD matrix in MR that is not equal to the true LD matrix representative of the discovery population can introduce bias in causal estimates using MRBEE. In this section, we demonstrate that misspecified LD of this type can cause inflation of test statistics corresponding to tests of the causal null hypothesis. MR methods that can allow for IVs that are in LD with each other are IVW [21], principal components analysis (PCA) [22], the conditional and joint (CoJo) algorithm [23], single-SNP [24, 25], LD pruning [26, 27], effective-median [28], and the JAM algorithm (joint analysis of marginal summary statistics) [29]. An estimate of the LD matrix between IVs is generally made using a reference panel and not the actual disease GWAS individual-level data. This is because reference panels are widely publicly available and individual-level data from many disease GWAS are not.

Using an independent reference panel to estimate LD between IVs used in MR may inflate test statistics and lead to a large false positive rate [28] if the reference population differs from the discovery (Figure 17) or if the reference panel is relatively small (Figure 3 in the main text). In the literature, only a

single solution to this problem has been documented [28], but it is only available for univariable MR with gene expression, which may be highly vulnerable to bias and its own inflation because of complex regulatory networks between the expression levels of multiple nearby genes. Additionally, this correction relies on resampling methods that cannot be scaled genome-wide because of the computational burden. There is currently no solution to this problem of inflation from misspecified LD that can be applied to multivariable MR with gene expression.

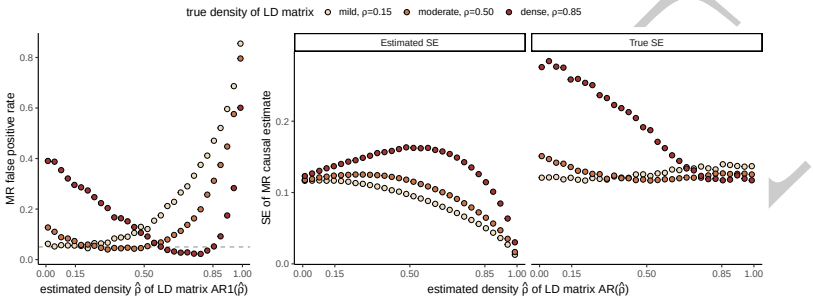


**Fig. 17** This figure demonstrates inflation in the Type I error rate of the IVW method [7] when applied to correlated IVs in simulation following a procedure similar to that described in Section 2.6.3. In this simulation, the size of the reference panel was fixed at 1,000 but the similarity in the true LD in the reference and discovery population varies. The true LD matrix $\mathbf{R}$ was of AR1($\rho$) structure and the true LD matrix in the reference panel was of AR1($\hat{\rho}$) structure. We display the false positive rate of IVW (left) for different value pairs of $\rho$ and $\hat{\rho}$, which are observed to surpass the nominal 0.05 level in many scenarios. We also display the IVW-estimated (middle) and corresponding true standard error (right) for the causal estimate in each of these scenarios. These results demonstrate that the estimated and true SEs are often unequal, which explains the inflation that is observed in the left panel.

We demonstrate that inflation in MR with gene expression is the result of relatively small reference panel sample sizes and systematic differences in genetic architecture between reference panel and discovery GWAS samples. Current methods with straightforward extensions to MVMR such as PCA and LD pruning are not guaranteed to control this inflation. We considered many potential solutions to this problem, the simulation results of which are presented in Sections 2.6.4 and 2.6.5.

### 2.6.2 Inflation correction (IFC)

In the main text, we proposed a method to correct for inflation in MR test statistics due to misspecified LD structure amongst the IVs used in MR that is presented in Figure 18. Here, we describe that method in greater detail. We propose to correct for inflation by using a data-driven approach by using the degree of inflation in surrounding null regions to adjust the standard errors for causal estimates in the target region. Here, 'target region' refers to a locus in which there is a hypothesized causal relationship between the expression of one

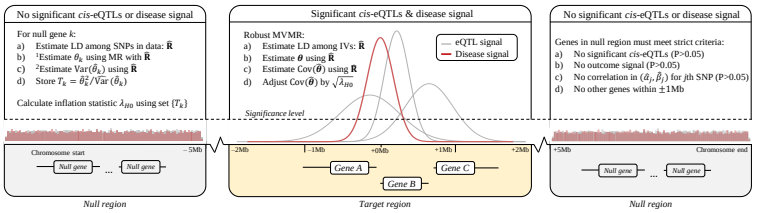


**Fig. 18** Visual description of the IFC method to correct for inflation in MVMR from misspecified LD. [1]: The causal effect $\theta_k$ can be estimated using any parametric multivariable MR method that allows the instrumental variables to be in LD. This should also be the same method that will later be used for inference in target gene regions. [2]: The variance of $\hat{\theta}_k$ should generally be estimated using robust methods, since it is unlikely to be true that the working LD matrix $\hat{\mathbf{R}}$ is exactly equal to the true LD matrix $\mathbf{R}_0$. This is because working LD matrices are typically estimated from reference samples of finite size and which may be ancestrally different from the discovery population.

