Deciphering causal proteins in Alzheimer's disease: A novel Mendelian randomization method integrated with AlphaFold3 for 3D structure prediction

Minhao Yao[1], Gary W. Miller[2], Badri N. Vardarajan[3], Andrea A. Baccarelli[4], Zijian Guo[5*], Zhonghua Liu[6*]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
[2] Department of Environmental Health Sciences, Columbia University, New York, NY, USA
[3] Taub Institute on Alzheimer's Disease and the Aging Brain, Department of Neurology, Columbia University, New York, NY, USA.
[4] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5] Department of Statistics, Rutgers University, Piscataway, NJ, USA.
[6] Department of Biostatistics, Columbia University, New York, NY, USA.

[*] Correspondence to: Zijian Guo (zijguo@stat.rutgers.edu) or Zhonghua Liu (Lead contact, zl2509@cumc.columbia.edu)

Summary

Hidden confounding biases hinder identifying causal protein biomarkers for Alzheimer's disease in non-randomized studies. While Mendelian randomization (MR) can mitigate these biases using protein quantitative trait loci (pQTLs) as instrumental variables, some pQTLs violate core assumptions, leading to biased conclusions. To address this, we propose MR-SPI, a novel MR method that selects valid pQTL instruments using the Anna Karenina Principle and performs robust post-selection inference. Integrating MR-SPI with AlphaFold3, we developed a computational pipeline to identify causal protein biomarkers and predict 3D structural changes. Applied to genome-wide proteomics data from 54,306 UK Biobank participants and 455,258 subjects (71,880 cases and 383,378 controls) for a genome-wide association study of Alzheimer's disease, we identified seven proteins (TREM2, PILRB, PILRA, EPHA1, CD33, RET, and CD55) with structural alterations due to missense mutations. These findings offer insights into the etiology and potential drug targets for Alzheimer's disease.

1. Introduction

Alzheimer's disease (AD) stands as the primary cause of dementia globally, exerting a considerable strain on healthcare resources[1,2]. Despite extensive efforts, the etiology and pathogenesis of AD are still unclear, and strategies aimed at impeding or delaying its clinical advancement have largely remained challenging to achieve[1,3,4]. The amyloid cascade hypothesis posits that AD begins with the accumulation and aggregation of amyloid-beta (Aβ) peptides in the brain, culminating in the formation of β-amyloid fibrils, leading to tau hyperphosphorylation, neurofibrillary tangle formation and neurodegeneration[5,6]. However, current AD therapies targeting Aβ production and amyloid formation offer only transient symptomatic relief and fail to halt disease progression, resulting in a lack of effective drugs for AD[1,7]. Therefore, it is imperative and urgent to identify causal protein biomarkers to elucidate the underlying mechanisms of AD, and to expedite the development of effective therapeutic interventions for AD.

In causal inference, randomized controlled trials (RCTs) serve as the gold standard for evaluating the causal effect of an exposure on the health outcome of interest. However, it might be neither feasible nor ethical to perform RCTs where protein levels are considered as the exposures. Mendelian randomization (MR) leverages the random assortment of genes from parents to offspring to mimic RCTs to establish causality in non-randomized studies[8-10]. MR uses genetic variants, typically single-nucleotide polymorphisms (SNPs), as instrumental variables (IVs) to assess the causal association between an exposure and a health outcome[11]. Recently, many MR methods have been developed to investigate causal relationships using genome-wide association study (GWAS) summary statistics data that consist of effect estimates of SNP-exposure and SNP-outcome associations from two sets of samples, which are commonly referred to as the two-sample MR designs[12-15]. Since summary statistics are often publicly available and provide abundant information of associations between genetic variants and complex traits/diseases, two-sample MR methods become increasingly popular[14,16-18]. In particular, recent studies with large-scale proteomics data have unveiled numerous protein quantitative traits loci (pQTLs) associated with thousands of proteins[19,20], facilitating the application of two-sample MR methods, where pQTLs serve as IVs and protein levels serve as exposures, to identify proteins as causal biomarkers for complex traits and diseases.

To employ MR for identifying causal protein biomarkers, conventional MR methods require the pQTLs included in the analysis to be valid IVs for reliable causal inference. A pQTL is called a valid IV if the following three core IV assumptions hold[9,21]:

(A1). **Relevance**: The pQTL is associated with the protein;

(A2). **Effective Random Assignment**: The pQTL is not associated with any unmeasured confounder of the protein-outcome relationship; and

(A3). **Exclusion Restriction**: The pQTL affects the outcome only through the protein in view.

Among the three core IV assumptions (A1) - (A3), only the first assumption (A1) can be tested empirically by selecting pQTLs significantly associated with the protein. However, assumptions (A2) and (A3) cannot be empirically verified in general and may be violated in practice, which may lead to a biased estimate of the causal effect. For example, violation of (A2) may occur due to the presence of population stratification[9,22]; and violation of (A3) may occur in the presence of horizontal pleiotropy[9,23], which is a widespread biological phenomenon that the pQTL IV affects the outcome through other biological pathways that do not involve the protein in view, for example, through alternative splicing or micro-RNA effects[24-26].

Recently, several MR methods have been proposed to handle invalid IVs under certain assumptions, as summarized in Table 1. Some of these additional assumptions for the identification of the causal effect in the presence of invalid IVs are listed below:

(i) The Instrument Strength Independent of Direct Effect (InSIDE) assumption: the pQTL-protein effect is asymptotically independent of the horizontal pleiotropic effect when the number of pQTL IVs goes to infinity. For example, the random-effects inverse-variance weighted (IVW) method[27], MR-Egger[28], MR-RAPS (Robust Adjusted Profile Score)[16], and the Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO) test[29].

(ii) Majority rule condition: up to 50% of the candidate pQTL IVs are invalid. For example, the weighted median method[30] and MR-PRESSO.

(iii) Plurality rule condition or the ZEro Modal Pleiotropy Assumption (ZEMPA)[15,31]: a plurality of the candidate IVs are valid, which is weaker than the majority rule condition. For example, the mode-based estimation[31], MRMix[32] and the contamination mixture method[33].

(iv) Other distributional assumptions. For example, MRMix and the contamination mixture method impose normal mixture distribution assumption on the genetic

associations and the ratio estimates, respectively.

Despite many existing efforts, current MR methods still face new challenges when dealing with pQTLs IVs for analyzing proteomics data. First, it's worth noting that the number of pQTLs for each protein tends to be small. For example, in two proteomics studies, the median number of pQTLs per protein is 4[20,34]. With such a limited number of IVs, those MR methods based on the InSIDE assumption which requires a large number of IVs or other distributional assumptions might yield unreliable results in the presence of invalid IVs[15,28]. Second, current MR methods require an ad-hoc set of pre-determined genetic IVs, which is often obtained by selecting genetic variants with strong pQTL-protein associations in proteomics data[35]. Since such traditional way of selecting IVs only requires the proteomics data, hence the same set of selected IVs is used for assessing the causal relationships between the protein in view and different health outcomes. Obviously, this one-size-fits-all strategy for selecting IVs might not work well for different outcomes because the underlying genetic architecture may vary across outcomes. For example, the pattern of horizontal pleiotropy might vary across different outcomes. Therefore, it is desirable to develop an automatic and computationally efficient algorithm to select a set of valid genetic IVs for a specific protein-outcome pair to perform reliable causal inference, especially when the number of candidate pQTL IVs is small.

In this paper, we develop a novel all-in-one pipeline for causal protein biomarker identification and 3D structural alteration prediction using large-scale genetics, proteomics and phenotype/disease data, as illustrated in Figure 1. Specifically, we propose a two-sample MR method and algorithm that can automatically Select valid pQTL IVs and then performs robust Post-selection Inference (MR-SPI) for the causal effect of proteins on the health outcome of interest. The key idea of MR-SPI is based on the Anna Karenina Principle which states that all valid instruments are alike, while each invalid instrument is invalid in its own way – paralleling Leo Tolstoy's dictum that "all happy families are alike; each unhappy family is unhappy in its own way"[36]. In other words, valid instruments will form a group and should provide similar ratio estimates of the causal effect, while the ratio estimates of invalid instruments are more likely to be different from each other. With the application of MR-SPI, we can not only identify the causal protein biomarkers associated with disease outcomes, but can also obtain missense genetic variations (used as pQTL IVs) for those identified causal

proteins. These missense pQTL IVs will induce changes of amino acids, leading to 3D structural changes of these proteins. A classic example of a missense mutation was found in sickle cell disease, where the mutation at SNP rs334, located on chromosome 11 (11p15.4), results in the change of codon 6 of the beta globin chain from [GAA] to [GTA] [37-39]. This substitution leads to the replacement of glutamic acid with valine at position 6 of the beta chain of the hemoglobin protein, altering the structure and function of hemoglobin protein. Consequently, red blood cells assume a crescent or sickle-shaped morphology, impairing blood flow to various parts of the body[40,41]. Moreover, to further offer novel biological insights into the interpretation of the causal effect at the molecular level, we incorporate AlphaFold3[42-45] into our pipeline to predict the 3D structural alteration resulting from the corresponding missense pQTL IVs for the causal proteins identified by MR-SPI. Our pipeline can elucidate the mechanistic underpinnings of how missense genetic variations translate into 3D structural alterations at the protein level, thereby advancing our understanding of disease etiology and potentially informing targeted therapeutic interventions.

Our pioneering pipeline for the first time integrates the identification of causal protein biomarkers for health outcomes and the subsequent analysis of their 3D structural alterations into a unified framework, leveraging increasingly publicly available GWAS summary statistics for health research. Within our framework, the proposed MR-SPI serves dual purposes: (1) identifying causal protein biomarkers; and (2) selecting valid missense pQTL IVs for subsequent 3D structural analysis. Compared to existing two-sample MR methods, MR-SPI is the first MR method that utilizes both exposure and outcome data to automatically select a set of valid IVs, especially when the number of candidate IVs is small in proteomics data, which is a prominent challenge with no satisfactory solution up to date. We note that while MR-PRESSO also selects valid IVs for MR analysis, it requires both the stronger majority rule condition and the InSIDE assumption, as well as a minimum number of four candidate IVs for implementation[29]. In contrast, our proposed MR-SPI does not require the InSIDE assumption and only requires the plurality rule condition that is weaker than the majority rule condition, and only requires a minimum number of three IVs for the proposed voting procedure. Therefore, MR-SPI is more suitable for analyzing proteomics data. Extensive simulations show that our MR-SPI method outperforms other competing MR methods under the plurality rule condition. We employ MR-SPI to perform omics MR (xMR) with 912 plasma proteins using the large-scale UK Biobank proteomics data in 54,306

UK Biobank participants[20] and find 7 proteins significantly associated with the risk of Alzheimer's disease. We further use AlphaFold3[42-45] to predict the 3D structural alterations of these 7 proteins due to missense genetic variations, and then illustrate the structural alterations graphically using the PyMOL software (https://pymol.org), providing new biological insights into their functional roles in AD development and may aid in identifying potential drug targets.

