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

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

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

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

Abstract

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

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

1 Background

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

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

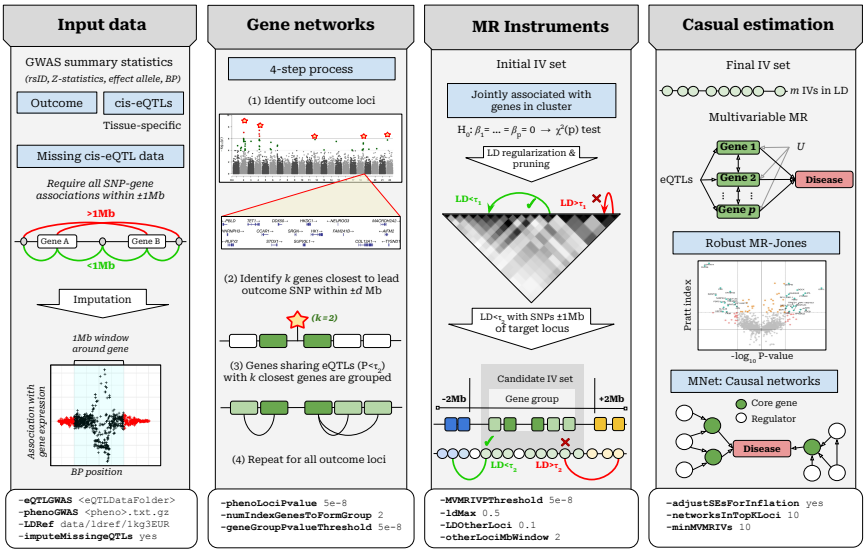

Fig. 1 Flowchart illustrating genome-wide causal gene searches using HORNET. Example options given to flags that the command line version of HORNET uses are at the bottom of each panel. In the 'Input data' section, ±1Mb is used because it is standard in many publicly available data such as GTEx [17]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

2 Methods

2.1 Data

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

2.2 IV selection and missing data

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

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

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

Algorithm 1 Pseudo-code of eQTL imputation.

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

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

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

Find $\mathbf{UDV}^\top = \widehat{\mathbf{B}}^{(t)}$ and define the kth singular value as $d_k^{(t)}$,