or more genes and the disease trait; 'null region' refers to a locus in which there is no evidence of any association between the genetic variants and the disease trait or the expression of genes. In the truly null regions, the causal effects of gene expression on the outcome trait are each 0. This is because in Equation 7, all elements in $\boldsymbol{\alpha}$ and $\mathbf{B}$ are 0, implying that $\boldsymbol{\theta} = \mathbf{0}$ in the MR equation $\boldsymbol{\alpha} = \mathbf{B}\boldsymbol{\theta}$. By calculating inflation in these null regions, we are calculating inflation under $H_0 : \boldsymbol{\theta} = \mathbf{0}$. Any inflation that is observed in these regions is at least partially due to misspecified LD, and we assume that the same degree of inflation will be present in target regions. Under this assumption, we can adjust standard errors of causal estimates in target regions by the inflation observed in null regions. Figure 19 demonstrates that this assumption is reasonable using AD and eQTLs in blood tissue, evidenced by stable inflation across multiple null regions within chromosome 2 and across the entire genome.

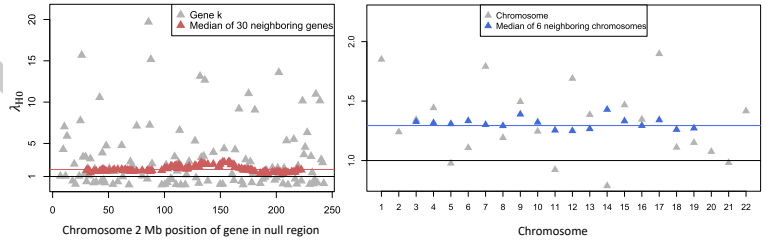


**Fig. 19** Left: observed inflation in null regions on chromosome 2. The horizontal black line is at the $\lambda_{H0} = 1$ position. The horizontal red line is at the position of the median of $\lambda_{H0}$ values across null regions on chromosome 2, which is 1.85. Right: observed inflation in null regions on all chromosomes. The horizontal black line is at the $\lambda_{H0} = 1$ position. The horizontal blue line is at the position of the median of $\lambda_{H0}$ across all chromosomes, which is at 1.29. These data are from eQTLs in blood tissue from [1] and AD GWAS summary statistics from [30].

Let $\hat{\theta}$ denote a causal estimate for the expression of a gene in a target region and $\lambda_{H0}$ denote the inflation observed in null regions. The corrected standard error estimate for $\hat{\theta}$ is the following:

$$\widehat{\mathrm{SE}}_{\lambda}(\hat{\theta}) = \widehat{\mathrm{SE}}(\hat{\theta})\sqrt{\lambda_{H0}}. \tag{45}$$

The obvious challenge lies in distinguishing truly null regions from those containing genes with extremely small causal effects. We therefore propose to use strict criteria for considering a genomic region as a null region. Firstly, it should be noted that we do not necessarily need to calculate inflation in null regions using multiple genes simultaneously in an MVMR framework. We may use one gene at a time to produce a set of causal estimate test statistics to be used in determining inflation. This is because under the condition that $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$ necessarily implies $\boldsymbol{\theta} = 0$, no negative confounding of $(\alpha_j, \beta_j)$ could exist to provide an alternative explanation for $\mathrm{E}(\hat{\theta}) \neq 0$. We therefore require that each gene to be used in determining inflation meet the following criteria: the SNPs within ±1Mb of the transcription start site are (i) not associated with the outcome, namely all P-values are greater than 0.05, (ii) not associated with the expression of the gene in the target tissue (all P>0.05), (iii) not within ±1Mb of any other genes, and (iv) the SNP associations with the outcome and gene expression are uncorrelated (P>0.05). In practice, conditions (i)-(iii) can be verified using the raw outcome phenotype and gene expression GWAS data. Condition (iv) can be verified by selected a set of SNPs for which *cis*-SNP association estimates are available for gene expression and the outcome, and calculating the empirical correlation. Generally, conditions (i)-(iv) should be satisfied after applying pruning to the raw LD matrix estimated by the reference panel. In our simulations below, we only consider SNPs that have LD coefficients less than 0.5 in absolute value.