2. Results

2.1 Overview of the pipeline

Our proposed all-in-one pipeline for the identification and 3D structural alteration prediction of causal protein biomarkers consists of three primary steps, as illustrated in Figure 1. First, for each protein biomarker, we employ our proposed MR-SPI to select valid pQTL IVs by incorporating the proteomics GWAS and disease outcome GWAS summary data together, and then estimate the causal effect of each protein on the outcome using the selected valid pQTL IVs. The main idea and more detailed implementation steps for MR-SPI is described in Section 2.2. Second, we perform Bonferroni correction[46] for the p-values of the estimated causal effects to identify putative causal protein biomarkers associated with the outcome. Third, for each identified protein, we apply AlphaFold3 to predict and compare the 3D structures of both the wild-type protein and mutated protein resulting from missense pQTL IVs.

2.2 MR-SPI selects valid genetic instruments by a voting procedure

MR-SPI is an automatic procedure to select valid pQTL instruments and perform robust causal inference using two-sample GWAS and proteomics data. In summary, MR-SPI consists of the following four steps, as illustrated in Figure 2:

(1). select relevant pQTL IVs that are strongly associated with the protein;

(2). each relevant pQTL IV provides a ratio estimate of the causal effect, and then all the other relevant pQTL IVs votes for it to be a valid IV if their degrees of violation of assumptions (A2) and (A3) are smaller than a data-dependent threshold as in equation (4);

(3). select valid pQTL IVs by majority/plurality voting or by finding the maximum clique of the voting matrix that encodes whether two relevant pQTL IVs mutually vote for each other to be valid (the voting matrix is defined in equation (6) in STAR (structured, transparent, accessible reporting) Methods);

(4). estimate the causal effect using the selected valid pQTL IVs and construct a

robust confidence interval with guaranteed nominal coverage even if in the presence of possible IV selection error in finite samples.

Most current two-sample MR methods only use step (1) to select (relevant) pQTL instruments for downstream MR analysis, while the selected pQTL instruments might violate assumptions (A2) and (A3), leading to possibly unreliable scientific findings. To address this issue, MR-SPI automatically select valid pQTL instruments for a specific protein-outcome pair by further incorporating the outcome GWAS data. Our key idea of selecting valid pQTL instruments is that, under the plurality rule condition, valid IVs will form the largest group and should give "similar" ratio estimates according to the Anna Karenina Principle (see STAR Methods). More specifically, we propose the following two criteria to measure the similarity between the ratio estimates of two pQTLs j and k in step (2):

C1 We say the kth pQTL "votes for" the jth pQTL to be a valid IV if, by assuming the jth pQTL is valid, the kth pQTL's degree of violation of assumptions (A2) and (A3) is smaller than a data-dependent threshold as in equation (4);

C2 We say the ratio estimates of two pQTLs j and k are "similar" if they mutually vote for each other to be valid.

In step (3), we construct a symmetric binary voting matrix to encode the votes that each relevant pQTL receives from other relevant pQTLs: the (k, j) entry of the voting matrix is 1 if pQTLs j and k mutually vote for each other to be valid, and 0 otherwise. We propose two ways to select valid pQTL IVs based on the voting matrix (see STAR Methods): (1) select relevant pQTLs who receive majority voting or plurality voting as valid IVs; and (2) use pQTLs in the maximum clique of the voting matrix as valid IVs[47]. Our simulation studies show that the maximum clique method can empirically offer lower false discovery rate (FDR)[48] and higher true positive proportion (TPP) as shown in Table S4 and Supplementary Section S6.

In step (4), we estimate the causal effect by fitting a zero-intercept ordinary least squares regression of pQTL-outcome associations on pQTL-protein associations using the set of selected valid pQTL IVs, and then construct a standard confidence interval for the causal effect using standard linear regression theory. In finite samples, some invalid IVs with small (but still nonzero) degrees of violation of assumptions (A2) and

(A3) might be incorrectly selected as valid IVs, commonly referred to as "locally invalid IVs"[49]. To address this possible issue, we propose to construct a robust confidence interval with a guaranteed nominal coverage even in the presence of IV selection error in finite-sample settings using a searching and sampling method[49], as described in Supplementary Figure S17 and STAR Methods.

2.3 Comparing MR-SPI to other competing MR methods in simulation studies

We conduct extensive simulations to evaluate the performance of MR-SPI in the presence of invalid IVs. We simulate data in a two-sample setting under four setups: (**S1**) majority rule condition holds, and no locally invalid IVs exist; (**S2**) plurality rule condition holds, and no locally invalid IVs exist; (**S3**) majority rule condition holds, and locally invalid IVs exist; (**S4**) plurality rule condition holds, and locally invalid IVs exist. More detailed simulation settings are described in STAR Methods. We compare MR-SPI to the following competing MR methods: (1) the random-effects IVW method[27], (2) MR-RAPS[16], (3) MR-PRESSO[29], (4) the weighted median method[30], (5) the mode-based estimation[31], (6) MRMix[32], and (7) the contamination mixture method[33]. We exclude MR-Egger in this simulation since it is heavily biased in our simulation settings. For simplicity, we shall use IVW to represent the random-effects IVW method hereafter.

In Figure 3, we present the percent bias, empirical coverage, and average lengths of 95% confidence intervals of those MR methods in simulated data with a sample size of 5,000 for both the exposure and the outcome. Additional simulation results under a range of sample sizes (n=5,000, 10,000, 20,000, 40,000, 80,000) can be found in Supplementary Figure S1 and Tables S1-S3. When the plurality rule condition holds and no locally invalid IVs exist, MR-SPI has small bias and short confidence interval, and the empirical coverage can attain the nominal level. When locally invalid IVs exist, the standard confidence interval might suffer from finite-sample IV selection error, and thus the empirical coverage is lower than 95% if the sample sizes are not large (e.g., 5,000). In practice, we can perform sensitivity analysis of the causal effect estimate by changing the threshold in the voting step (see STAR Methods and Supplementary Figure S14). If the causal effect estimate is sensitive to the choice of the threshold, then there might exists finite-sample IV selection error. In such cases, the proposed robust confidence interval of MR-SPI can still attain the 95% coverage level and thus is recommended for use. We also examine the performance of MR-SPI in overlapped

samples mimicking real data settings with simulation set-up and results given in Supplementary Section S8, and we find that our MR-SPI can still provide valid statistical inference.

2.4 Identifying plasma proteins associated with the risk of Alzheimer's disease

Omics MR (xMR) aims to identify omics biomarkers (e.g., proteins) causally associated with complex traits and diseases. In particular, xMR with proteomics data enables the identification of disease-associated proteins, facilitating crucial advancements in disease diagnosis, monitoring, and novel drug target discovery. In this section, we apply MR-SPI to identify putative causal plasma protein biomarkers associated with the risk of Alzheimer's disease (AD). The proteomics data used in our analysis comprises 54,306 participants from the UK Biobank Pharma Proteomics Project (UKB-PPP)[20]. In the UKB-PPP consortium, up to 22.6 million imputed autosomal variants across 1,463 proteins post quality control were analyzed, discovering 10,248 primary associations through LD (Linkage Disequilibrium) clumping $\pm$1Mb around the significant variants, including 1,163 in the *cis* region and 9,085 in the *trans* region[20]. As described in Sun, et al. [20], the following filtering steps are used to retain pQTLs in the UKB-PPP summary level proteomics data: (1)genome-wide significant (p-value$< 3.40 \times 10^{-11}$), after Bonferroni correction; and (2) independent pQTLs using LD clumping ($r^2 < 0.01$). Thus, all these candidates pQTL IVs are independent and strongly associated with the proteins. Summary statistics for AD are obtained from a meta-analysis of GWAS studies for clinically diagnosed AD and AD-by-proxy, comprising 455,258 samples in total[50]. For MR method comparison, we analyze 912 plasma proteins that share four or more candidate pQTLs within the summary statistics for AD, because the implementation of MR-PRESSO requires a minimum of four candidate IVs [29].

As presented in Figure 4(A), MR-SPI identifies 7 proteins that are significantly associated with AD after Bonferroni correction, including CD33, CD55, EPHA1, PILRA, PILRB, RET, and TREM2. The detailed information of the selected pQTL IVs for these 7 proteins can be found in Supplementary Table S6. Among them, four proteins (CD33, PILRA, PILRB, and RET) are positively associated with the risk of AD while the other three proteins (CD55, EPHA1, and TREM2) are negatively associated with the risk of AD. We also note that some competing MR methods may detect additional proteins, which are likely spurious due to invalid pQTL IVs, as demonstrated in Supplementary Section 11. Previous studies have revealed that some

of those 7 proteins and the corresponding protein-coding genes might contribute to the pathogenesis of AD[51-56], as shown in Supplementary Table S8. For example, it has been found that CD33 plays a key role in modulating microglial pathology in AD, with TREM2 acting downstream in this regulatory pathway[53]. Besides, a recent study has shown that a higher level of soluble TREM2 is associated with protection against the progression of AD pathology[57]. Additionally, RET at mitochondrial complex I is activated during ageing, which might contribute to an increased risk of ageing-related diseases including AD[55]. Using the UniProt database[58], we also find that genes encoding these 7 proteins are overexpressed in tissues including hemopoietic tissues and brain, as well as cell types including microglial, macrophages and dendritic cells. These findings highlight the potential therapeutic opportunities that target these proteins for the treatment of AD. Furthermore, in the Therapeutic Target Database (TTD)[59] and DrugBank database[60], we find existing US Food and Drug Administration (FDA)-approved drugs that target these proteins identified by MR-SPI. For example, gemtuzumab ozogamicin is a drug that targets CD33 and has been approved by FDA for acute myeloid leukemia therapy[61,62]. Besides, pralsetinib and selpercatinib are two RET inhibitors that have been FDA-approved for the treatment of non-small-cell lung cancers[63,64]. Therefore, these drugs might be potential drug repurposing candidates for the treatment of AD.

