

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

Noah Lorincz-Comi¹, Yihe Yang¹, Jayakrishnan Ajayakumar¹, Makaela Mews¹, ... ¹, William Bush¹
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

11	1 Missing data	3
12	1.1 Demonstration of missingness	3
13	1.2 Support from cis-eQTLs in a larger window	5
14	1.3 Multivariate Imputation	5
15	1.3.1 Procedure	5
16	1.3.2 Simulations	7
17	1.4 Power after imputing missing values	8
18	2 Accounting for LD in eQTL-MVMR	10
19	2.1 CHP bias from LD	10
20	2.1.1 Notation	10
21	2.1.2 Models	11
22	2.1.3 CHP bias in traditional MR methods	13
23	2.1.4 HORNET CHP correction	18
24	2.2 MRBEE bias-correction under LD	19
25	2.3 Heritability estimation	21
26	2.4 Source of bias in MRBEE from a misspecified LD matrix	21
27	2.5 Finding LD blocks	24
28	2.6 Misspecified LD	25

2 CONTENTS

29	2.6.1	Background	25
30	2.6.2	Inflation correction (IFC)	26
31	2.6.3	Simulation setup	28
32	2.6.4	Type I error	29
33	2.6.5	Power	30
34	3	Estimating bias-correction terms	31
35	4	Prioritizing tissues	32
36	4.1	Heritability scores	32
37	4.2	Running <code>tissue_chooser.py</code> to prioritize tissues	33
38	4.3	Limitations	34
39	5	Software	34

40 1 Missing data

41 1.1 Demonstration of missingness

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

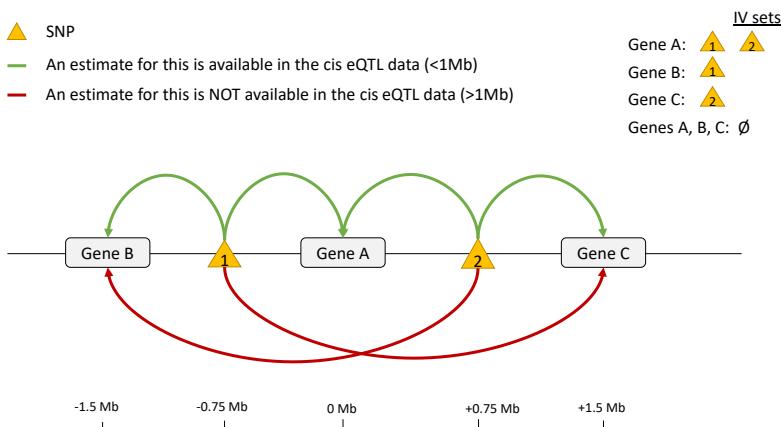


Fig. 1 This is an example representation of the data that is available in the summary-level eQTLGen [1] and GTEx [2] cis-eQTL public data sets. Only associations between SNPs and the expression of genes within $\pm 1\text{Mb}$ of those SNPs have estimates included in the available data. Since in multivariable 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.

4 CONTENTS

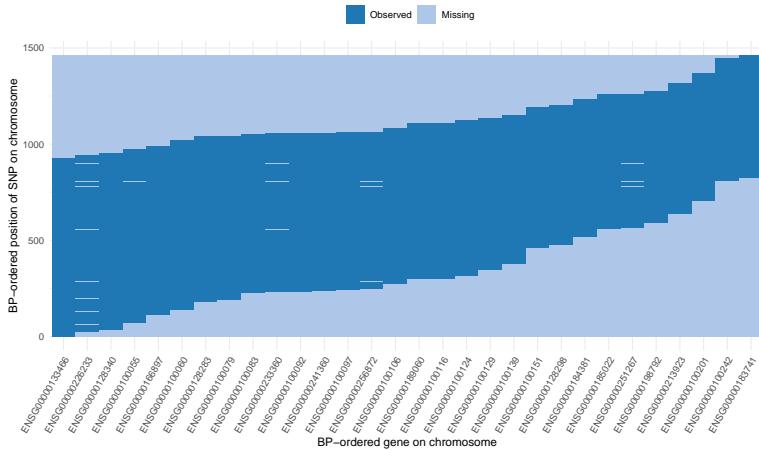


Fig. 2 Determinations of missing values for each SNP (in order by BP position on y-axis) that was used in multivariable MR with the 30 genes ordered by BP position on CHR 22 (x-axis), as an example. These genes were grouped using the procedure described in **Methods**. It was stated in **Methods** that the nature of the *cis*-eQTL data is such that only SNPs within $\pm 1\text{Mb}$ 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 $\mathcal{M}_k^1 \dots \mathcal{M}_k^{p_k}$. 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.

57 1.2 Support from cis-eQTLs in a larger window

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

69 1.3 Multivariate Imputation

70 1.3.1 Procedure

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

6 CONTENTS

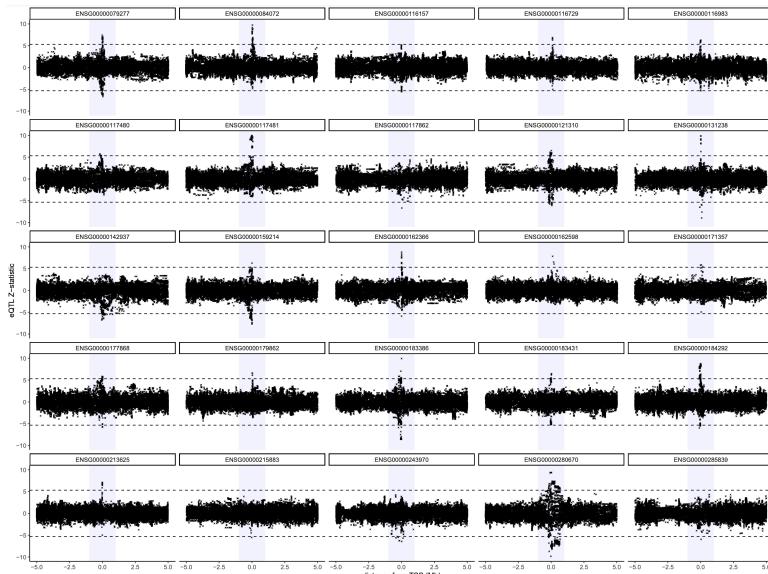


Fig. 4 Displayed are Z-statistics for eQTLs that are within $\pm 5\text{Mb}$ around 25 genes on chromosome 1 in blood tissue and non-Hispanic White individuals from the *** cohort. The shaded blue regions represent the $\pm 1\text{Mb}$ region around the transcription start site (TSS) for each gene. Horizontal 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 α . 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 5×10^{-8} outside of the $\pm 1\text{Mb}$ window from the TSS.

refers to the nonzero variance of $\hat{\beta}_{j\ell}$, the *estimated* association of the j th SNP with the ℓ 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 $\Sigma_{W_\beta W_\beta}$ denote the $p \times p$ variance-covariance matrix of $\hat{\beta}_j$, 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 $\Sigma_{W_\beta W_\beta}$ from the singular values of an initialized $\widehat{\mathbf{B}}$. This method requires the tuning parameter λ 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 $\mathcal{N}(\mathbf{B}, \mathbf{R}, \Sigma_{W_\beta W_\beta})$, 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 ρ which was in the set $\{0.0, 0.1, \dots, 0.8, 0.9\}$. The matrix $\Sigma_{W_\beta W_\beta}$ 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 λ 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 $\pm 1\text{Mb}$ 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.