### 2.6.3 Simulation setup

In this section, we perform simulation to demonstrate the roles of $n_{\text{ref.}}$ and $\boldsymbol{\Psi}$ in inflating test statistics corresponding to causal effect estimates made using MVMR. These simulations used real data wherever possible. These data came from the Alzheimer's disease (AD) GWAS by [30] (n=455k) for the outcome trait and from the eQTLGen Consortium [1] (n=32k) for cis-eQTLs in blood tissue. We first identified a gene regulatory network in the 2q37.1 region that contained 7 genes and selected 484 candidate IVs for these genes using the following procedures. These IVs were jointly associated with the expression of at least one of the seven genes in blood tissue and were not in LD of $r^2$>0.1 with any other SNPs ±1Mb away from the network. We then estimated LD for these IVs using the 438k non-related European individuals in the UK Biobank [18] using the PLINKv1.9 software [15]. These data respectively provided the following quantities: $\hat{\boldsymbol{\alpha}} : 484 \times 1$, $\hat{\mathbf{B}} : 484 \times 7$, and $\widetilde{\mathbf{R}}_0 : 484 \times 484$. The values in $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{B}}$ were Z-scores, i.e., association estimates divided by their standard errors. The original LD matrix for the 484 SNPs estimated using UKBB was

not positive definite. We applied LD pruning to $\widetilde{\mathbf{R}}_0$ using the threshold $|\tilde{r}_{ij}| < 0.85$ to generate the positive definite matrix $\mathbf{R}_0$. This subsetted our data from 484 to 168 SNPs.

From these data, we estimated genetic correlation between the columns of $\hat{\mathbf{B}}$ denoted as $\mathbf{S}$. We fixed heritability of gene expression at 0.05 [16] for each gene and at 0.01 for AD. We then perturbed the true LD matrix $\mathbf{R}_0$ and randomly drew it from a Wishart distribution to emulate real world conditions in which $\mathbf{R}_0$ is estimated from an external and sometimes relatively small reference panel. We did using the models:

$$\hat{\mathbf{R}} \sim \text{Wishart}(n_{\text{ref.}}, \mathbf{R}) \tag{46}$$

$$\mathbf{R} = \xi\mathbf{R}_0 + (1-\xi)\mathbf{I}_m \tag{47}$$

$$\xi \in \{0.0, 0.1, ..., 0.9, 1.0\} \tag{48}$$

$$n_{\text{ref.}} \in \{350, 450, ..., 950\}, \tag{49}$$

The minimum value in the set $n_{\text{ref.}}$ was chosen to be equal to the smallest population-specific sample size in 1000 Genomes Phase 3 [4], which corresponds to Hispanic individuals.

We therefore drew GWAS summary data for gene expression and AD from the following matrix normal distribution:

$$(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{B}}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}, \hat{\mathbf{R}}), \qquad m \times (1+p) \tag{50}$$

where

$$\boldsymbol{\Sigma} = \mathbf{D}\begin{pmatrix} 1 & \boldsymbol{\sigma}_{\mathbf{x}Y}^\top \\ \boldsymbol{\sigma}_{\mathbf{x}Y} & \mathbf{S} \end{pmatrix}\mathbf{D}, \qquad (1+p) \times (1+p) \tag{51}$$

$$\mathbf{D} = \text{diag}(0.01, 0.05, ..., 0.05), \qquad (1+p) \times (1+p) \tag{52}$$

and the quantity $\boldsymbol{\sigma}_{\mathbf{x}Y}$ was controlled to reflect the degree of causality between gene expressions $\mathbf{x}$ and AD $Y$. For example, $\boldsymbol{\sigma}_{\mathbf{x}Y} = \mathbf{0}$ implies no causality between $\mathbf{x}$ and $Y$ and was used to evaluate Type I error.

For each $(\xi, n_{\text{ref.}})$ pair, we drew $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{B}})$, applied the causal estimation methods of PCA [22], LD pruning [26], single SNP [24, 25], and our proposed IFC and recorded power and Type I error. For our IFC method, we require additional data beyond that which is provided by the 484 IVs. These data were the remaining gene expression and AD GWAS summary data on chromosome 2 that met the criteria for null regions as described in Section 2.6.2. We estimated Type I error when $\boldsymbol{\sigma}_{\mathbf{x}Y} = \mathbf{0}$ and power when $\sigma_{\mathbf{x}Y} = (\rho\sqrt{0.01 \times 0.05})$ where $\rho \in \{0.1, 0.2, 0.3\}$.