In Figure 4(B), we present the 3D structural alterations of CD33 due to missense genetic variation of pQTL rs2455069, as predicted by AlphaFold3[42,43,45]. The 3D structures are shown in blue when the allele is A, and in red when the allele is G at pQTL rs2455069 A/G, which is a cis-SNP located on chromosome 19 (19q13.41) and is selected as a valid IV by MR-SPI. The presence of the G allele at pQTL rs2455069 results in the substitution of the 69th amino acid of CD33, changing it from Arginine (colored in green if the allele is A) to Glycine (colored in yellow if the allele is G), consequently causing a local change in the structure of CD33 (R69G). Previous studies have found that CD33 is overexpressed in microglial cells in the brain[65], and the substitution of Arginine to Glycine in the 69th amino acid of CD33 might lead to the accumulation of amyloid plaques in the brain[66], thus the presence of the G allele at pQTL rs2455069 might contribute to an increased risk of AD. We also apply AlphaFold3 to predict the 3D structures of the other proteins that are detected to be significantly associated with AD by MR-SPI, which are presented in Supplementary Figure S16.

In Figure 4(C), we present the point estimates and 95% confidence intervals of the causal effects (on the log odds ratio scale) of these 7 proteins on AD using the other competing MR methods. In Figure 4(C), these proteins are identified by most of the competing MR methods, confirming the robustness of our findings. To the best of our knowledge, there may be two reasons for the differences in results between MRMix and other MR methods for some proteins: (1) MRMix assumes that the pQTL-protein and pQTL-outcome associations follow a bivariate normal-mixture model with four mixture components while the contamination mixture models assume that the ratio estimator follows a normal distribution with two mixture components, and therefore it may be more challenging to obtain reliable causal effect estimates using the MRMix model with a small number of pQTLs per protein; and (2) the default grid search values implemented in the MRMix R package might not be optimal for some proteins. Notably, MR-SPI detects one possibly invalid IV pQTL rs10919543 for TREM2-AD relationship, which is associated with red blood cell count according to PhenoScanner[67]. Red blood cell count is a known risk factor for AD[68,69], and thus pQTL rs10919543 might exhibit pleiotropy in the relationship of TREM2 on AD. After excluding this potentially invalid IV, MR-SPI suggests that TREM2 is negatively associated with the risk of AD ($\hat{\beta} = -0.04$, p-value= 1.20×10^{-18}). Additionally, we perform the gene ontology (GO) enrichment analysis using the g:Profiler web server[70] (https://biit.cs.ut.ee/gprofiler/gost) to gain more biological insights for the 7 proteins identified by MR-SPI, and the results are presented in Figure 4(D) and Supplementary Table S7. After Bonferroni correction, the GO analysis indicates that these 7 proteins are significantly enriched in 20 GO terms, notably, the positive regulation of phosphorus metabolic process and major histocompatibility complex (MHC) class I protein binding. It has been found that increased phosphorus metabolites (e.g., phosphocreatine) are associated with aging, and that defects in metabolic processes for phospholipid membrane function is involved in the pathological progression of Alzheimer's disease[71,72]. In addition, MHC class I proteins may play a crucial role in preserving brain integrity during post-developmental stages, and modulation of the stability of MHC class I proteins emerges as a potential therapeutic target for restoring synaptic function in AD[73-75].

3. Discussion

In this paper, we develop a novel integrated pipeline that combines our proposed MR-SPI method with AlphaFold3 to identify putative causal protein biomarkers for complex

traits/diseases and to predict the 3D structural alterations induced by missense pQTL IVs. Specifically, MR-SPI is an automatic algorithm to select valid pQTL IVs under the plurality rule condition for a specific protein-outcome pair from two-sample GWAS summary statistics. MR-SPI first selects relevant pQTL IVs with strong pQTL-protein associations to minimize weak IV bias, and then applies the proposed voting procedure to select valid pQTL IVs whose ratio estimates are similar to each other. In the possible presence of locally invalid IVs in finite-sample settings, MR-SPI further provides a robust confidence interval constructed by the searching and sampling method[49], which is immune to finite-sample IV selection error. The valid pQTL IVs selected by MR-SPI serve dual purposes: (1) facilitating more reliable scientific discoveries in identifying putative causal proteins associated with diseases; and (2) shedding new light on the molecular-level mechanism of causal proteins in disease etiology through the 3D structural alterations of mutated proteins induced by missense pQTL IVs. We employ MR-SPI to conduct xMR analysis with 912 plasma proteins using the proteomics data in 54,306 UK Biobank participants and identify 7 proteins significantly associated with the risk of Alzheimer's disease. The 3D structural changes in these proteins, as predicted by AlphaFold3 in response to missense genetic variations of selected pQTL IVs, offering new insights into their biological functions in the etiology of Alzheimer's disease. We also found existing FDA-approved drugs that target some of our identified proteins, which provide opportunities for potential existing drug repurposing for the treatment of Alzheimer's disease. These findings highlight the great potential of our proposed pipeline for identifying protein biomarkers as new therapeutic targets and drug repurposing for disease prevention and treatment.

We emphasize three main advantages of MR-SPI. First, MR-SPI incorporates both proteomics and outcome data to automatically select a set of valid pQTL instruments in genome-wide studies, and the selection procedure does not rely on any additional distributional assumptions on the genetic effects nor require a large number of candidate IVs. Therefore, MR-SPI is the first method to offer such a practically robust approach to selecting valid pQTL IVs for a specific exposure-outcome pair from GWAS studies for more reliable MR analyses, which is especially advantageous in the presence of wide-spread horizontal pleiotropy and when only a small number of candidate IVs are available in xMR studies. While our real data application specifically focuses on the identification of putative causal protein biomarkers for Alzheimer's disease through the integration of MR-SPI with AlphaFold3, it's important to highlight that MR-SPI holds

426 broader applicability in elucidating causal relationships across complex traits and
427 diseases. For additional data analysis results and insights into the utility of MR-SPI in
428 this context, please refer to Supplementary Sections S9 and S10. Second, we propose a
429 robust confidence interval for the causal effect using the searching and sampling
430 method, which is immune to finite-sample IV selection error. Therefore, when locally
431 invalid IVs are incorrectly selected, MR-SPI can still provide reliable statistical
432 inference for the causal effect using the proposed robust confidence interval. Third,
433 MR-SPI is computationally efficient. The average computation time for constructing
434 the standard CI and the robust CI with 20 candidate IVs is 0.02 seconds and 10.60
435 seconds, respectively, using a server equipped with an Intel Xeon Silver 4116 CPU and
436 64 GB RAM memory.
437
438 In conclusion, MR-SPI is a powerful tool for identifying putative causal protein
439 biomarkers for complex traits and diseases. The integration of MR-SPI with
440 AlphaFold3 as a computationally efficient pipeline can further predict the 3D structural
441 alterations caused by missense pQTL IVs, improving our understanding of molecular-
442 level disease mechanisms. Therefore, our pipeline holds promising implications for
443 drug target discovery, drug repurposing, and therapeutic development.
444
445 **Limitations of the Study**
446 MR-SPI has some limitations. First, MR-SPI uses independent pQTLs as candidate IVs
447 after LD clumping, which might exclude strong and valid pQTL IVs. We plan to extend
448 MR-SPI to include correlated pQTLs with arbitrary LD structure to increase statistical
449 power. Second, the proposed robust confidence interval is slightly more conservative,
450 which is the price to pay for the gained robustness to finite-sample IV selection error.
451 We plan to construct less conservative confidence intervals with improved power to
452 detect more putative causal proteins. Third, we will incorporate colocalization
453 analysis[76-79] into our pipeline to better understand the shared genetic architecture
454 between proteins and disease outcomes when unfiltered GWAS summary statistics are
455 available in future studies.
456
457 **4. Resource Availability**
458 **Lead Contact**
459 Further information and requests for resources and reagents should be directed to and
460 will be fulfilled by the lead contact, Zhonghua Liu (zl2509@cumc.columbia.edu)

Materials Availability

The materials that support the findings of this study are available from the corresponding authors upon reasonable request. Please contact the lead contact, Zhonghua Liu (zl2509@cumc.columbia.edu) for additional information.

Data and Code Availability

All the GWAS data analyzed are publicly available with the following URLs:

- GWAS for Alzheimer's disease: https://ctg.cncr.nl/software/summary_statistics;
- UK Biobank proteomics data: https://www.biorxiv.org/content/10.1101/2022.06.17.496443v1.supplementary-material

The R package **MR.SPI** is publicly available at https://github.com/MinhaoYaooo/MR-SPI.

5. Acknowledgments

We thank the UK Biobank Pharma Proteomics Project (http://ukb-ppp.gwas.eu/) for providing the proteomics data. We thank the Center for Neurogenomics and Cognitive Research (CNCR, https://cncr.nl/) for providing GWAS summary statistics for Alzheimer's disease.