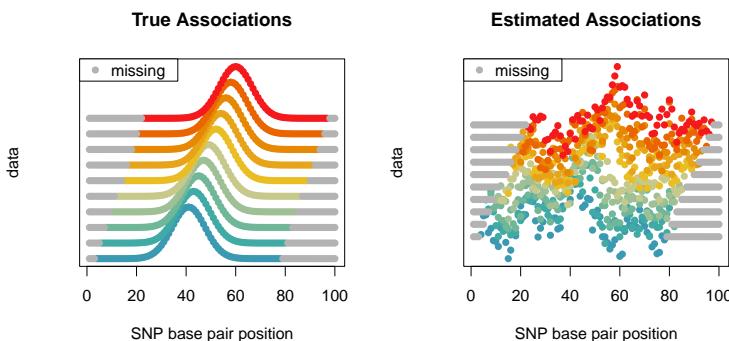


Fig. 5 This figure displays the data we generated to perform simulation using our eQTL imputation method described in Algorithm 1 in the main text. Positions on the x-axis correspond to unique SNPs. Each gene is represented by a different color, where gray always represents missing values. For each gene, 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.

8 CONTENTS

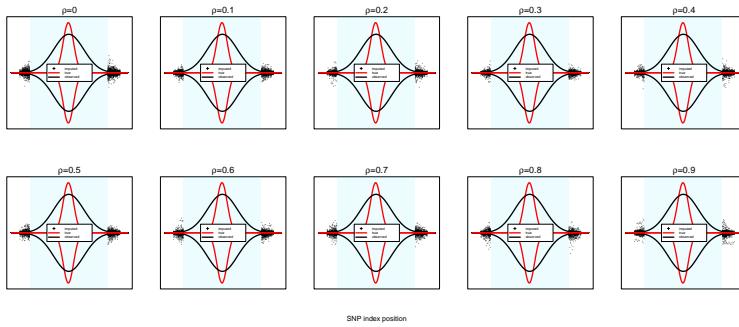


Fig. 6 These are the results of simulations in which LD structure among the 150 SNPs varied from an AR1 structure with correlation parameter $\rho = 0$ to $\rho = 0.9$. Y-axis display the relative strength of association between a SNP indexed on the x-axis and the expression of the first gene in 10 simulation. 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.

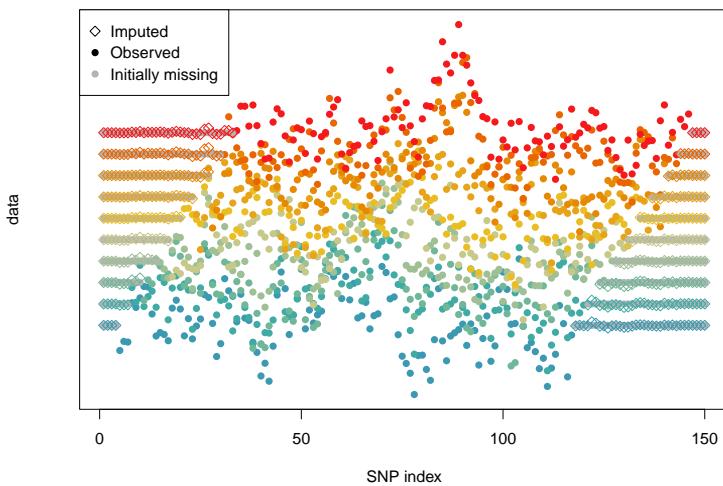


Fig. 7 These are the results of simulations which are described in the text above this figure when **R** has an AR1 structure with correlation parameter $\rho = 0.5$. These results demonstrate that our imputation method well approximates the true underlying association values when the data is missing,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

119 1.4 Power after imputing missing values

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

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

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

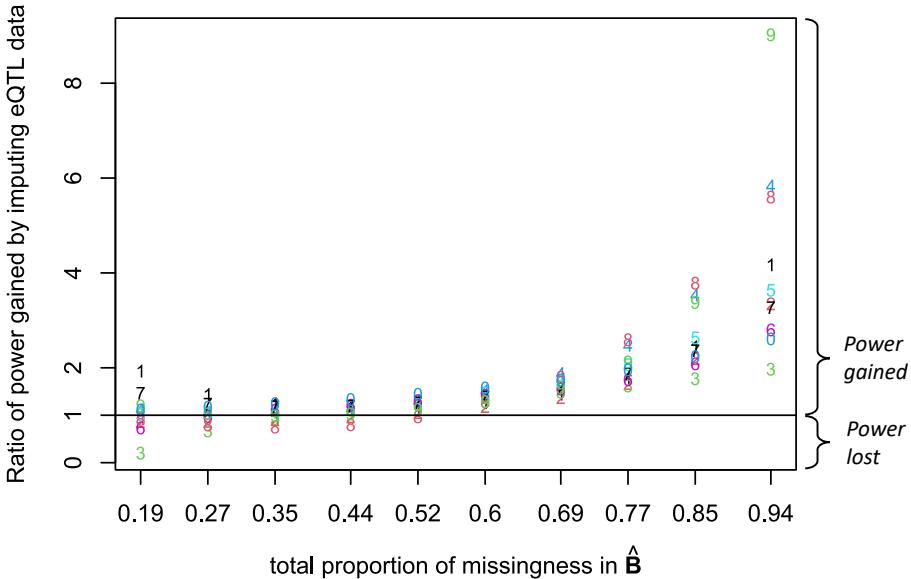


Fig. 8 This figure displays differences in the power of eQTL-MVMR using IVW with completely observed vs imputed eQTL summary statistics. The x-axis is the proportion of missingness in the working design matrix $\hat{\mathbf{B}}$ and the y-axis is the power when using imputed data divided by the power 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 $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{R})$, where $\mathbf{\Sigma} : p \times p$ 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 $\text{vec}(\mathbf{X}) \sim \mathcal{N}(\text{vec}(\boldsymbol{\mu}), \mathbf{\Sigma} \otimes \mathbf{R})$.