### 2.6.4 Type I error

The results of these simulations suggest that the IVW [21] method has inflated Type I error when the true LD in the reference panel is sparser than that

in the discovery population. IVW also has deflated Type I error when the reference and discovery populations have the same LD structure but the size of the reference panel is less than 1,000 individuals. Pruning at the $|r| < 0.5$ level reduced some of the Type I error inflation and deflation that was present in IVW, but did not bring Type I error to nominal levels (i.e., 0.05) in all simulation scenarios. Pruning at the $|r| < 0.3$ level removed the Type I error deflation, but not the inflation. The PCA method [22] had drastically inflated Type I error rates in all simulation scenarios. Using pruning at the $|r| < 0.3$ level then applying IFC controlled Type I error better than any other method or combination of methods and did not deflate Type I error below the nominal 0.05 level. This approach only had inflation of Type I error when the true LD matrix in the reference panel was much more dense than the true LD matrix in the discovery population, a situation which is unlikely to occur in practice. Importantly, pruning + IFC also controls Type I error when the size of the LD reference panel is small. Jackknifing methods generally still had inflated Type I error, though to a lesser extent than IVW, pruning alone, or PCA.

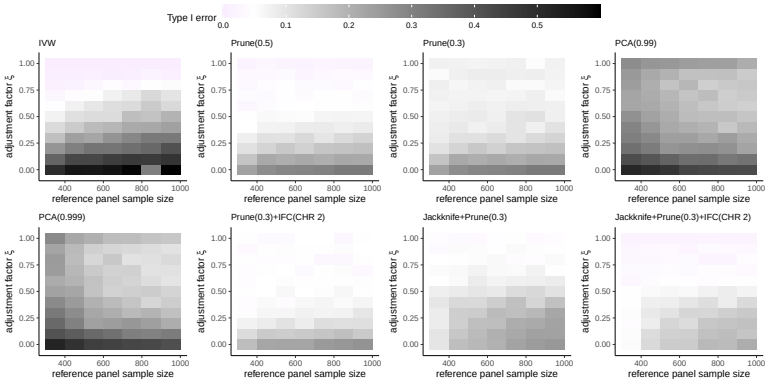


**Fig. 20** These results display Type I error for different methods for performing MR with instrumental variables that are in LD with each other. Full simulation settings are described in the text. Type I error is displayed for the first of 3 exposures. The PCA methods use univariable MR with only the first exposure; the other methods use multivariable MR with all three.

### 2.6.5 Power

We also investigated power of each method when $\boldsymbol{\sigma}_{\mathbf{x}Y} = \boldsymbol{\theta} = (0.1)$ the results of which are displayed in Figure 21. These results demonstrate that the IVW method is generally most powerful at very specific combinations of the adjustment factor $\xi$ and reference panel sample size (see Figure 21). The power of IWV is low even when the true LD structure in the reference panel is the same as in the discovery population (i.e., $\xi = 1$), but the size of the reference panel is less than 1k. Only as the size of the reference panel increases can the IVW method achieve greater power when $\xi = 1$. Pruning at the $|r| < 0.5$ the $|r| < 0.3$ thresholds have similar power which increases as $\xi$ approaches 1 and the reference panel sample size increases. These methods can achieve greater

power than IVW when the reference panel is relatively small. PCA methods actually have lower power as the reference panel sample size increases, and greater power as LD in the reference panel approaches the identity matrix, i.e. $\xi$ approaches 0. Our pruning and IFC corrective method generally has power than increases with $\xi$ approaching 1 and the reference panel sample size increasing. This approach generally has less power than alternative methods, which is the sacrifice made for controlling Type I error. Jackknife methods can generally be more powerful than all methods except pruning at $|r| < 0.5$ when LD in the reference panel is sparser than LD in the discovery population (i.e., $\xi < 0.5$). Overall, these results confirm that our corrective method of pruning and IFC does not sacrifice substantial power to achieve controlled Type I error.