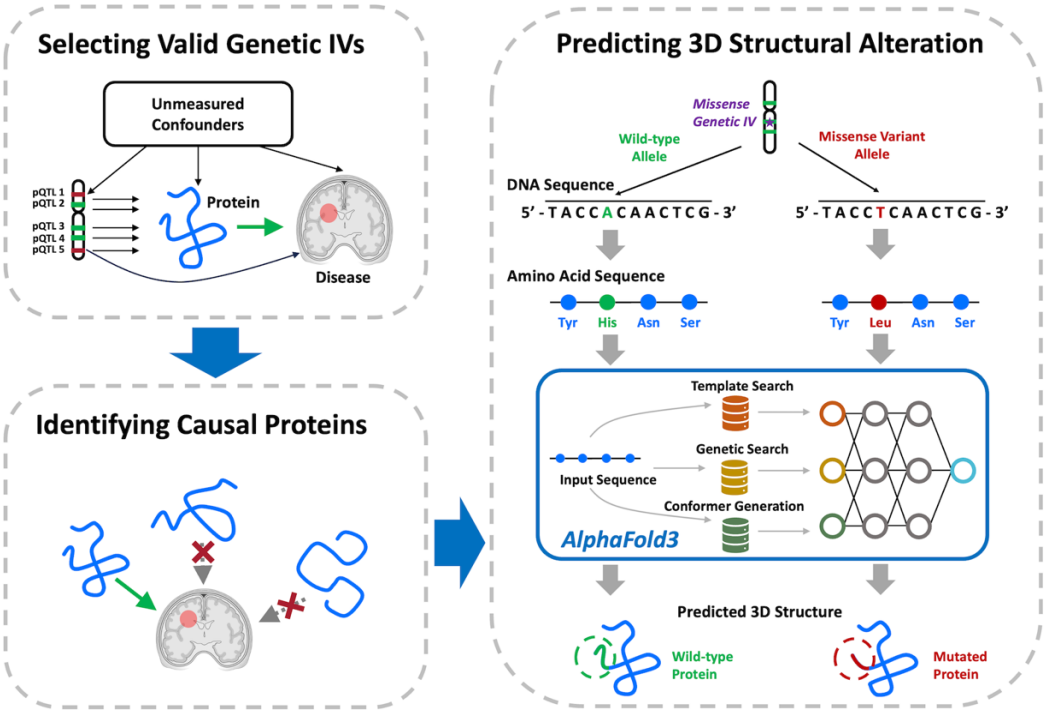
6. Author Contributions

Conceptualization, Zijian Guo and Zhonghua Liu; methodology, Minhao Yao, Zijian Guo and Zhonghua Liu; software, Minhao Yao; formal analysis: Minhao Yao; writing (original draft), Minhao Yao, Zijian Guo and Zhonghua Liu; writing (review & editing), Minhao Yao, Gary W. Miller, Badri N. Vardarajan, Andrea A. Baccarelli, Zijian Guo and Zhonghua Liu; supervision, Zijian Guo and Zhonghua Liu.

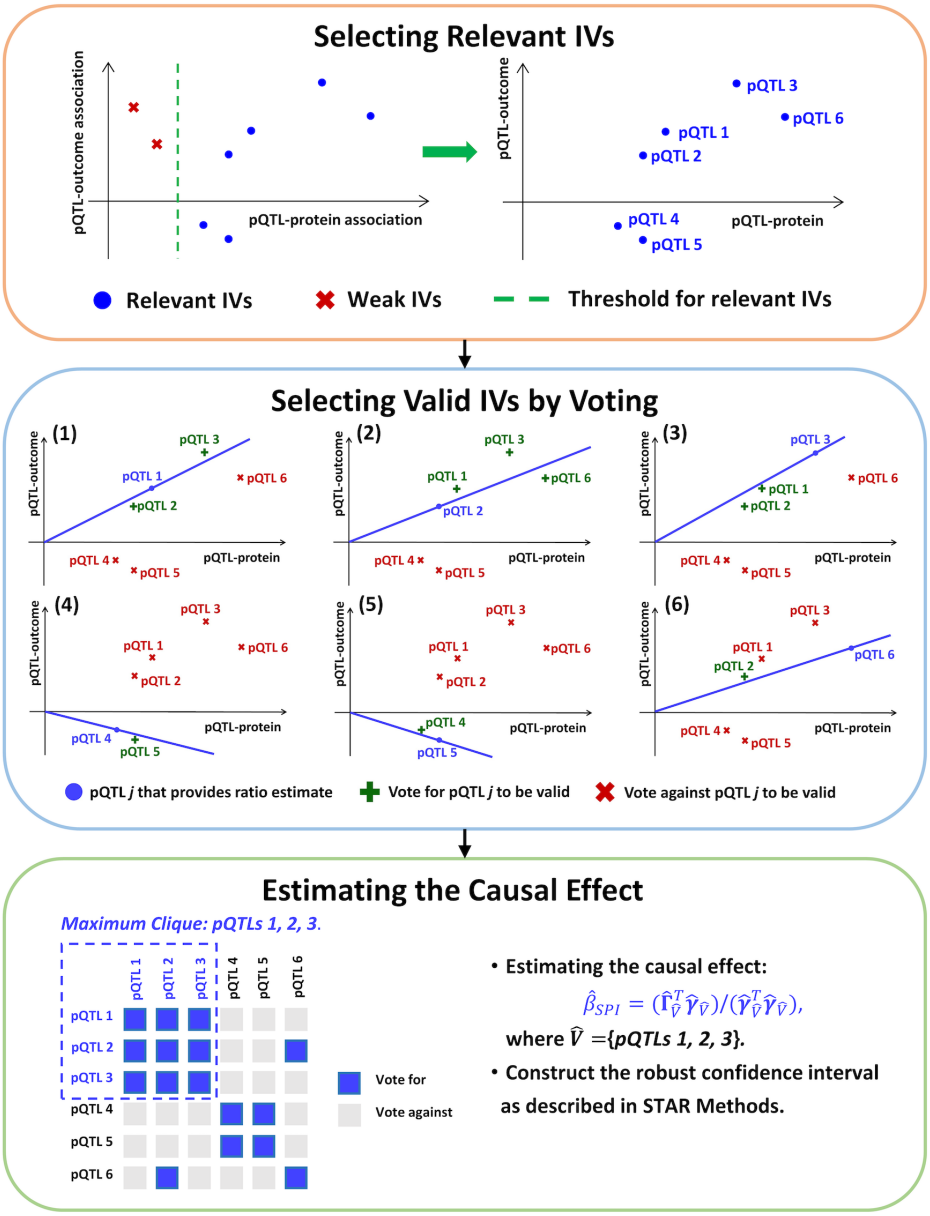
7. Declaration of Interests

The authors declare no competing interests.

8. Main Figure Titles and Legends

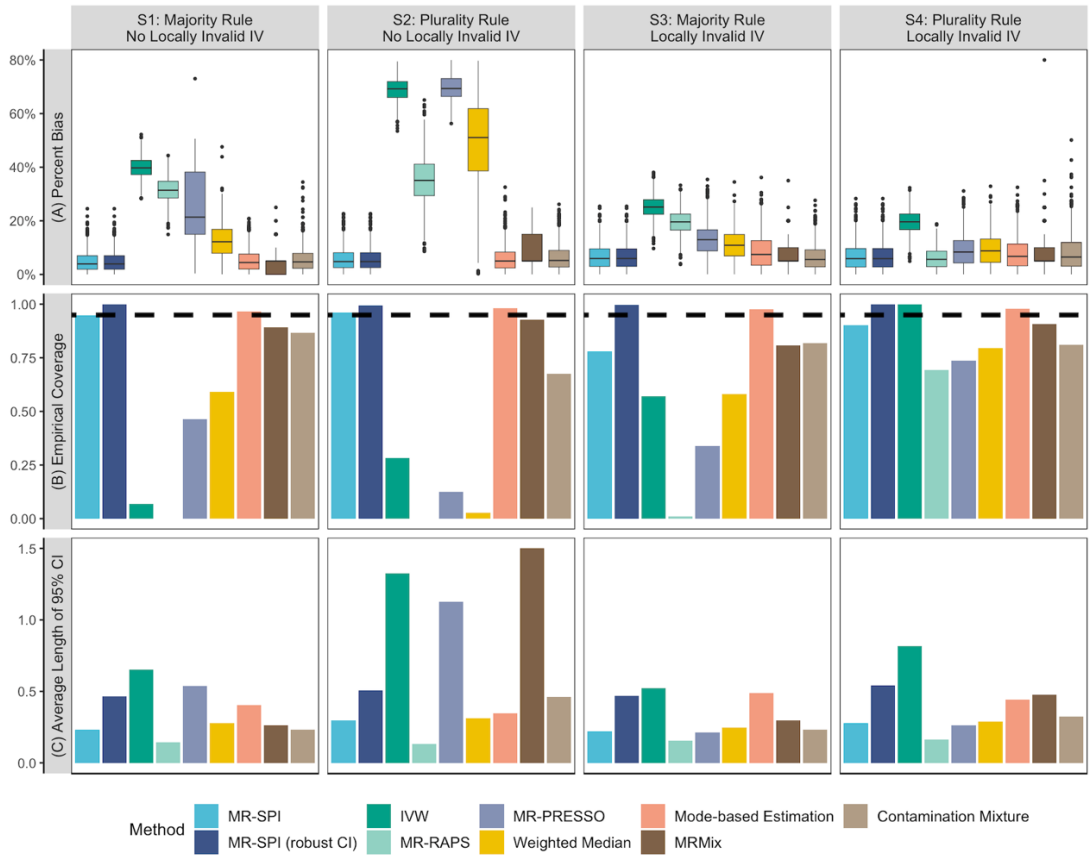


Main Figure 1. Overview of the pipeline.
First, we apply MR-SPI for each protein to (1) select valid pQTL IVs under the plurality condition, and (2) estimate the causal effect on the outcome of interest. Second, we perform the Bonferroni correction procedure for causal protein identification. Third, for each causal protein biomarker, we apply AlphaFold3 to predict the 3D structural alterations due to missense pQTL IVs selected by MR-SPI.