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

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

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

end while

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

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

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

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

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

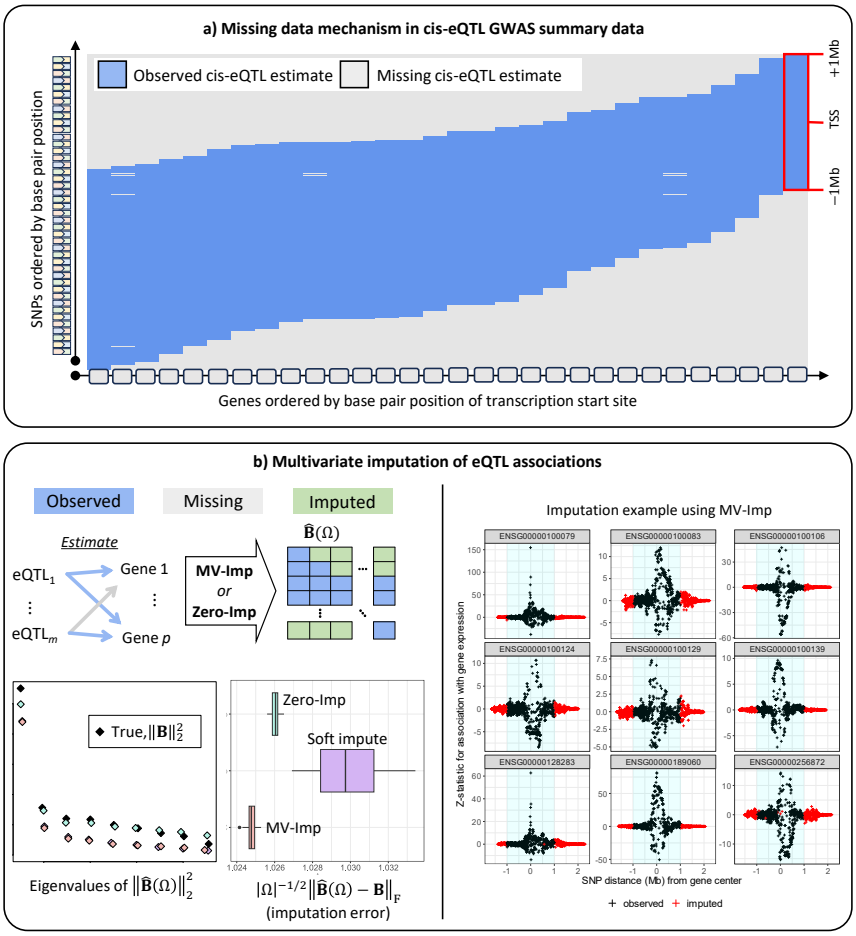

Fig. 2 This figure illustrates the mechanism in summary cis-eQTL GWAS data that leads to missing data in eQTL-MVMR and how this missing data can be addressed using imputation. a) Only SNP-gene pairs within a defined distance have association estimates present in cis-eQTL summary data. This figure demonstrates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

2.3 LD

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

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

2.3.1 CHP from LD between loci

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

2.3.2 Inflation from misspecified LD

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

IFC estimates inflation in surrounding null loci to adjust for inflation in a target locus. 'Null' loci have no SNPs that are significantly associated either with the expressions of genes or the disease phenotype (i.e., all $P > 0.01$). These loci are also at least 5Mb away from known disease-associated loci, detected as those with $P < 5 \times 10^{-8}$, an are at least 1Mb away from the transcription start sites of other genes. In these loci, we identify a set of eQTLs to use as IVs for a single gene and perform univariable MR. We repeat this process for all genes meeting the above criteria within chromosome C and calculate the genomic control inflation factor [34], denoted λ_0^C. We then return to causal estimation using eQTL-MVMR in a target locus of p genes to find the causal estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_k)_{k=1}^p$ and corresponding standard error estimates $\widehat{\text{SE}}(\widehat{\theta}_1), ..., \widehat{\text{SE}}(\widehat{\theta}_p)$. The IFC correction is applied by the inflation-corrected

standard errors

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

for statistical inference. We demonstrate in **Supplement Sections 2.6.2** using real data from the eQTLGen Consortium [19] that inflation statistics λ_0^C are stable across chromosomes and the genome. In these data, the mean inflation value across all chromosomes was 1.29, ranging from 0.89 to 1.91. Panel D of Figure 3 demonstrates that applying IFC and pruning [26, 27] controls the Type I error rate better than pruning alone across a range of scenarios in which the size of the reference panel and its concordance with the discovery data changes.