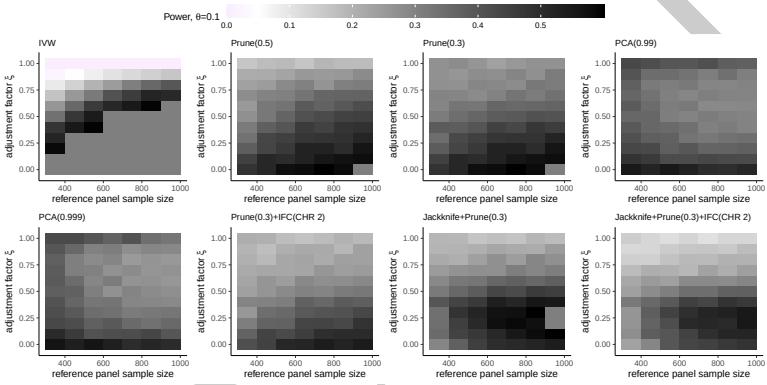


**Fig. 21** These results display power for different methods for performing MR with instrumental variables that are in LD with each other and $\theta = 0.1$. Full simulation settings are described in the text. Power is displayed for the first of three exposures. The PCA methods use univariable MR with only the first exposure; the other methods use multivariable MR with all three.

# 3 Estimating bias-correction terms

MR-Jones is an extension of the MR with unbiased estimating equations (MRBEE) method [12] to the high dimensional setting. MRBEE corrects for bias from weak instruments [31] that is introduced by measurement error in the GWAS associations. Let $(\hat{\alpha}_j, \hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jp}) = (\hat{\alpha}_j, \hat{\boldsymbol{\beta}}_j^\top)$ be a pair of associations between the $j$th IV and the outcome and expression of $p$ genes, respectively. It was shown in [12] that when estimating $\boldsymbol{\theta}$ from

$$\hat{\alpha}_j = \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\theta} + \varepsilon_j \tag{53}$$

using the standard IVW method [7], there is downward bias due to nonzero variance of $\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j$, which we denote as $\boldsymbol{\Sigma}_{W_\beta W_\beta}$. MRBEE estimates $\boldsymbol{\Sigma}_{W_\beta W_\beta}$, denoted as $\widehat{\boldsymbol{\Sigma}}_{W_\beta W_\beta}$, to correct for bias in IVW [7]. The estimate $\widehat{\boldsymbol{\Sigma}}_{W_\beta W_\beta}$ is highly precise when there are many SNPs available that have no evidence of association with the expression of any of the $p$ genes in the locus. However,

when there are relatively few SNPs available that meet this criteria for any pair of genes, the corresponding estimate in $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ may be imprecise. When the total number of SNPs available to estimate $\mathbf{\Sigma}_{W_\beta W_\beta}$ is less than 50, HORNET automatically treats the corresponding elements in $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ as missing and employs the maximum determinant method (MaxDet; [32]) to impute the missing values. In practice, $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ is converted to its corresponding correlation matrix before MaxDet is applied.

Consider that $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ is a $p \times p$ correlation matrix with 2 missing values corresponding to the correlation between measurement errors for a single pair of genes. MaxDet estimates the missing value in $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ as that which maximizes the determinant of $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ while retaining positive definiteness. The method is essentially an imputation procedure for the correlation matrix. In the real data, there may be missing measurement error correlation estimates for $k$ pairs of genes. In this case, HORNET applies MaxDet to each of $k$ sub-matrices of $\widehat{\mathbf{\Sigma}}_{W_\beta W_\beta}$ that contains only non-missing and non-imputed values.

# 4 Prioritizing tissues

In this section, we describe how to use the `tissue_chooser.py` command-line tool to identify tissues in which a pre-defined candidate set of genes have the strongest eQTL signals. From GTEx data of 54 tissues [2], we estimated heritability scores for the expression of each available gene in each tissue using all significant cis-eQTLs. These heritability scores were calculated from cis SNPs within ±1Mb and are proportional to SNP heritability. We then created a matrix in which each row was a gene, each column was a tissue, and each value was a heritability score. The `tissue_chooser.py` tool simply receives a comma-separated list of gene IDs, or a header-less file in which each row is a gene ID, and returns the tissues for which the eQTLs in GTEx v8 are the strongest. We first show how heritability scores were calculated, then demonstrate how to use the tool from the command line.

## 4.1 Heritability scores

Heritability scores are calculated in the following way. First, GTEx v8 associations between SNPs and gene expression in each tissue were calculated using fastQTL [33]. All associations that were significant at a corrected threshold were recorded and placed into the `GTEx_Analysis_v8_eQTL.tar` file available at https://gtexportal.org/home/datasets [2]. These data provided us with Z-statistics for association between the $k$ SNP and the $G$th gene in the $\mathcal{T}$th tissue, denoted here as $z_k^G(\mathcal{T})$. From these association estimates, we created the vectors $\mathbf{z}^G(\mathcal{T}) = [z_k^G(\mathcal{T})]_{k=1}^{p_G}$ of varying length $p_G$ that were gene-specific. In other words, these vectors contained all SNP-gene association estimates that were significant at a specific threshold in a specific tissue, and we created them for each gene-tissue pair. It was shown in [34] that $\text{Cov}[\mathbf{z}^G(\mathcal{T})] = \mathbf{R}$, which is

the LD matrix between the SNPs whose associations with gene expression are stored in $\mathbf{z}^G(\mathcal{T})$.