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 p_2 SNPs, where all of the m_1 SNPs are *cis*-eQTLs for at least one gene in their locus. Denote $\hat{\mathbf{B}}_1 = (\beta_{ij})_{i,j=1}^{m_1, p_1} : m_1 \times p_1$ and $\hat{\mathbf{B}}_2 : m_2 \times p_2$ 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 $\hat{\boldsymbol{\alpha}}_1$ 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

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

161 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:

$$\mathbf{x}_1 = \boldsymbol{\gamma}_1^\top \mathbf{g}_1 + \boldsymbol{\gamma}_2^\top \mathbf{g}_2 + \boldsymbol{\epsilon}_1 \quad (1)$$

$$= (\boldsymbol{\gamma}_1^\top + \boldsymbol{\gamma}_2^\top \boldsymbol{\lambda}_{12}^\top) \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_1, \quad (2)$$

$$= \mathbf{B}_1^\top \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_1 \quad (3)$$

$$y = \boldsymbol{\theta}^\top \mathbf{x}_1 + \boldsymbol{\pi}^\top \mathbf{g}_2 + \boldsymbol{\epsilon}_y \quad (4)$$

$$= \boldsymbol{\theta}^\top \mathbf{B}_1^\top \mathbf{g}_1 + \boldsymbol{\pi}^\top \boldsymbol{\lambda}_{12}^\top \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_y \quad (5)$$

$$= \boldsymbol{\alpha}_1^\top \mathbf{g}_1 + \tilde{\boldsymbol{\epsilon}}_y, \quad (6)$$

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

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

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

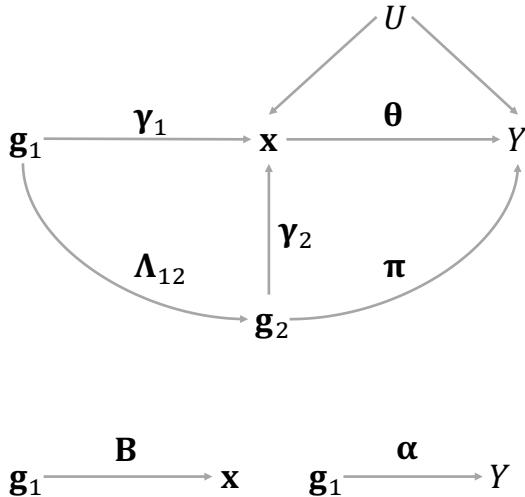


Fig. 9 DAG representing the models specified in section 3.1. $\mathbf{g}_1 : m_1 \times 1$ is a vector of SNP genotypes used as IVs in MR to estimate $\boldsymbol{\theta}$, $\mathbf{g}_2 : m_2 \times 1$ is a generic vector of genotypes for other SNPs no in \mathbf{g}_1 , $\mathbf{x} : p \times 1$ is 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, $\Lambda_{12} \neq \mathbf{0}$. 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 $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbf{B}}\boldsymbol{\theta} + \Lambda_{12}^\top \boldsymbol{\pi} + \tilde{\boldsymbol{\epsilon}}, \quad (8)$$

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 $\Lambda_{12}^\top \boldsymbol{\pi} \neq \mathbf{0}$, which may be considered unbalanced if $m^{-1}\mathbf{1}_m^\top \Lambda_{12}^\top \boldsymbol{\pi} \neq 0$ and correlated horizontal pleiotropy (CHP) if the correlation between $\Lambda_{12}^\top \boldsymbol{\pi}$ 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 $(\hat{\mathbf{B}}, \Lambda_{12}^\top \boldsymbol{\pi})$ 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 $\mathbf{B} = (\beta_k)_{k=1}^p$ and \mathbf{W}_β are random. As in [8, 9], let

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

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

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

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 $\hat{\mathbf{B}} = (\hat{\beta}_k)_{k=1}^p$, whose columns will have the distribution

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

Recall that CHP bias can arise when $\hat{\mathbf{B}}$ is correlated with $\Lambda_{12}^\top \pi$. Let $\Lambda_{12}^\top (\tau^\top \otimes \mathbf{R}_{12}^\top) := \text{Cov}(\text{vec}[\mathbf{B}], \Lambda_{12}^\top \pi)$ where $\tau^\top : p \times 1$ represents genetic covariance between π and the columns of \mathbf{B} . For example, consider $\text{Cov}(\beta_k, \pi) := \tau_k \mathbf{R}_{12}^\top = [\text{E}(\beta_{jk} \pi_s r_{js})]_{j,s=1}^{m_1, m_2}$. Since this covariance has a kronecker product structure, it can be zero if either one of two conditions are met, namely if (i) $\tau = \mathbf{0}$ 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 π 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 $\mathbf{R}_{k,-k}$, $\Lambda_{k,-k} \approx \mathbf{R}_{k,-k}^\top \mathbf{R}_k^{-1}$, and π_{-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:

$$I_1 = \|\mathbf{R}_k^{-1/2} \Lambda_{k,-k}^\top \pi_{-k}\|_2^2, \quad (12)$$

$$I_2 = \frac{1}{m_k} \mathbf{1}_{m_k}^\top \mathbf{R}_k^{-1/2} \Lambda_{k,-k}^\top \pi_{-k}, \quad (13)$$

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

14 CONTENTS

where θ_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:

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

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

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

where

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

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

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

and

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

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 θ_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 θ_k , we also applied the following restrictions on the IV set: (i) P-value for association with gene expression less than 5×10^{-5} , (ii) absolute LD between SNPs used to instrument expression of the k th gene less than 0.9, (iii) ≥ 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

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

205 Under these restrictions to obtain a valid IV set, univariable MR using
 206 bias-corrected SMR (i.e., single-exposure MRBEE [9, 12]) could only be per-
 207 formed for 194 of the 752 (25%) total genes with *cis*-eQTLs on CHR 19. This
 208 is another major limitation of univariable MR - the valid IV set can be reduced
 209 so small that causal estimation becomes unreliable and therefore should not
 210 even be performed. Figures 10, 11, 12, and 13 provide some inference for I_1 ,
 211 I_2 , and I_3 , respectively. These results indicate substantial nonzero unbal-
 212 anced horizontal pleiotropy across CHR 19. Subsequently, there is widespread
 213 evidence of differences in causal estimates made using univariable vs multi-
 214 variable causal estimates, suggesting the presence off OV bias that may be due
 215 to CHP. We found that 37.6% of genes tested using univariable MR on CHR
 216 19 (73/194) had evidence ($P < 5 \times 10^{-5}$) of nonzero horizontal pleiotropy
 217 ($I_1 \neq 0$), 13.7% of which had evidence of imbalance ($I_2 \neq 0$), and 48.2% of
 218 genes had multivariable causal estimates that differed from univariable causal
 219 estimates ($I_3 \neq 0$ where the test was available [see Figure 13 caption]). See
 220 the corresponding figure captions for more details.

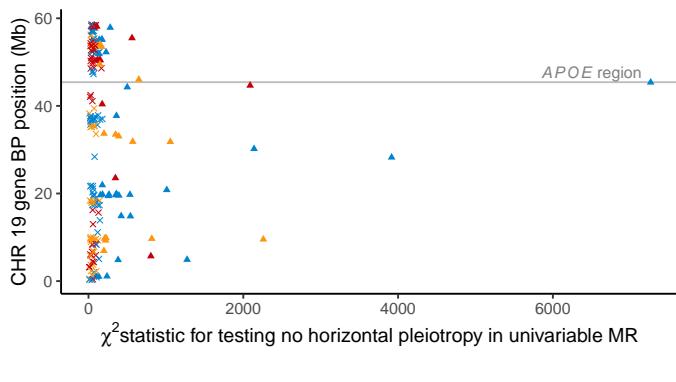


Fig. 10 These results display χ^2 test statistics for testing $H_0 : I_1 = 0$ as stated above. Each point represents a gene. This is a test for nonzero horizontal pleiotropy in the IV set in univariable MR using MRBEE [9, 12], which can be considered a version of SMR [10, 14] corrected for bias from 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).