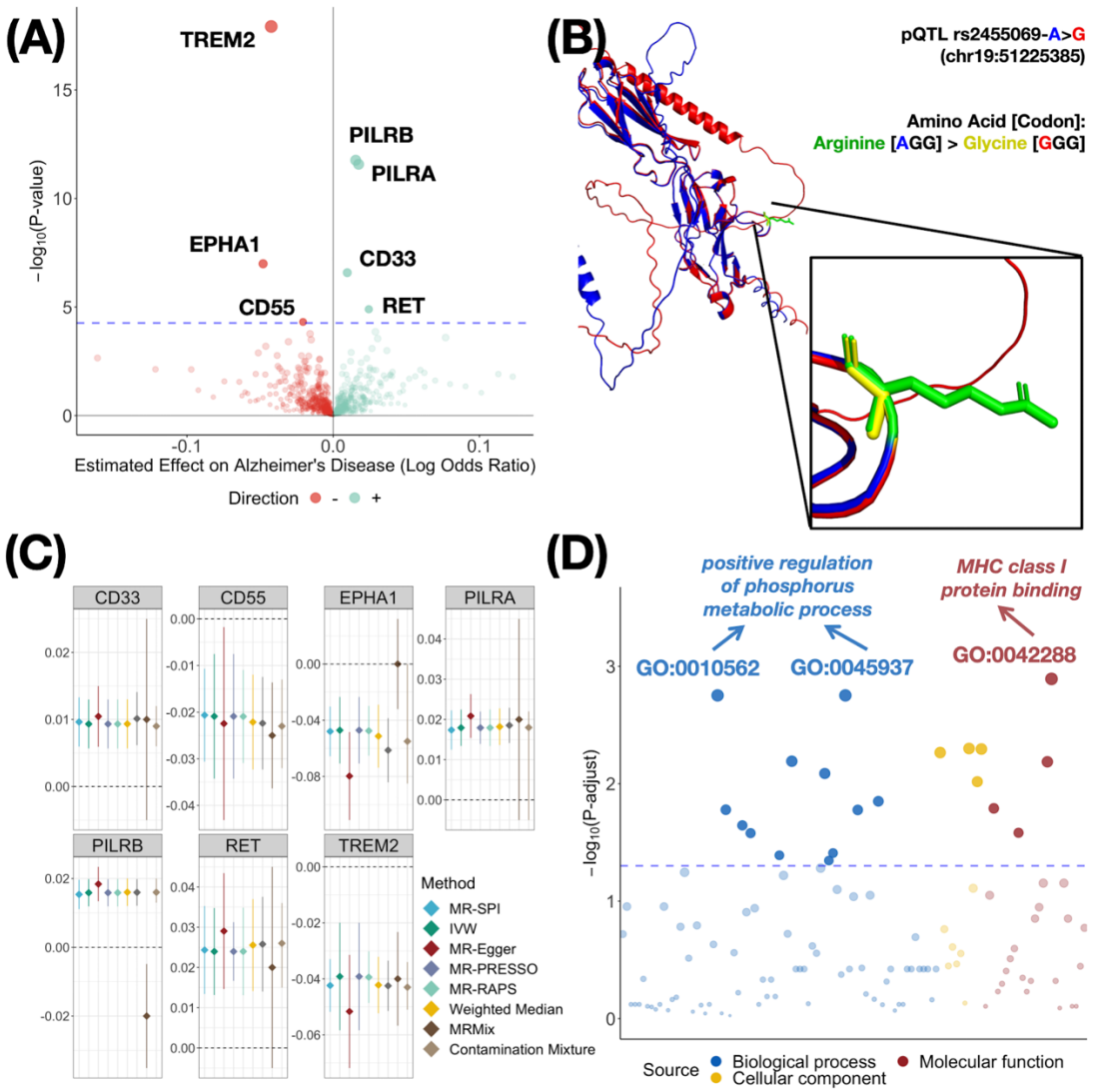
Main Figure 2. The MR-SPI framework.

First, MR-SPI selects relevant IVs with strong pQTL-protein associations. Second, each relevant IV provides a ratio estimate of the causal effect and then receives votes on itself to be valid from the other relevant IVs whose degrees of violation of (A2) and (A3) are small under this ratio estimate of causal effect. For example, by assuming pQTL 1 is valid, the slope of the line connecting pQTL 1 and the origin represents the ratio estimate of pQTL 1, and pQTLs 2 and 3 vote for pQTL 1 to be valid because they are close to that line, while pQTLs 4, 5 and 6 vote against it since they are far away from that line. Third, MR-SPI estimates the causal effect by fitting a zero-intercept OLS regression of pQTL-outcome associations on pQTL-protein associations and construct the robust confidence interval using selected valid pQTL IVs in the maximum clique of the voting matrix, which encodes whether two pQTLs mutually vote for each other to be valid IVs.



Main Figure 3. Empirical performance of MR-SPI and the other competing MR methods in simulated data with sample size 5,000.

(A) Boxplot of the percent bias in causal effect estimates. (B) Empirical coverage of 95% confidence intervals. The black dashed line in (B) represents the nominal level (95%). (C) Average lengths of 95% confidence intervals.



Main Figure 4. Causal effect estimates, gene ontology analysis and structural alteration prediction for putatively causal proteins.

(A) Volcano plot of associations of plasma proteins with Alzheimer's disease using MR-SPI. The horizontal axis represents the estimated effect size (on the log odds ratio scale), and the vertical axis represents the $-\log_{10}(p\text{-value})$. Positive and negative associations are represented by green and red points, respectively. The size of a point is proportional to $-\log_{10}(p\text{-value})$. The blue dashed line represents the significance threshold using Bonferroni correction ($p\text{-value} < 5.48 \times 10^{-5}$). (B) 3D Structural alterations of CD33 predicted by AlphaFold3 due to missense genetic variation of pQTL rs2455069. The ribbon representation of 3D structures of CD33 with Arginine and Glycine at position 69 are colored in blue and red, respectively. The amino acids at position 69 are displayed in stick representation, with Arginine and Glycine colored in green and yellow, respectively. The predicted template modeling (pTM) yields a score of 0.6 for both structures, which suggests that AlphaFold3 provides good predictions for these two 3D structures. (C) Forest plot of significant associations of proteins with Alzheimer's disease identified by MR-SPI. Confidence intervals are clipped to vertical axis limits. (D) Bubble plot of GO analysis results using the 7 significant proteins

detected by MR-SPI. The horizontal axis represents the z-score of the enriched GO term, and the vertical axis represents the $-\log_{10}(p\text{-value})$ after Bonferroni correction. Each point represents one enriched GO term. The blue dashed line represents the significance threshold (adjusted p-value< 0.05 after Bonferroni correction).

9. Main Tables and Legends

Main Table 1: Comparison of MR methods and the underlying assumptions for handling invalid IVs. Balanced pleiotropy means on average the pleiotropic effects have zero mean. NOME assumption refers to NO Measurement Error in the exposure data.

MR Method	**InSIDE assumption**	**Majority/plurality rule condition**	**Other distributional assumptions**
Random-effects IVW	Required	Not required	Balanced pleiotropy, NOME assumption
MR-Egger	Required	Not required	NOME assumption
MR-RAPS	Required	Not required	Balanced pleiotropy
MR-PRESSO	Required	Majority rule condition	Not required
Weighed median	Not required	Majority rule condition	Not required
Mode-based estimation	Not required	Plurality rule condition	Not required
MRMix	Not required	Plurality rule condition	Direct effects of a SNP on the exposure and outcome follow a mixture of normal distributions
Contamination mixture method	Not required	Plurality rule condition	Ratio estimate of a SNP follow a mixture of normal distributions, NOME assumption
MR-SPI	**Not required**	**Plurality rule condition**	**Not required**

STAR Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK Biobank proteomics data	Sun, et al. [20]	http://ukb-ppp.gwas.eu/
Genome-wide association studies of Alzheimer's disease	Jansen, et al. [50]	https://www.nature.com/articles/s41588-018-0311-9
Software and algorithms		
AlphaFold3	Abramson, et al. [45]	https://alphafoldserver.com
g:Profiler	Raudvere, et al. [70]	https://biit.cs.ut.ee/gprofiler/gost
MendelianRandomization	Yavorska and Burgess [80]	https://github.com/cran/MendelianRandomization
MRMix	Qi and Chatterjee [32]	https://github.com/gqi/MRMix
MR-PRESSO	Verbanck, et al. [29]	https://github.com/rondolab/MR-PRESSO
MR-RAPS	Zhao, et al. [16]	https://github.com/qingyuanzhao/mr.raps
MR-SPI	This study	https://github.com/MinhaoYaooo/MR-SPI
PLINK	Purcell, et al. [81]	https://www.cog-genomics.org/plink/1.9
PyMol	PyMOL Molecular Graphics	https://pymol.org

	System	
R	The R Foundation for Statistical Computing	https://www.r-project.org/

Method details

Two-sample GWAS summary statistics

Suppose that we obtain p independent pQTLs $\boldsymbol{Z} = (Z_1, \cdots, Z_p)^\top$ by using LD clumping that retains one representative pQTL per LD region[81]. We also assume that the pQTLs are standardized[82] such that $\mathbb{E}Z_j = 0$ and $\mathrm{Var}(Z_j) = 1$ for $1 \le j \le p$. Let D denote the exposure and Y denote the outcome. We assume that D and Y follow the exposure model $D = \boldsymbol{Z}^\top \boldsymbol{\gamma} + \delta$ and the outcome model $Y = D\beta + \boldsymbol{Z}^\top \boldsymbol{\pi} + e$, respectively, where β represents the causal effect of interest, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^\top$ represents the IV strength, and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_p)^\top$ encodes the violation of assumptions (A2) and (A3)[83,84]. If assumptions (A2) and (A3) hold for pQTL j, then $\pi_j = 0$ and otherwise $\pi_j \neq 0$ (see Supplementary Section S1 for details). The error terms δ and e with respective variances σ_δ^2 and σ_e^2 are possibly correlated due to unmeasured confounding factors. By plugging the exposure model into the outcome model, we obtain the reduced-form outcome model $Y = \boldsymbol{Z}^\top(\beta\boldsymbol{\gamma} + \boldsymbol{\pi}) + \epsilon$, where $\epsilon = \beta\delta + e$. Let $\boldsymbol{\Gamma} = (\Gamma_1, \cdots, \Gamma_p)^\top$ denote the pQTL-outcome associations, then we have $\boldsymbol{\Gamma} = \beta\boldsymbol{\gamma} + \boldsymbol{\pi}$. If $\gamma_j \neq 0$, then pQTL j is called a relevant IV. If both $\gamma_j \neq 0$ and $\pi_j = 0$, then pQTL j is called a valid IV. Let $\mathcal{S} = \{j: \gamma_j \neq 0, 1 \le j \le p\}$ denote the set of all relevant IVs, and $\mathcal{V} = \{j: \gamma_j \neq 0 \text{ and } \pi_j = 0, 1 \le j \le p\}$ denote the set of all valid IVs. The majority rule condition can be expressed as $|\mathcal{V}| > \frac{1}{2}|\mathcal{S}|$[84], and the plurality rule condition can be expressed as $|\mathcal{V}| > \max_{c \neq 0} |\{j \in \mathcal{S}: \pi_j/\gamma_j = c\}|$ [83]. If the plurality rule condition holds, then valid IVs with the same ratio of pQTL-outcome effect to pQTL-protein effect will form a plurality. Based on this key observation, our proposed MR-SPI selects the largest group of pQTLs as valid IVs with similar ratio

estimates of the causal effect using a voting procedure described in detail in the next subsection.