We calculated heritability scores as

$$H_S(G,\mathcal{T}) = \left[\mathbf{z}^G(\mathcal{T})\right]^\top \widehat{\mathbf{R}}^{-1}\left[\mathbf{z}^G(\mathcal{T})\right], \tag{54}$$

where $\widehat{\mathbf{R}}$ is an estimate of the corresponding LD matrix between the SNPs whose association estimates are in $\mathbf{z}^G(\mathcal{T})$, which we made using the full 1000 Genomes Phase 3 reference panel [4]. The heritability scores $H_S(G,\mathcal{T})$ were previously shown to be proportional to SNP heritability [35]. Since the tissue prioritizing tool that we present in this section is only intended to provide a ranked list of tissues in which the strongest eQTL signals for a pre-specified list of genes are, the heritability scores $H_S(G,\mathcal{T})$ are sufficient for accomplishing this task. Heritability scores for each gene-tissue pair were then stored in the file `hscores.txt`, which can be found in the `tissue_priority` directory of the HORNET software (see https://github.com/noahlorinczcomi/HORNET).

## 4.2 Running `tissue_chooser.py` to prioritize tissues

In this subsection, we demonstrated how to use our `tissue_chooser.py` tool to automatically search `hscores.txt` for tissues with the strongest eQTLs. Please see the 'Choosing tissues' branch at https://github.com/noahlorinczcomi/HORNET for a complete demonstration of how to use this tool. Briefly, this tool receives either a comma-separated list or file of gene IDs to its `--candidateGenes` flag, a 'yes' or 'no' to its `--saveAsFile` flag indicating if the results should be saved in a file in addition to being printed to the console, and the output filepath to `--outFile` if you put 'yes' to the `--candidateGenes` flag. Note, the top results will always be printed to the console, unless you want to suppress them by setting the `--printResults` flag to be 'no'. An example of output that could be generated by this tool for the *HMGCR*, *CETP*, and *FES* genes is displayed in Figure 22.

The top 10 tissues for these genes are the following:

| Tissue | Genes | GeneCount | nSignifSNPs | Maxh2Score |
|---|---|---|---|---|
| Lung | ENSG00000087237,ENSG00000182511 | 2 | 16,35 | 1208.473765 |
| Cells_Cultured_fibroblasts | ENSG00000087237,ENSG00000182511 | 2 | 6,23 | 2949.393100 |
| Muscle_Skeletal | ENSG00000113161 | 1 | 76 | 1548.315195 |
| Whole_Blood | ENSG00000182511 | 1 | 25 | 837.521937 |
| Thyroid | ENSG00000182511 | 1 | 19 | 1983.885576 |
| Pancreas | ENSG00000182511 | 1 | 18 | 1139.772883 |
| Heart_Atrial_Appendage | ENSG00000087237 | 1 | 9 | 230.349358 |
| Small_Intestine_Terminal_Ileum | ENSG00000087237 | 1 | 9 | 227.685715 |
| Stomach | ENSG00000087237 | 1 | 8 | 184.174806 |
| Adipose_Visceral_Omentum | ENSG00000113161 | 1 | 3 | 28.204502 |

**Fig. 22** This is the output of running the

```
python tissue_chooser.py --candidateGenes ENSG00000113161,ENSG00000087237,ENSG00000182511
```

command in the `HORNET` directory. All values are aggregated within tissues. `Maxh2score` is the maximum heritability score for the tissue. `nSignifSNPs` is the number of SNPs significantly associated with gene expression after adjustment for multiple comparisons in GTEx v8 [2]. `TissueCount` represents the number of genes for which the specific tissue is in the top 5 tissues with the strongest cis-eQTLs. `Genes` represents the genes for which each tissue contains one of the top 5 strongest eQTLs.