16 CONTENTS

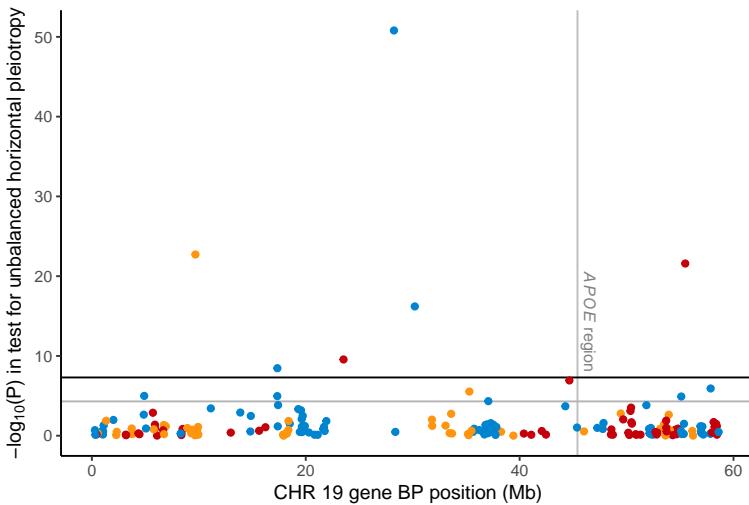


Fig. 11 This is a test for unbalanced (i.e., nonzero mean) horizontal pleiotropy in univariable MR with MRBEE [9, 12], which can be considered a version of SMR [10, 14] corrected for bias from horizontal pleiotropy, weak instruments, sample overlap, and measurement/estimation error. Each point represents 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).

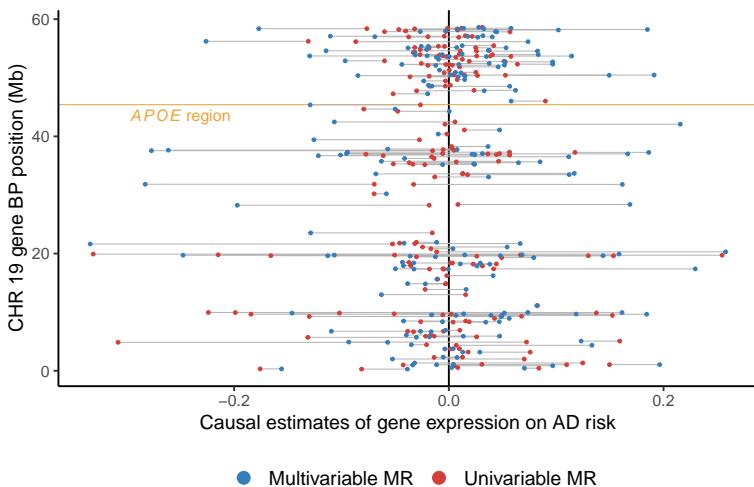


Fig. 12 These results display differences between causal estimates made for each gene on chromosome (CHR) 19 using univariable MR vs multivariable MR. Each pair of points (paired by horizontal grey lines) corresponds to a single gene. Causal estimates were made using MRBEE [9, 12], which can be considered 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 omitted 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).

18 CONTENTS

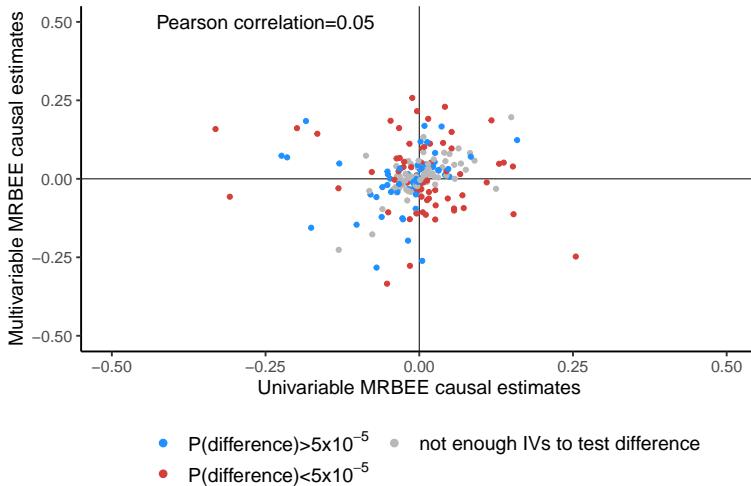


Fig. 13 These results display the bivariate association between univariable MR causal estimates and indications of the significance in testing $H_0 : I_3 = 0$ from the text above. Red and blue points respectively represent genes for which this null hypothesis was rejected and not rejected at $P < 5 \times 10^{-5}$. 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 $\hat{\sigma}_\theta^2$ (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.

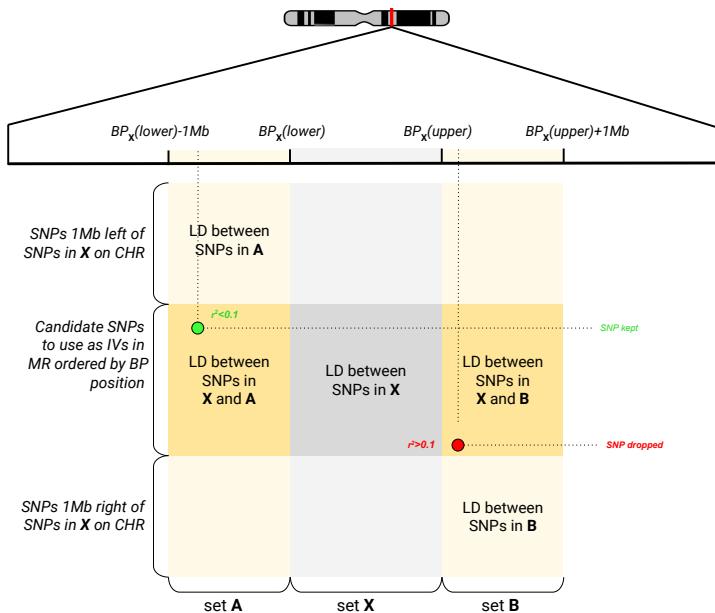


Fig. 14 Visual depiction of how the final IV sets were pruned based on their LD with non-IVs within surrounding Mb windows of defined size (1Mb as the example in this figure). It is shown in section ?? that horizontal pleiotropy bias may be present in MR if the IVs used are in LD with SNPs that are conditionally 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 **B**. Let $\hat{\mathbf{B}} = (\hat{\beta}_j)_{j=1}^m = (\hat{\beta}_{jk})_{j=1,k=1}^{m,p}$ 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 $\mathbf{R} = (r_{js})_{j,s=1}^m$, 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 $(\hat{\beta}_j, \hat{\alpha}_j)$ as in [9, 12] in the following way:

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

where the errors in our measurements of $(\hat{\beta}_j, \hat{\alpha}_j)$ were due only to sampling error introduced by finite GWAS sample sizes, and (β_j, α_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

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

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 $-E(\mathbf{W}_\beta^\top \mathbf{W}_\beta) + E(\mathbf{W}_\beta^\top \mathbf{w}_\alpha)$. Under the normality assumption in Equation 22,