2.3.3 Non-positive definite LD matrix

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

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

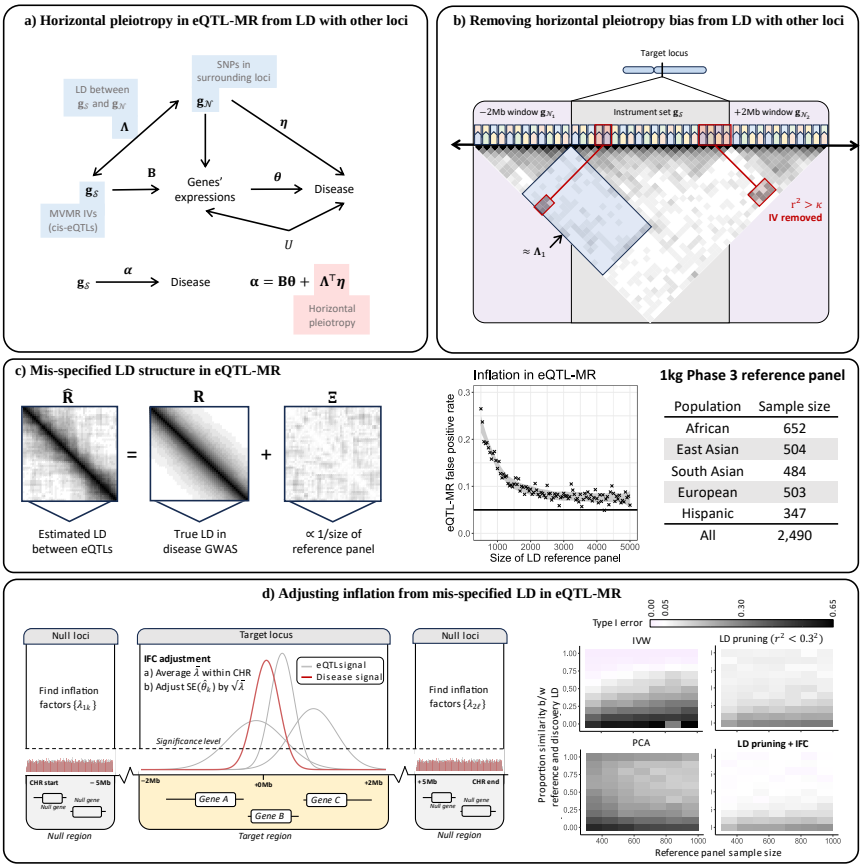

Fig. 3 This figure illustrates the adjustments for CHP and inflation that are introduced when the eQTLs used in MR are in LD and researchers only have access to relatively small reference panels. a) The goal of eQTL-MVMR is to estimate $\boldsymbol{\theta}$, which may be subject to bias when $\boldsymbol{\Lambda}$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

2.4 Causal inference

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

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

2.5 Computation

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

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

3 Simulations

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

3.1 Imputing missing data

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

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

3.2 Inflation in eQTL-MR

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

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

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

3.3 Correcting inflation from misspecified LD

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

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

4 Real data analysis with schizophrenia

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

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

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

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

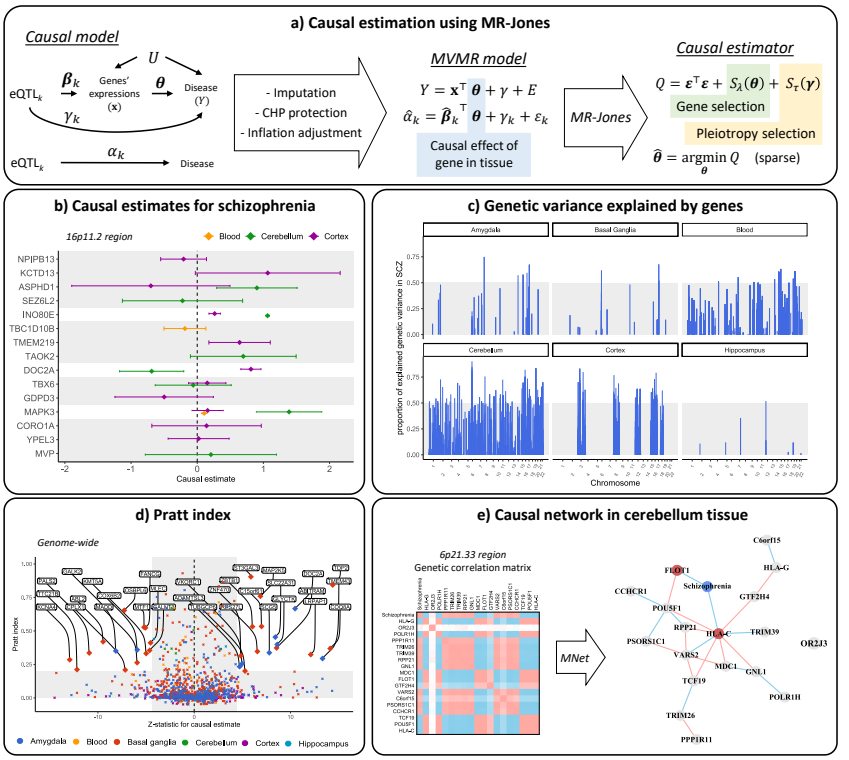

Fig. 4 This figure presents the results of using HORNET to search for genes modifying schizophrenia risk when expressed in different tissues. a) Description of the MR-Jones causal model, MVMR model, and estimator. b) Causal estimates for multiple genes in blood, cerebellum, and cortex tissues in the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

5 Conclusion

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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