## 4.3 Limitations

This tool has the following limitations. First, this tool relies solely on GTEx v8 data for the inferences it intends to supply. Second, these heritability scores are proportional to SNP heritability but are also proportional to the true total number of causal SNPs, which in this case may not be reliably estimated from only the cis-eQTL data. The tool therefore implicitly assumes constant numbers of causal SNPs across all genes in all tissues. In this context, 'causal SNPs' refers to those SNPs that cause variation in gene expression in a specific tissue. Thirdly, any prioritization of certain tissues over others is completely agnostic to the outcome phenotype. It therefore may be true that a gene with a very strong causal effect on a disease when expressed in one particular tissue may not have strong enough eQTLs in that tissue to give it a relatively high ranking by our `tissue_chooser.py` tool. Researchers should therefore only consult this tool as one of many forms of guidance in choosing the most appropriate tissues for their analysis. Fifthly, we used strictly GTEx summary data [2] when constructing the reference data set `hscores.txt` on which the `tissue_chooser.py` relies. The GTEx v8 sample size for whole blood tissue is 670, whereas the sample size for cis-eQTLs in the eQTLGen Consortium [1] is 31k, which provides more statistical power for detecting cis-eQTLs than GTEx v8. Since whole blood tissue is generally considered in all analyses anyway, we omitted calculation of heritability scores using eQTL GWAS in whole blood from the eQTLGen Consortium.

# 5 Software

The HORNET software is available as a command line tool and desktop application for Linux, Windows, and Mac machines. Complete tutorials demonstrating how to download and use the HORNET software are present at https://github.com/noahlorinczcomi/HORNET under the 'README.md'

and 'HORNET_Desktop.md' files for the command line and desktop versions, respectively.

# References

[1] U. Võsa, A. Claringbould, H.-J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, *et al.*, "Large-scale cis- and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression," *Nature genetics*, vol. 53, no. 9, pp. 1300–1310, 2021.

[2] G. Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, *et al.*, "The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, 2015.

[3] N. de Klein, E. A. Tsai, M. Vochteloo, D. Baird, Y. Huang, C.-Y. Chen, S. van Dam, R. Oelen, P. Deelen, O. B. Bakker, *et al.*, "Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases," *Nature genetics*, vol. 55, no. 3, pp. 377–388, 2023.

[4] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[5] A. Taylor-Weiner, F. Aguet, N. J. Haradhvala, S. Gosai, S. Anand, J. Kim, K. Ardlie, E. M. Van Allen, and G. Getz, "Scaling computational genomics to millions of individuals with gpus," *Genome biology*, vol. 20, no. 1, pp. 1–5, 2019.

[6] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[7] S. Burgess and J. Bowden, "Integrating summarized data from multiple genetic variants in mendelian randomization: bias and coverage properties of inverse-variance weighted methods," *arXiv preprint arXiv:1512.04486*, 2015.

[8] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, L. Duncan, J. R. Perry, N. Patterson, E. B. Robinson, *et al.*, "An atlas of genetic correlations across human diseases and traits," *Nat. Genet.*, vol. 47, no. 11, pp. 1236–1241, 2015.

[9] Y. Yang, N. Lorincz-Comi, and X. Zhu, "Unbiased estimation and asymptotically valid inference in multivariable mendelian randomization with many weak instrumental variables," *arXiv preprint arXiv:2301.05130*,

2023.

[10] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, *et al.*, "Integration of summary data from gwas and eqtl studies predicts complex trait gene targets," *Nature genetics*, vol. 48, no. 5, pp. 481–487, 2016.

[11] K. J. Gleason, F. Yang, and L. S. Chen, "A robust two-sample mendelian randomization method integrating gwas with multi-tissue eqtl summary statistics," *bioRxiv*, pp. 2020–06, 2020.

[12] N. Lorincz-Comi, Y. Yang, G. Li, and X. Zhu, "Mrbee: A novel bias-corrected multivariable mendelian randomization method," *bioRxiv*, pp. 2023–01, 2023.

[13] X. Zhu, X. Li, R. Xu, and T. Wang, "An iterative approach to detect pleiotropy and perform mendelian randomization analysis using gwas summary statistics," *Bioinformatics*, vol. 37, no. 10, pp. 1390–1400, 2021.

[14] Y. Wu, J. Zeng, F. Zhang, Z. Zhu, T. Qi, Z. Zheng, L. R. Lloyd-Jones, R. E. Marioni, N. G. Martin, G. W. Montgomery, *et al.*, "Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits," *Nature communications*, vol. 9, no. 1, p. 918, 2018.

[15] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "Plink: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.