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

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

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

and

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

The proof for $E(\mathbf{W}_\beta^\top \mathbf{P} \mathbf{W}_\beta)$ 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

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

It follows that

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

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

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

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

227 2.3 Heritability estimation

Assume the random model for $\beta_k : M \times 1$ 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_k^2 . In our analyses, all association estimates were standardized by estimated standard error in GWAS such that $\hat{\zeta}_k \approx \sqrt{n_k} \hat{\beta}_k$ for GWAS sample size n_k and $\hat{\zeta}_k$ was the unit of analysis. Under the assumptions in model 9,

$$h_k^2 = \left[\frac{\text{E}(\hat{\zeta}_k^\top \hat{\zeta}_k)}{M} - 1 \right] \frac{\tilde{m}_k}{n_k}. \quad (30)$$

This is seen immediately from the following result

$$\text{E}(\hat{\zeta}_k^\top \hat{\zeta}_k) = \text{trace}\left(n_k \left[\frac{h_k^2}{\tilde{m}_k} \mathbf{I} + \frac{1}{n_k} \mathbf{R} \right]\right) \quad (31)$$

from the original model in Equation 11. A natural estimate of h_k^2 is

$$\hat{h}_k^2 = \left[\frac{\hat{\zeta}_k^\top \hat{\zeta}_k - \frac{1}{n_k}}{M} - 1 \right] \frac{\hat{m}_k}{n_k} \quad (32)$$

228 where the $-1/n_k$ term is introduced as a measurement/estimation error bias-
 229 correction [9, 12]. In practice, \tilde{m}_k is rarely known and so must be estimated
 230 from the data. We estimated this quantity using a procedure similar to that
 231 used by PLINK [16] where we let $2\hat{m}_k$ be the number of SNPs of the total
 232 M with association $P < 5 \times 10^{-5}$ for $\hat{\zeta}_{jk}$ that are independent ($r^2 < 0.05$)
 233 of all other $M - 1$ SNPs for the gene group. We assume the factor 2 on \tilde{m}_k
 234 consistent with results in [17? ?] that *cis*-eQTLs explain approximately 1/3rd
 235 of the SNP heritability and *trans*-eQTLs explain the rest. Note that whether
 236 you assume a random or fixed effects model for β_k , the result in Equation 30
 237 is the same. To see this, let $\hat{\zeta}_k \sim N(\zeta_k, n_k^{-1} \mathbf{R})$ and use the same technique as
 238 in Equation 31 then rearrange to arrive at the result in Equation 32.

239 2.4 Source of bias in MRBEE from a misspecified LD 240 matrix

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

22 CONTENTS

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 [18]. 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

$$\hat{\theta}_{\text{MRBEE}} = \left(\hat{\mathbf{B}}^\top \hat{\mathbf{R}}^{-1} \hat{\mathbf{B}} - m \Sigma_{W_\beta W_\beta} \right)^{-1} \hat{\mathbf{B}}^\top \hat{\mathbf{R}}^{-1} \hat{\alpha}. \quad (33)$$

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 $\hat{\alpha} \sim \mathcal{N}(\boldsymbol{\alpha}, \mathbf{R})$. If $\hat{\mathbf{R}} = \mathbf{R}$, then MRBEE is not biased by $\hat{\mathbf{R}}$. This follows from Equation 24 under the assumption that $\text{Var}(\hat{\alpha}) = \mathbf{R}$. However, if $\hat{\mathbf{R}} \neq \mathbf{R}$ then the bias-correction to IVW that MRBEE makes is not correct and therefore $\hat{\theta}_{\text{MRBEE}}$ 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

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

which MRBEE assumes to be $-m(\Sigma_{W_\beta W_\beta} \boldsymbol{\theta} - \sigma_{W_\beta w_\alpha})$. 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

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

$$\Sigma = \mathbf{D} \begin{pmatrix} 1.0 & 0.2 & 0.2 \\ 0.2 & 1.0 & 0.5 \\ 0.2 & 0.5 & 1.0 \end{pmatrix} \mathbf{D} \quad (36)$$

$$\mathbf{D} = 0.1 \mathbf{I}_3 \quad (37)$$

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

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

$$\hat{\mathbf{R}} = \text{AR1}(\xi), \quad \xi \in \{0.0, 0.1, \dots, 0.8, 0.9\} \quad (40)$$

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

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

(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 ξ 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 to 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, except 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.

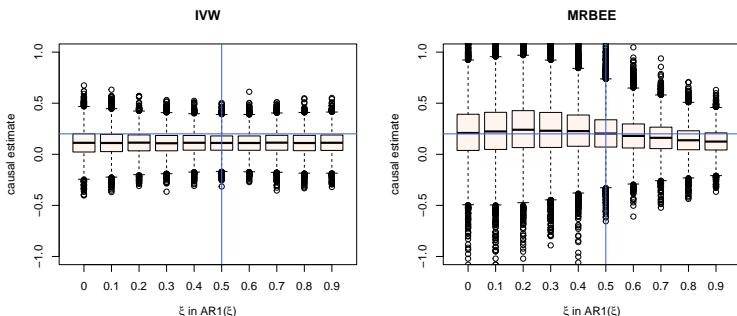


Fig. 15 These are the results of 10,000 simulations, the settings for which are fully described in the text above. Displayed are boxplots of causal estimates made by IVW (left) and MRBEE (right) in different scenarios of assumed LD structure. The horizontal blue lines are positioned at the true causal effect, 0.2. The vertical blue lines are positioned at the value of ρ 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 (τ, b) are hyperparameters 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 [16] 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 τ 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

$$P_{(\tau, b, k)} = -\log \|\boldsymbol{\Omega}_1\|_F \log p_1 + \log \|\boldsymbol{\Omega}_{-1}\|_F \log(n - p_1) + \log k, \quad (44)$$

where k is the number of blocks to use and $\boldsymbol{\Omega}_1 = (r_{ij}^2)_{i,j=1}^{p_1}$. In practice, (τ, 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 `noahlorinczcom/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 [19] and was converted to a positive definite matrix using the method of [20]. 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 [21].

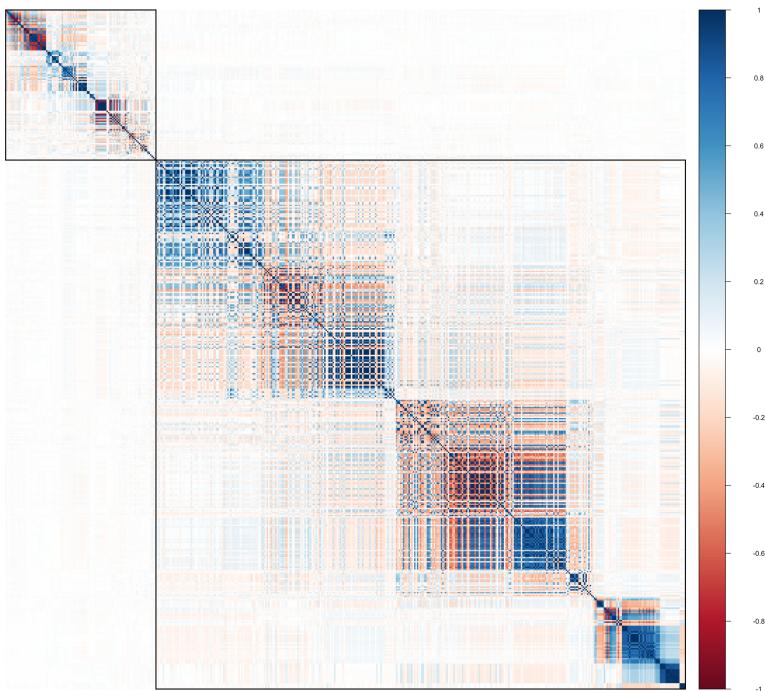


Fig. 16 The LD matrix for the 484 SNPs displayed here was estimated from 438k non-related European individuals in the UK Biobank. The black lines are generated by our block-finding algorithm outlined in the text.