Let $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ be the estimated marginal effects of pQTL j on the protein and the outcome, and $\hat{\sigma}_{\gamma_j}$ and $\hat{\sigma}_{\Gamma_j}$ be the corresponding estimated standard errors respectively. Let $\hat{\boldsymbol{\gamma}} = \left(\hat{\gamma}_1, \cdots, \hat{\gamma}_p\right)^\top$ and $\hat{\boldsymbol{\Gamma}} = \left(\hat{\Gamma}_1, \cdots, \hat{\Gamma}_p\right)^\top$ denote the vector of estimated pQTL-protein and pQTL-outcome associations, respectively. In the two-sample setting, the summary statistics $\left\{\hat{\gamma}_j, \hat{\sigma}_{\gamma_j}\right\}_{1 \le j \le p}$ and $\left\{\hat{\Gamma}_j, \hat{\sigma}_{\Gamma_j}\right\}_{1 \le j \le p}$ are calculated from two non-overlapping samples with sample sizes n_1 and n_2 respectively. When all the pQTLs are independent of each other, the joint asymptotic distribution of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Gamma}}$ is

$$\begin{pmatrix} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma} \end{pmatrix} \xrightarrow{d} N\left[\mathbf{0}, \begin{pmatrix} \frac{1}{n_1}\mathbf{V}_\gamma & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\mathbf{V}_\Gamma \end{pmatrix}\right]$$

where the diagonal entries of $\mathbf{V}_\gamma$ and $\mathbf{V}_\Gamma$ are $\mathbf{V}_{\gamma,jj} = \mathrm{Var}\left(Z_{ij}^2\right)\gamma_j^2 + \sum_{l \ne j}\gamma_l^2 + \sigma_\delta^2$ and $\mathbf{V}_{\Gamma,jj} = \mathrm{Var}\left(Z_{ij}^2\right)\Gamma_j^2 + \sum_{l \ne j}\Gamma_l^2 + \sigma_\epsilon^2$, respectively, and the off-diagonal entries of $\mathbf{V}_\gamma$ and $\mathbf{V}_\Gamma$ are $\mathbf{V}_{\gamma,j_1j_2} = \gamma_{j_1}\gamma_{j_2}$ and $\mathbf{V}_{\Gamma,j_1j_2} = \Gamma_{j_1}\Gamma_{j_2}(j_1 \ne j_2)$, respectively. The derivation of the limit distribution can be found in Supplementary Section S2. Therefore, with the summary statistics of the protein and the outcome, we estimate the covariance matrices $\frac{1}{n_1}\mathbf{V}_\gamma$ and $\frac{1}{n_2}\mathbf{V}_\Gamma$ as:

$$\frac{1}{n_1}\widehat{\mathbf{V}}_{\gamma,j_1j_2} = \begin{cases} \hat{\sigma}_{\gamma_{j_1}}^2 & \text{if } j_1 = j_2, \\ \frac{1}{n_1}\hat{\gamma}_{j_1}\hat{\gamma}_{j_2} & \text{if } j_1 \ne j_2. \end{cases} \text{ and } \frac{1}{n_2}\widehat{\mathbf{V}}_{\Gamma,j_1j_2} = \begin{cases} \hat{\sigma}_{\Gamma_{j_1}}^2 & \text{if } j_1 = j_2 \\ \frac{1}{n_2}\hat{\Gamma}_{j_1}\hat{\Gamma}_{j_2} & \text{if } j_1 \ne j_2 \end{cases} \quad (1)$$

After obtaining $\left\{\hat{\boldsymbol{\gamma}}, \widehat{\mathbf{V}}_\gamma, \hat{\boldsymbol{\Gamma}}, \widehat{\mathbf{V}}_\Gamma\right\}$, we then perform the proposed IV selection procedure as illustrated in Figure 2 in the main text.

Selecting valid instruments by voting

The first step of MR-SPI is to select relevant pQTLs with large IV strength using

proteomics data. Specifically, we estimate the set of relevant IVs $\mathcal{S}$ by:

$$\hat{\mathcal{S}} = \left\{1 \leq j \leq p: \frac{|\hat{\gamma}_j|}{\hat{\sigma}_{\gamma_j}} > \Phi^{-1}\left(1 - \frac{\alpha^*}{2}\right)\right\} \tag{2}$$

where $\hat{\sigma}_{\gamma_j}$ is the standard error of $\hat{\gamma}_j$ in the summary statistics, $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution, and α^* is the user-specified threshold with the default value of 1×10^{-6}. This step is equivalent to filtering the pQTLs in the proteomics data with p-value $< \alpha^*$, and is adopted by most of the current two-sample MR methods to select (relevant) genetic instruments for downstream MR analysis. Note that the selected pQTL instruments may not satisfy the IV independence and exclusion restriction assumptions and thus maybe invalid. In contrast, our proposed MR-SPI further incorporates the outcome data to automatically select a set of valid genetic instruments from $\hat{\mathcal{S}}$ for a specific protein-outcome pair.

Under the plurality rule condition, valid pQTL instruments with the same ratio of pQTL-outcome effect to pQTL-protein effect (i.e., Γ_j/γ_j) will form a plurality and yield "similar" ratio estimates of the causal effect. Based on this key observation, MR-SPI selects a plurality of relevant IVs whose ratio estimates are "similar" to each other as valid IVs. Specifically, we propose the following two criteria to measure the similarity between the ratio estimates of two pQTLs j and k :

C1: We say the kth pQTL "votes for" the jth pQTL to be a valid IV if, by assuming the jth pQTL is valid, the kth pQTL's degree of violation of assumptions (A2) and (A3) is smaller than a threshold as in equation (4);

C2: We say the ratio estimates of two pQTLs j and k are "similar" if they mutually vote for each other to be valid IVs.

The ratio estimate of the jth pQTL is defined as $\hat{\beta}^{[j]} = \hat{\Gamma}_j/\hat{\gamma}_j$. By assuming the jth pQTL is valid, the plug-in estimate of the kth pQTL's degree of violation of (A2) and (A3) can be obtained by

$$\hat{\pi}_k^{[j]} = \hat{\Gamma}_k - \hat{\beta}^{[j]}\hat{\gamma}_k = \left(\hat{\beta}^{[k]} - \hat{\beta}^{[j]}\right)\hat{\gamma}_k \tag{3}$$

as we have $\Gamma_k = \beta\gamma_k + \pi_k$ for the true causal effect β, and $\hat{\Gamma}_k = \hat{\beta}^{[k]}\hat{\gamma}_k$ for the ratio

estimate $\hat{\beta}^{[k]}$ of the kth pQTL. From equation (3), $\hat{\pi}_k^{[j]}$ has two noteworthy implications. First, $\hat{\pi}_k^{[j]}$ measures the difference between the ratio estimates of pQTLs j and k (multiplied by the kth pQTL-protein effect estimate $\hat{\gamma}_k$), and a small $\hat{\pi}_k^{[j]}$ implies that the difference scaled by $\hat{\gamma}_k$ is small. Second, $\hat{\pi}_k^{[j]}$ represents the kth IV's degree of violation of assumptions (A2) and (A3) by regarding the jth pQTL's ratio estimate $\hat{\beta}^{[j]}$ as the true causal effect, thus a small $\hat{\pi}_k^{[j]}$ implies a strong evidence that the kth IV supports the jth IV to be valid. Therefore, we say the kth IV votes for the jth IV to be valid if:

$$\frac{\left|\hat{\pi}_k^{[j]}\right|}{\widehat{\mathrm{SE}}\left(\hat{\pi}_k^{[j]}\right)} \leq \sqrt{\log\ \min(n_1, n_2)} \tag{4}$$

where $\widehat{\mathrm{SE}}\left(\hat{\pi}_k^{[j]}\right)$ is the standard error of $\hat{\pi}_k^{[j]}$, which is given by:

$$\widehat{\mathrm{SE}}\left(\hat{\pi}_k^{[j]}\right) = \sqrt{\frac{1}{n_2}\left(\hat{\mathbf{V}}_{\Gamma,kk} + \left(\frac{\hat{\gamma}_k}{\hat{\gamma}_j}\right)^2 \hat{\mathbf{V}}_{\Gamma,jj} - 2\frac{\hat{\gamma}_k}{\hat{\gamma}_j}\hat{\mathbf{V}}_{\Gamma,jk}\right) + \frac{1}{n_1}\left(\hat{\beta}^{[j]}\right)^2\left(\hat{\mathbf{V}}_{\gamma,kk} + \left(\frac{\hat{\gamma}_k}{\hat{\gamma}_j}\right)^2 \hat{\mathbf{V}}_{\gamma,jj} - 2\frac{\hat{\gamma}_k}{\hat{\gamma}_j}\hat{\mathbf{V}}_{\gamma,jk}\right)} \tag{5}$$

and the term $\sqrt{\log\ \min(n_1, n_2)}$ in equation (4) ensures that the violation of (A2) and (A3) can be correctly detected with probability one as the sample sizes go to infinity, as shown in Supplementary Section S3.

For each relevant IV in $\hat{\mathcal{S}}$, we collect all relevant IVs' votes on whether it is a valid IV according to equation (4). Then we construct a voting matrix $\hat{\mathbf{\Pi}} \in \mathbb{R}^{|\hat{\mathcal{S}}|\times|\hat{\mathcal{S}}|}$ to summarize the voting results and evaluate the similarity of two pQTLs' ratio estimates according to criterion C2. Specifically, we define the (k, j) entry of $\hat{\mathbf{\Pi}}$ as:

$$\hat{\Pi}_{k,j} = I\left(\max\left\{\frac{\left|\hat{\pi}_k^{[j]}\right|}{\widehat{\mathrm{SE}}\left(\hat{\pi}_k^{[j]}\right)}, \frac{\left|\hat{\pi}_j^{[k]}\right|}{\widehat{\mathrm{SE}}\left(\hat{\pi}_j^{[k]}\right)}\right\} \leq \sqrt{\log \min(n_1, n_2)}\right) \tag{6}$$

where $I(\cdot)$ is the indicator function such that $I(A) = 1$ if event A happens and $I(A) = 0$ otherwise. From equation (6), we can see that the voting matrix $\hat{\mathbf{\Pi}}$ is

symmetric, and the entries of $\widehat{\mathbf{\Pi}}$ are binary: $\widehat{\Pi}_{k,j} = 1$ represents pQTLs j and k vote for each other to be a valid IV, i.e., the ratio estimates of these two pQTLs are close to each other; $\widehat{\Pi}_{k,j} = 0$ represents that they do not. For example, in Figure 2, $\widehat{\Pi}_{1,2} = 1$ since the ratio estimates of pQTLs 1 and 2 are similar, while $\widehat{\Pi}_{1,4} = 0$ because the ratio estimates of pQTLs 1 and 4 differ substantially, as pQTLs 1 and 4 mutually "vote against" each other to be valid according to equation (4).