[16] K. G. Ouwens, R. Jansen, M. G. Nivard, J. van Dongen, M. J. Frieser, J.-J. Hottenga, W. Arindrarto, A. Claringbould, M. van Iterson, H. Mei, *et al.*, "A characterization of cis-and trans-heritability of rna-seq-based gene expression," *European Journal of Human Genetics*, vol. 28, no. 2, pp. 253–263, 2020.

[17] S. Fairley, E. Lowy-Gallego, E. Perry, and P. Flicek, "The international genome sample resource (igsr) collection of open human genomic variation resources," *Nucleic Acids Research*, vol. 48, no. D1, pp. D941–D947, 2020.

[18] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, p. e1001779, 2015.

[19] Y.-G. Choi, J. Lim, A. Roy, and J. Park, "Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage,"

*Journal of Multivariate Analysis*, vol. 171, pp. 234–249, 2019.

[20] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Stat.*, vol. 36, no. 1, pp. 199–227, 2008.

[21] S. Burgess, A. Butterworth, and S. G. Thompson, "Mendelian randomization analysis with multiple genetic variants using summarized data," *Genet. Epidemiol.*, vol. 37, no. 7, pp. 658–665, 2013.

[22] S. Burgess, V. Zuber, E. Valdes-Marquez, B. B. Sun, and J. C. Hopewell, "Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables," *Genetic epidemiology*, vol. 41, no. 8, pp. 714–725, 2017.

[23] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, G. I. of ANthropometric Traits (GIANT) Consortium, D. G. Replication, M. analysis (DIAGRAM) Consortium, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, "Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits," *Nature genetics*, vol. 44, no. 4, pp. 369–375, 2012.

[24] R. Sofat, A. D. Hingorani, L. Smeeth, S. E. Humphries, P. J. Talmud, J. Cooper, T. Shah, M. S. Sandhu, S. L. Ricketts, S. M. Boekholdt, *et al.*, "Separating the mechanism-based and off-target actions of cholesteryl ester transfer protein inhibitors with cetp gene polymorphisms," *Circulation*, vol. 121, no. 1, pp. 52–62, 2010.

[25] D. I. Swerdlow, D. Preiss, K. B. Kuchenbaecker, M. V. Holmes, J. E. Engmann, T. Shah, R. Sofat, S. Stender, P. C. Johnson, R. A. Scott, *et al.*, "Hmg-coenzyme a reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials," *The Lancet*, vol. 385, no. 9965, pp. 351–361, 2015.

[26] F. Dudbridge and P. J. Newcombe, "Accuracy of gene scores when pruning markers by linkage disequilibrium," *Human heredity*, vol. 80, no. 4, pp. 178–186, 2016.

[27] A. F. Schmidt, C. Finan, M. Gordillo-Marañón, F. W. Asselbergs, D. F. Freitag, R. S. Patel, B. Tyl, S. Chopade, R. Faraway, M. Zwierzyna, *et al.*, "Genetic drug target validation using mendelian randomisation," *Nature communications*, vol. 11, no. 1, p. 3255, 2020.

[28] L. Jiang, L. Miao, G. Yi, X. Li, C. Xue, M. J. Li, H. Huang, and M. Li, "Powerful and robust inference of complex phenotypes' causal genes with dependent expression quantitative loci by a median-based mendelian randomization," *The American Journal of Human Genetics*, vol. 109, no. 5, pp. 838–856, 2022.

[29] P. J. Newcombe, D. V. Conti, and S. Richardson, "Jam: a scalable bayesian framework for joint analysis of marginal snp effects," *Genetic epidemiology*, vol. 40, no. 3, pp. 188–201, 2016.

[30] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu, *et al.*, "Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk," *Nature genetics*, vol. 51, no. 3, pp. 404–413, 2019.

[31] I. Andrews, J. H. Stock, and L. Sun, "Weak instruments in instrumental variables regression: Theory and practice," *Annu. Rev. Econom.*, vol. 11, no. 1, 2019.

[32] D. I. Georgescu, N. J. Higham, and G. W. Peters, "Explicit solutions to correlation matrix completion problems, with an application to risk management and insurance," *Royal Society open science*, vol. 5, no. 3, p. 172348, 2018.

[33] H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau, "Fast and efficient qtl mapper for thousands of molecular phenotypes," *Bioinformatics*, vol. 32, no. 10, pp. 1479–1485, 2016.

[34] Y. Zou, P. Carbonetto, G. Wang, and M. Stephens, "Fine-mapping from summary data with the "sum of single effects" model," *PLoS Genetics*, vol. 18, no. 7, p. e1010299, 2022.

[35] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "Ld score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nat. Genet.*, vol. 47, no. 3, pp. 291–295, 2015.