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 [22], principal components analysis (PCA) [23], the conditional and joint (CoJo) algorithm [24], single-SNP [25, 26], LD pruning [27, 28], effective-median [29], and the JAM algorithm (joint analysis of marginal summary statistics) [30]. 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 [29] 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

26 CONTENTS

single solution to this problem has been documented [29], 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.

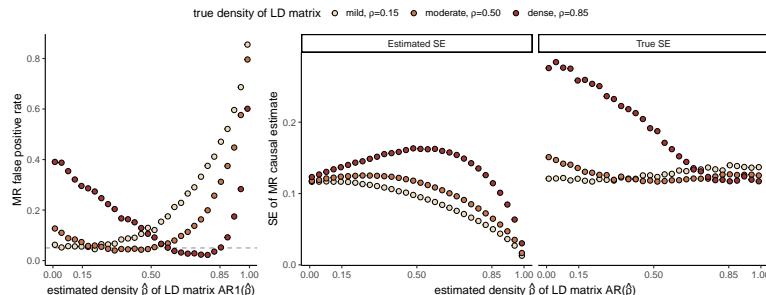


Fig. 17 This figure demonstrates inflation in the Type I error rate of the IVW method [7] when applied to correlated IVs in simulation following a procedure similar to that described in Section 2.6.3. In this simulation, the size of the reference panel was fixed at 1,000 but the similarity in the true 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 ρ 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

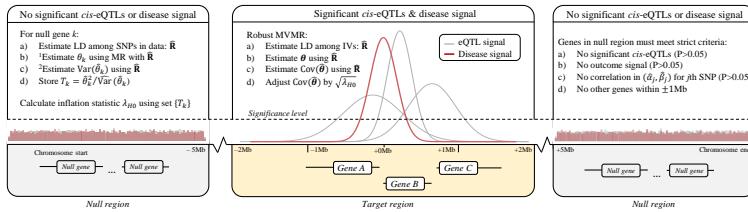


Fig. 18 Visual description of the IFC method to correct for inflation in MVMR from misspecified LD. [1]: The causal effect θ_k can be estimated using any parametric multivariable MR method that allows the instrumental variables to be in LD. This should also be the same method that will later be used 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.

340 or more genes and the disease trait; ‘null region’ refers to a locus in which there
 341 is no evidence of any association between the genetic variants and the disease
 342 trait or the expression of genes. In the truly null regions, the causal effects of
 343 gene expression on the outcome trait are each 0. This is because in Equation 7,
 344 all elements in α and \mathbf{B} are 0, implying that $\boldsymbol{\theta} = \mathbf{0}$ in the MR equation $\alpha = \mathbf{B}\boldsymbol{\theta}$.
 345 By calculating inflation in these null regions, we are calculating inflation under
 346 $H_0 : \boldsymbol{\theta} = \mathbf{0}$. Any inflation that is observed in these regions is at least partially
 347 due to misspecified LD, and we assume that the same degree of inflation will
 348 be present in target regions. Under this assumption, we can adjust standard
 349 errors of causal estimates in target regions by the inflation observed in null
 350 regions. Figure 19 demonstrates that this assumption is reasonable using AD
 351 and eQTLs in blood tissue, evidenced by stable inflation across multiple null
 352 regions within chromosome 2 and across the entire genome.

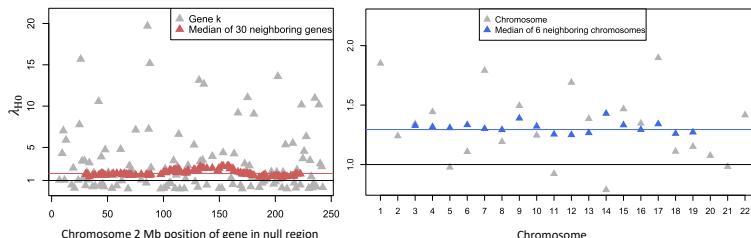


Fig. 19 Left: observed inflation in null regions on chromosome 2. The horizontal black line is at the $\lambda_{H0} = 1$ position. The horizontal red line is at the position of the median of λ_{H0} values across null regions on chromosome 2, which is 1.85. Right: observed inflation in null regions on 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 λ_{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 [15].

28 CONTENTS

Let $\hat{\theta}$ denote a causal estimate for the expression of a gene in a target region and λ_{H_0} denote the inflation observed in null regions. The corrected standard error estimate for $\hat{\theta}$ is the following:

$$\widehat{SE}_\lambda(\hat{\theta}) = \widehat{SE}(\hat{\theta})\sqrt{\lambda_{H_0}}. \quad (45)$$

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 $\alpha = \beta = \mathbf{0}$ necessarily implies $\theta = 0$, no negative confounding of (α_j, β_j) could exist to provide an alternative explanation for $E(\hat{\theta}) \neq 0$. We therefore require that each gene to be used in determining inflation meet the following criteria: the SNPs within $\pm 1\text{Mb}$ 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 $\pm 1\text{Mb}$ 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 selecting 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 Ψ 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 [15] ($n=455k$) 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 $\pm 1\text{Mb}$ away from the network. We then estimated LD for these IVs using the 438k non-related European individuals in the UK Biobank [19] using the PLINKv1.9 software [16]. These data respectively provided the following quantities: $\hat{\alpha} : 484 \times 1$, $\hat{\mathbf{B}} : 484 \times 7$, and $\tilde{\mathbf{R}}_0 : 484 \times 484$. The values in $\hat{\alpha}$ 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

391 not positive definite. We applied LD pruning to $\tilde{\mathbf{R}}_0$ using the threshold $|\tilde{r}_{ij}| <$
 392 0.85 to generate the positive definite matrix \mathbf{R}_0 . This subsetted our data from
 393 484 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 [17] 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:

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

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

$$\xi \in \{0.0, 0.1, \dots, 0.9, 1.0\} \quad (48)$$

$$n_{\text{ref.}} \in \{350, 450, \dots, 950\}, \quad (49)$$

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

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

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

where

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

$$\mathbf{D} = \text{diag}(0.01, 0.05, \dots, 0.05), \quad (1 + p) \times (1 + p) \quad (52)$$

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

400 For each $(\xi, n_{\text{ref.}})$ pair, we drew $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{B}})$, applied the causal estimation meth-
 401 ods of PCA [23], LD pruning [27], single SNP [25, 26], and our proposed IFC
 402 and recorded power and Type I error. For our IFC method, we require addi-
 403 tional data beyond that which is provided by the 484 IVs. These data were
 404 the remaining gene expression and AD GWAS summary data on chromosome
 405 2 that met the criteria for null regions as described in Section 2.6.2. We esti-
 406 mated Type I error when $\boldsymbol{\sigma}_{\mathbf{x}Y} = \mathbf{0}$ and power when $\boldsymbol{\sigma}_{\mathbf{x}Y} = (\rho \sqrt{0.01 \times 0.05})$
 407 where $\rho \in \{0.1, 0.2, 0.3\}$.