After constructing the voting matrix $\widehat{\mathbf{\Pi}}$, we select the valid IVs by applying majority/plurality voting or finding the maximum clique of the voting matrix[47]. Let $\mathbf{VM}_k = \sum_{j \in \mathcal{S}} \widehat{\Pi}_{k,j}$ be the total number of pQTLs whose ratio estimates are similar to pQTL k. For example, $\mathbf{VM}_1 = 3$ in Figure 2, since three pQTLs (including pQTL 1 itself) yield similar ratio estimates to pQTL 1 according to criterion **C2**. A large $\mathbf{VM}_k$ implies strong evidence that pQTL k is a valid IV, since we assume that valid IVs form a plurality of the relevant IVs. Let $\hat{\mathcal{V}}_M = \{k \in \hat{\mathcal{S}} : \mathbf{VM}_k > |\hat{\mathcal{S}}|/2\}$ denote the set of IVs with majority voting, and $\hat{\mathcal{V}}_P = \{k \in \hat{\mathcal{S}} : \mathbf{VM}_k = \max_{l \in \hat{\mathcal{S}}} \mathbf{VM}_l\}$ denote the set of IVs with plurality voting, then the union $\hat{\mathcal{V}} = \hat{\mathcal{V}}_M \cup \hat{\mathcal{V}}_P$ can be a robust estimate of $\mathcal{V}$ in practice. Alternatively, we can also find the maximum clique in the voting matrix as an estimate of $\mathcal{V}$. A clique in the voting matrix is a group of IVs who mutually vote for each other to be valid, and the maximum clique is the clique with the largest possible number of IVs[47].

Estimation and inference of the causal effect

After selecting the set of valid pQTL instruments $\hat{\mathcal{V}}$, the causal effect β is estimated by

$$\hat{\beta}_{\mathrm{SPI}} = \frac{\hat{\mathbf{\Gamma}}_{\hat{\mathcal{V}}}^{T} \hat{\boldsymbol{\gamma}}_{\hat{\mathcal{V}}}}{\hat{\boldsymbol{\gamma}}_{\hat{\mathcal{V}}}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}_{\hat{\mathcal{V}}}} \tag{7}$$

where $\hat{\boldsymbol{\gamma}}_{\hat{\mathcal{V}}}$ and $\hat{\mathbf{\Gamma}}_{\hat{\mathcal{V}}}$ are the estimates of pQTL-protein associations and pQTL-outcome associations of the selected valid IVs in $\hat{\mathcal{V}}$, respectively. The MR-SPI estimator in equation (7) is the regression coefficient obtained by fitting a zero-intercept ordinary least squares regression of $\hat{\mathbf{\Gamma}}_{\hat{\mathcal{V}}}$ on $\hat{\boldsymbol{\gamma}}_{\hat{\mathcal{V}}}$. Since the pQTLs are standardized, the genetic

associations $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ are scaled by $\sqrt{2f_j(1-f_j)}$ (compared to the genetic associations calculated using the unstandardized pQTLs, denoted by $\check{\gamma}_j$ and $\check{\Gamma}_j$), where f_j is the minor allele frequency of pQTL j. As $f_j(1-f_j)$ is approximately proportional to the inverse variance of $\check{\Gamma}_j$ when each pQTL IV explains only a small proportion of variance in the outcome[85], the MR-SPI estimator of the causal effect in equation (7) is approximately equal to the inverse-variance weighted estimator[27] calculated with $\{\check{\gamma}_j, \check{\Gamma}_j\}_{j\in\hat{\mathcal{V}}}$.

Let $\alpha \in (0,1)$ be the significance level and $z_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the standard normal distribution, then the $(1-\alpha)$ confidence interval for β is given by:

$$\mathrm{CI} = \left(\hat{\beta}_{\mathrm{SPI}} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{SPI}})}, \hat{\beta}_{\mathrm{SPI}} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{SPI}})}\right) \tag{8}$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{SPI}})$ is the estimated variance of $\hat{\beta}_{\mathrm{SPI}}$, which can be found in Supplementary Section S4. As $\min\{n_1, n_2\} \to \infty$, we have $\mathbb{P}\left\{\beta \in \left(\hat{\beta}_{\mathrm{SPI}} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{SPI}})}, \hat{\beta}_{\mathrm{SPI}} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{SPI}})}\right)\right\} \to 1-\alpha$ under the plurality rule condition, as shown in Supplementary Section S5. Hence, MR-SPI provides a theoretical guarantee for the asymptotic coverage probability of the confidence interval under the plurality rule condition.

We summarize the proposed procedure of selecting valid IVs and constructing the corresponding confidence interval by MR-SPI in Algorithm 1.

A robust confidence interval via searching and sampling

In finite-sample settings, the selected set of relevant IVs $\hat{\mathcal{S}}$ might include some invalid IVs whose degrees of violation of (A2) and (A3) are small but nonzero, and we refer to them as "locally invalid IVs"[49]. When locally invalid IVs exist and are incorrectly selected into $\hat{\mathcal{V}}$, the confidence interval in equation (8) becomes unreliable, since its

validity (i.e., the coverage probability attains the nominal level) requires that the invalid IVs are correctly filtered out. In practice, we can multiply the threshold $\sqrt{\log \min(n_1, n_2)}$ in the right-hand side of equation (4) by a scaling factor η to examine whether the confidence interval calculated by equation (8) is sensitive to the choice of the threshold. If the confidence interval varies substantially to the choice of the scaling factor η, then there might exist finite-sample IV selection error especially with locally invalid IVs. We demonstrate this issue with two numerical examples presented in Supplementary Figure S14. Supplementary Figure S14(a) shows an example in which MR-SPI provides robust inference across different values of the scaling factor, while Supplementary Figure S14(b) shows an example that MR-SPI might suffer from finite-sample IV selection error, as the causal effect estimate and the corresponding confidence interval are sensitive to the choice of the scaling factor η. This issue motivates us to develop a more robust confidence interval.

To construct a confidence interval that is robust to finite-sample IV selection error, we borrow the idea of searching and sampling[49], with main steps described in Supplementary Figure S17. The key idea is to sample the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ repeatedly from the following distribution:

$$\begin{pmatrix} \widehat{\boldsymbol{\gamma}}^{(m)} \\ \widehat{\boldsymbol{\Gamma}}^{(m)} \end{pmatrix} \sim N\left[\begin{pmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\Gamma}} \end{pmatrix}, \begin{pmatrix} \frac{1}{n_1}\widehat{\mathbf{V}}_\gamma & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\widehat{\mathbf{V}}_\Gamma \end{pmatrix}\right], m = 1, \cdots, M \tag{9}$$

where M is the number of sampling times (by default, we set $M = 1{,}000$). Since $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ follow distributions centered at $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$, there exists m^* such that $\widehat{\boldsymbol{\gamma}}^{(m^*)}$ and $\widehat{\boldsymbol{\Gamma}}^{(m^*)}$ are close enough to the true values $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ when the number of sampling times M is sufficiently large, and thus the confidence interval obtained by using $\widehat{\boldsymbol{\gamma}}^{(m^*)}$ and $\widehat{\boldsymbol{\Gamma}}^{(m^*)}$ instead of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ might have a larger probability of covering β.

For each sampling, we construct the confidence interval by searching over a grid of β values such that more than half of the selected IVs in $\widehat{\mathcal{V}}$ are detected as valid. As for the choice of grid, we start with the smallest interval $[L, U]$ that contains all the following intervals:

$$\left(\hat{\beta}^{[j]} - \sqrt{\log \min(n_1, n_2)\widehat{\mathrm{Var}}\left(\hat{\beta}^{[j]}\right)}, \hat{\beta}^{[j]} + \sqrt{\log \min(n_1, n_2)\widehat{\mathrm{Var}}\left(\hat{\beta}^{[j]}\right)}\right) \text{ for } j \in \widehat{\mathcal{V}} \tag{10}$$

where $\hat{\beta}^{[j]}$ is the ratio estimate of the jth pQTL IV, $\widehat{\mathrm{Var}}(\widehat{\beta^{[j]}}) = \left(\widehat{\mathbf{V}}_{\Gamma,jj}/n_2 + \left(\hat{\beta}^{[j]}\right)^2 \widehat{\mathbf{V}}_{\gamma,jj}/n_1\right)/\hat{\gamma}_j^2$ is the variance of $\hat{\beta}^{[j]}$, and $\sqrt{\log \min(n_1, n_2)}$ serves the same purpose as in equation (4). Then we discretize $[L, U]$ into $\mathcal{B} = \{b_1, b_2, \cdots, b_K\}$ as the grid set such that $b_1 = L, b_K = U$ and $|b_{k+1} - b_k| = n_{\min}^{-0.6}$ for $1 \le k \le K-2$, where $n_{\min} = \min(n_1, n_2)$. We set the grid size $n_{\min}^{-0.6}$ so that the error caused by discretization is smaller than the parametric rate $n_{\min}^{-1/2}$.