408 2.6.4 Type I error

409 The results of these simulations suggest that the IVW [22] method has inflated
 410 Type I error when the true LD in the reference panel is sparser than that

30 CONTENTS

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 [23] 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.

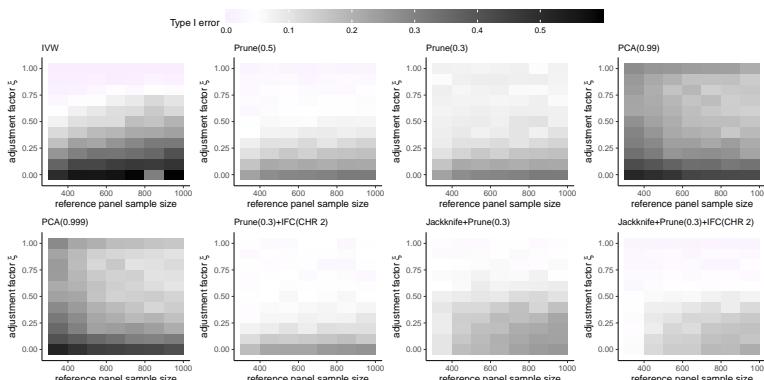


Fig. 20 These results display Type I error for different methods for performing MR with instrumental variables that are in LD with each other. Full simulation settings are described in the text. Type I error is displayed for the first of 3 exposures. The PCA methods use univariable MR with only the first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

427 **2.6.5 Power**

We also investigated power of each method when $\sigma_{XY} = \theta = (0.1)$ 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 ξ and reference panel sample size (see Figure 21). The power of IVW 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 ξ 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. ξ approaches 0. Our pruning and IFC corrective method generally has power that increases with ξ 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.

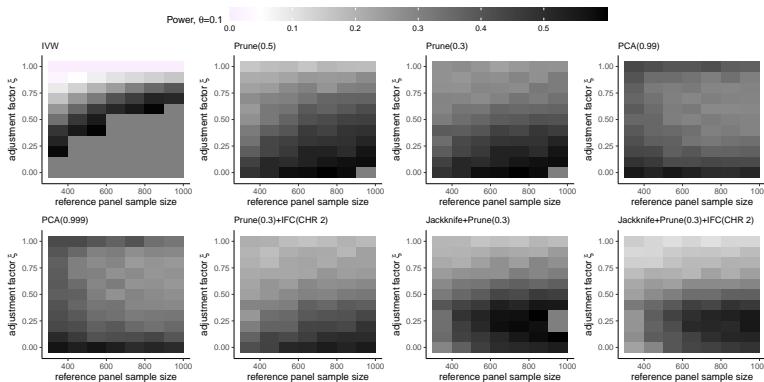


Fig. 21 These results display power for different methods for performing MR with instrumental variables that are in LD with each other and $\theta = 0.1$. Full simulation settings are described in the text. Power is displayed for the first of three exposures. The PCA methods use univariable MR with only 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 $(\hat{\alpha}_j, \hat{\beta}_{j1}, \hat{\beta}_{j2}, \dots, \hat{\beta}_{jp}) = (\hat{\alpha}_j, \hat{\beta}_j^\top)$ 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

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

using the standard IVW method [7], there is downward bias due to nonzero variance of $\hat{\beta}_j - \hat{\beta}_j$, which we denote as $\Sigma_{W_\beta W_\beta}$. MRBEE estimates $\Sigma_{W_\beta W_\beta}$, denoted as $\widehat{\Sigma}_{W_\beta W_\beta}$, to correct for bias in IVW [7]. The estimate $\widehat{\Sigma}_{W_\beta W_\beta}$ 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,

32 CONTENTS

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

462 Consider that $\widehat{\Sigma}_{W_\beta W_\beta}$ is a $p \times p$ correlation matrix with 2 missing values cor-
 463 responding to the correlation between measurement errors for a single pair of
 464 genes. MaxDet estimates the missing value in $\widehat{\Sigma}_{W_\beta W_\beta}$ as that which maximizes
 465 the determinant of $\widehat{\Sigma}_{W_\beta W_\beta}$ while retaining positive definiteness. The method
 466 is essentially an imputation procedure for the correlation matrix. In the real
 467 data, there may be missing measurement error correlation estimates for k pairs
 468 of genes. In this case, HORNET applies MaxDet to each of k sub-matrices of
 469 $\widehat{\Sigma}_{W_\beta W_\beta}$ that contains only non-missing and non-imputed values.

470

4 Prioritizing tissues

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

483

4.1 Heritability scores

484 Heritability scores are calculated in the following way. First, GTEx v8 associa-
 485 tions between SNPs and gene expression in each tissue were calculated using
 486 fastQTL [33]. All associations that were significant at a corrected threshold
 487 were recorded and placed into the `GTEx_Analysis_v8_eQTL.tar` file available
 488 at <https://gtexportal.org/home/datasets> [2]. These data provided us with Z-
 489 statistics for association between the k SNP and the G th gene in the \mathcal{T} th
 490 tissue, denoted here as $z_k^G(\mathcal{T})$. From these association estimates, we created
 491 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.
 492 In other words, these vectors contained all SNP-gene association estimates that
 493 were significant at a specific threshold in a specific tissue, and we created them
 494 for each gene-tissue pair. It was shown in [34] that $\text{Cov}[\mathbf{z}^G(\mathcal{T})] = \mathbf{R}$, which is

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

We calculated heritability scores as

$$H_S(G, \mathcal{T}) = [\mathbf{z}^G(\mathcal{T})]^\top \hat{\mathbf{R}}^{-1} [\mathbf{z}^G(\mathcal{T})], \quad (54)$$

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

507 4.2 Running `tissue_chooser.py` to prioritize tissues

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

34 CONTENTS

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'

547 and ‘HORNET/Desktop.md’ files for the command line and desktop versions,
548 respectively.

549 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*,

36 CONTENTS

584 2023.

- 585 [10] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W.
 586 Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, *et al.*, “Integration
 587 of summary data from gwas and eqtl studies predicts complex trait
 588 gene targets,” *Nature genetics*, vol. 48, no. 5, pp. 481–487, 2016.
- 589 [11] K. J. Gleason, F. Yang, and L. S. Chen, “A robust two-sample mendelian
 590 randomization method integrating gwas with multi-tissue eqtl summary
 591 statistics,” *bioRxiv*, pp. 2020–06, 2020.
- 592 [12] N. Lorincz-Comi, Y. Yang, G. Li, and X. Zhu, “Mrbee: A novel
 593 bias-corrected multivariable mendelian randomization method,” *bioRxiv*,
 594 pp. 2023–01, 2023.
- 595 [13] X. Zhu, X. Li, R. Xu, and T. Wang, “An iterative approach to detect
 596 pleiotropy and perform mendelian randomization analysis using gwas
 597 summary statistics,” *Bioinformatics*, vol. 37, no. 10, pp. 1390–1400, 2021.
- 598 [14] Y. Wu, J. Zeng, F. Zhang, Z. Zhu, T. Qi, Z. Zheng, L. R. Lloyd-Jones,
 599 R. E. Marioni, N. G. Martin, G. W. Montgomery, *et al.*, “Integrative
 600 analysis of omics summary data reveals putative mechanisms underlying
 601 complex traits,” *Nature communications*, vol. 9, no. 1, p. 918, 2018.
- 602 [15] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams,
 603 S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu, *et al.*,
 604 “Genome-wide meta-analysis identifies new loci and functional pathways
 605 influencing alzheimer’s disease risk,” *Nature genetics*, vol. 51, no. 3,
 606 pp. 404–413, 2019.
- 607 [16] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender,
 608 J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, “Plink: a tool set
 609 for whole-genome association and population-based linkage analyses,”
 610 *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.
- 611 [17] K. G. Ouwens, R. Jansen, M. G. Nivard, J. van Dongen, M. J. Frieser,
 612 J.-J. Hottenga, W. Arindrarto, A. Claringbould, M. van Iterson, H. Mei,
 613 *et al.*, “A characterization of cis-and trans-heritability of rna-seq-based
 614 gene expression,” *European Journal of Human Genetics*, vol. 28, no. 2,
 615 pp. 253–263, 2020.
- 616 [18] S. Fairley, E. Lowy-Gallego, E. Perry, and P. Flicek, “The international
 617 genome sample resource (igsr) collection of open human genomic variation
 618 resources,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D941–D947, 2020.
- 619 [19] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh,
 620 P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, “Uk biobank: an open

- 621 access resource for identifying the causes of a wide range of complex dis-
622 eases of middle and old age,” *PLoS Med.*, vol. 12, no. 3, p. e1001779,
623 2015.
- 624 [20] Y.-G. Choi, J. Lim, A. Roy, and J. Park, “Fixed support positive-
625 definite modification of covariance matrix estimators via linear shrinkage,”
626 *Journal of Multivariate Analysis*, vol. 171, pp. 234–249, 2019.
- 627 [21] P. J. Bickel and E. Levina, “Regularized estimation of large covariance
628 matrices,” *Ann. Stat.*, vol. 36, no. 1, pp. 199–227, 2008.
- 629 [22] S. Burgess, A. Butterworth, and S. G. Thompson, “Mendelian random-
630 ization analysis with multiple genetic variants using summarized data,”
631 *Genet. Epidemiol.*, vol. 37, no. 7, pp. 658–665, 2013.
- 632 [23] S. Burgess, V. Zuber, E. Valdes-Marquez, B. B. Sun, and J. C. Hopewell,
633 “Mendelian randomization with fine-mapped genetic data: choosing from
634 large numbers of correlated instrumental variables,” *Genetic epidemiol-
635 ogy*, vol. 41, no. 8, pp. 714–725, 2017.
- 636 [24] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, G. I. of ANthropo-
637 metric Traits (GIANT) Consortium, D. G. Replication, M. analysis
638 (DIAGRAM) Consortium, P. A. Madden, A. C. Heath, N. G. Martin,
639 G. W. Montgomery, *et al.*, “Conditional and joint multiple-snp analysis of
640 gwas summary statistics identifies additional variants influencing complex
641 traits,” *Nature genetics*, vol. 44, no. 4, pp. 369–375, 2012.
- 642 [25] R. Sofat, A. D. Hingorani, L. Smeeth, S. E. Humphries, P. J. Talmud,
643 J. Cooper, T. Shah, M. S. Sandhu, S. L. Ricketts, S. M. Boekholdt, *et al.*,
644 “Separating the mechanism-based and off-target actions of cholesteroyl
645 ester transfer protein inhibitors with ctep gene polymorphisms,” *Circula-
646 tion*, vol. 121, no. 1, pp. 52–62, 2010.
- 647 [26] D. I. Swerdlow, D. Preiss, K. B. Kuchenbaecker, M. V. Holmes, J. E.
648 Engmann, T. Shah, R. Sofat, S. Stender, P. C. Johnson, R. A. Scott,
649 *et al.*, “Hmg-coenzyme a reductase inhibition, type 2 diabetes, and body-
650 weight: evidence from genetic analysis and randomised trials,” *The Lancet*,
651 vol. 385, no. 9965, pp. 351–361, 2015.
- 652 [27] F. Dudbridge and P. J. Newcombe, “Accuracy of gene scores when prun-
653 ing markers by linkage disequilibrium,” *Human heredity*, vol. 80, no. 4,
654 pp. 178–186, 2016.
- 655 [28] A. F. Schmidt, C. Finan, M. Gordillo-Marañón, F. W. Asselbergs, D. F.
656 Freitag, R. S. Patel, B. Tyl, S. Chopade, R. Faraway, M. Zwierzyna, *et al.*,
657 “Genetic drug target validation using mendelian randomisation,” *Nature
658 communications*, vol. 11, no. 1, p. 3255, 2020.

38 CONTENTS

- 659 [29] L. Jiang, L. Miao, G. Yi, X. Li, C. Xue, M. J. Li, H. Huang, and M. Li,
660 “Powerful and robust inference of complex phenotypes’ causal genes with
661 dependent expression quantitative loci by a median-based mendelian ran-
662 domization,” *The American Journal of Human Genetics*, vol. 109, no. 5,
663 pp. 838–856, 2022.
- 664 [30] P. J. Newcombe, D. V. Conti, and S. Richardson, “Jam: a scalable
665 bayesian framework for joint analysis of marginal snp effects,” *Genetic*
666 *epidemiology*, vol. 40, no. 3, pp. 188–201, 2016.
- 667 [31] I. Andrews, J. H. Stock, and L. Sun, “Weak instruments in instrumental
668 variables regression: Theory and practice,” *Annu. Rev. Econom.*, vol. 11,
669 no. 1, 2019.
- 670 [32] D. I. Georgescu, N. J. Higham, and G. W. Peters, “Explicit solutions
671 to correlation matrix completion problems, with an application to risk
672 management and insurance,” *Royal Society open science*, vol. 5, no. 3,
673 p. 172348, 2018.
- 674 [33] H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau,
675 “Fast and efficient qtl mapper for thousands of molecular phenotypes,”
676 *Bioinformatics*, vol. 32, no. 10, pp. 1479–1485, 2016.
- 677 [34] Y. Zou, P. Carbonetto, G. Wang, and M. Stephens, “Fine-mapping from
678 summary data with the “sum of single effects” model,” *PLoS Genetics*,
679 vol. 18, no. 7, p. e1010299, 2022.
- 680 [35] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Pat-
681 terson, M. J. Daly, A. L. Price, and B. M. Neale, “Ld score regression
682 distinguishes confounding from polygenicity in genome-wide association
683 studies,” *Nat. Genet.*, vol. 47, no. 3, pp. 291–295, 2015.