For each grid value $b \in \mathcal{B}$ and sampling index $1 \le m \le M$, we propose an estimate of π_j by $\hat{\pi}_j^{(m)}(b) = \left(\hat{\Gamma}_j^{(m)} - b\hat{\gamma}_j^{(m)}\right) \cdot I\left(\left|\hat{\Gamma}_j^{(m)} - b\hat{\gamma}_j^{(m)}\right| \ge \lambda \hat{\rho}_j(b, \alpha)\right)$ for $j \in \hat{\mathcal{V}}$, where $\hat{\rho}_j(b, \alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{2|\hat{\mathcal{V}}|}\right)\sqrt{\left(\widehat{\mathbf{V}}_{\Gamma,jj}/n_2 + b^2 \widehat{\mathbf{V}}_{\gamma,jj}/n_1\right)}$ is a data-dependent threshold, $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution, $\alpha \in (0,1)$ is the significance level, and $\lambda = (\log \min(n_1, n_2)/M)^{\frac{1}{2|\hat{\mathcal{V}}|}}$($\lambda < 1$ when M is sufficiently large) is a scaling factor to make the thresholding more stringent so that the confidence interval in each sampling is shorter, as we will show shortly. Here, $\hat{\pi}_j^{(m)}(b) = 0$ indicates that the jth pQTL is detected as a valid IV in the m th sampling if we take $\left\{\hat{\boldsymbol{\gamma}}^{(m)}, \hat{\boldsymbol{\Gamma}}^{(m)}\right\}$ as the estimates of genetic associations and b as the true causal effect. Let $\hat{\boldsymbol{\pi}}_{\hat{\mathcal{V}}}^{(m)}(b) = \left(\hat{\pi}_j^{(m)}(b)\right)_{j \in \hat{\mathcal{V}}}$, then we construct the m th sampling's pseudo confidence interval $\mathrm{pCI}^{(m)}$ by searching for the smallest and largest $b \in \mathcal{B}$ such that more than half of pQTLs in $\hat{\mathcal{V}}$ are detected to be valid. Define $\beta_{\min}^{(m)} = \min\left\{b \in \mathcal{B}: \left\|\hat{\boldsymbol{\pi}}_{\hat{\mathcal{V}}}^{(m)}(b)\right\|_0 < |\hat{\mathcal{V}}|/2\right\}$ and $\beta_{\max}^{(m)} = \max\left\{b \in \mathcal{B}: \left\|\hat{\boldsymbol{\pi}}_{\hat{\mathcal{V}}}^{(m)}(b)\right\|_0 < |\hat{\mathcal{V}}|/2\right\}$, then the m th sampling's pseudo confidence interval is constructed as $\mathrm{pCI}^{(m)} = \left(\beta_{\min}^{(m)}, \beta_{\max}^{(m)}\right)$.

From the definitions of $\hat{\pi}_j^{(m)}(b)$ and $\mathrm{pCI}^{(m)}$, we can see that, when λ is smaller, there will be fewer pQTLs in $\hat{\mathcal{V}}$ being detected as valid for a given $b \in \mathcal{B}$, which leads to fewer $b \in \mathcal{B}$ satisfying $\left\|\hat{\boldsymbol{\pi}}_{\hat{\mathcal{V}}}^{(m)}(b)\right\|_0 < |\hat{\mathcal{V}}|/2$, thus the pseudo confidence interval in each sampling will be shorter. If there does not exist $b \in \mathcal{B}$ such that the majority

of IVs in $\hat{\mathcal{V}}$ are detected as valid, we set $\mathrm{pCI}^{(m)} = \emptyset$. Let $\mathcal{M} = \{1 \leq m \leq M: \mathrm{pCI}^{(m)} \neq \emptyset\}$ denote the set of all sampling indexes corresponding to non-empty searching confidence intervals, then the proposed robust confidence interval is given by:

$$\mathrm{CI}^{\mathrm{robust}} = \left(\min_{m\in\mathcal{M}} \beta_{\min}^{(m)}, \max_{m\in\mathcal{M}} \beta_{\max}^{(m)}\right) \tag{11}$$

We summarize the procedure of constructing the proposed robust confidence interval in Algorithm 2.

Simulation settings

We set the number of candidate IVs $p = 10$, as the average number of candidate pQTL IVs for the plasma proteins in the UK Biobank proteomics data is around 7.4. We set the sample sizes $n_1 = n_2 = 5{,}000, 10{,}000, 20{,}000, 40{,}000,$ or $80{,}000$. We generate the j th genetic instruments Z_j and X_j independently from a binomial distribution $\mathrm{Bin}\left(2, f_j\right)$, where $f_j \sim U(0.05{,}0.50)$ is the minor allele frequency of pQTL j. Then we generate the protein level $\boldsymbol{D} = \left(D_1, \cdots, D_{n_1}\right)^\top$ and the outcome $\boldsymbol{Y} = \left(Y_1, \cdots, Y_{n_2}\right)^\top$ according to the exposure model and the outcome model, respectively. Finally, we calculate the genetic associations and their corresponding standard errors for the protein and the outcome, respectively. As for the parameters, we fix the causal effect $\beta = 1$, and we consider 4 settings for $\gamma \in \mathbb{R}^p$ and $\boldsymbol{\pi} \in \mathbb{R}^p$:

(**S1**): set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\top$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_6, \mathbf{1}_4)^\top$.

(**S2**): set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\top$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_4, \mathbf{1}_3, -\mathbf{1}_3)^\top$.

(**S3**): set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\top$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_6, \mathbf{1}_2, 0.25{,}0.25)^\top$.

(**S4**): set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\top$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_4, \mathbf{1}_2, 0.25, \mathbf{1}_2, -0.25)^\top$.

Settings (**S1**) and (**S3**) satisfy the majority rule condition, while (**S2**) and (S4) only satisfy the plurality rule condition. In addition, (**S3**) and (**S4**) simulate the cases where locally invalid IVs exist, as we shrink some of the pQTLs' violation degrees of assumptions (A2) and (A3) down to 0.25 times in these two settings. In total, we run 1,000 replications in each setting.

Implementation of existing MR methods

We compare the performance of MR-SPI with eight other MR methods in simulation studies and real data analyses. These methods are implemented as follows:

- Random-effects IVW, MR-Egger, the weighted median method, the mode-based estimation and the contamination mixture method are implemented in the R package "MendelianRandomization" (https://github.com/cran/MendelianRandomization). The mode-based estimation is run with "iteration $= 1000$". All other methods are run with the default parameters.

- MR-PRESSO is implemented in the R package "MR-PRESSO" (https://github.com/ rondolab/MR-PRESSO) with outlier test and distortion test.

- MR-RAPS is performed using the R package "mr.raps" (https://github.com/qingyuanzhao/ mr.raps) with the default options.

- MRMix is run with the R package "MRMix" (https://github.com/gqi/MRMix) using the default options.

Algorithm 1: Selecting pQTL IVs and Performing Causal Inference by MR-SPI

Input: Summary statistics of independent pQTLs $\left\{\hat{\gamma}_j, \hat{\sigma}_{\gamma_j}, \hat{\Gamma}_j, \hat{\sigma}_{\Gamma_j}\right\}_{1 \le j \le p}$; Sample sizes n_1 for the proteomics data and n_2 for the outcome; Threshold α^* for selecting relevant IVs; Significance level $\alpha \in (0,1)$.

Output: An estimate of the set of valid IVs $\hat{\mathcal{V}}$, the causal effect estimate $\hat{\beta}_{\mathrm{SPI}}$ and the corresponding confidence interval CI.

Estimate the variance-covariance matrices $\widehat{\mathbf{V}}_{\gamma}$ and $\widehat{\mathbf{V}}_{\Gamma}$ as in equation (1);

Select the set of relevant IVs $\hat{\mathcal{S}}$ as in equation (2);

for $j \in \hat{\mathcal{S}}$ **do**

    Calculate $\hat{\beta}^{[j]} = \hat{\Gamma}_j / \hat{\gamma}_j$ and $\hat{\pi}_k^{[j]} = \hat{\Gamma}_k - \hat{\beta}^{[j]} \hat{\gamma}_k$ for $k \in \hat{\mathcal{S}}$;

    Each relevant IV $k \in \hat{\mathcal{S}}$ votes for the j th IV to be valid if $\left|\hat{\pi}_k^{[j]}\right| / \widehat{\mathrm{SE}}\left(\hat{\pi}_k^{[j]}\right) \le \sqrt{\log \min(n_1, n_2)}$;

end

Construct the symmetric voting matrix $\widehat{\mathbf{\Pi}} \in \mathbb{R}^{|\hat{\mathcal{S}}| \times |\hat{\mathcal{S}}|}$ as in equation (6);

Select the set of valid IVs $\hat{\mathcal{V}}$ by majority voting, plurality voting or finding the maximum clique in the voting matrix;

Estimate the causal effect as in equation (7), and construct the corresponding confidence interval as in equation (8) using the selected valid IVs in $\hat{\mathcal{V}}$.

Algorithm 2: Constructing A Robust Confidence Interval via Searching and Sampling

Input: Summary statistics of independent pQTLs $\left\{\hat{\gamma}_j, \hat{\sigma}_{\gamma_j}, \hat{\Gamma}_j, \hat{\sigma}_{\Gamma_j}\right\}_{1 \le j \le p}$; Sample sizes n_1 for the proteomics data and n_2 for the outcome; Threshold α^* for selecting relevant IVs; Significance level $\alpha \in (0,1)$; Sampling number M.

Output: The robust confidence interval $\mathrm{CI}^{\mathrm{robust}}$.

Estimate the set of valid IVs $\hat{\mathcal{V}}$ as in Algorithm 1;

Construct the initial interval $[L, U]$ as in equation (10) and obtain the corresponding grid set $\mathcal{B}$;

for $m \leftarrow 1$ to M **do**

  Sample $\hat{\boldsymbol{\gamma}}^{(m)}$ and $\hat{\boldsymbol{\Gamma}}^{(m)}$ from the distribution in equation (9);

  Calculate $\left\{\hat{\boldsymbol{\pi}}_{\hat{\mathcal{V}}}^{(m)}(b)\right\}_{b \in \mathcal{B}}$ by $\hat{\pi}_j^{(m)}(b) = \left(\hat{\Gamma}_j^{(m)} - b\hat{\gamma}_j^{(m)}\right) \cdot I\left(\left|\hat{\Gamma}_j^{(m)} - b\hat{\gamma}_j^{(m)}\right| \ge \lambda\hat{\rho}_j(b, \alpha)\right), j \in \hat{\mathcal{V}}$;

  Construct $\mathrm{pCI}^{(m)}$ by grid search over $\mathcal{B}$;

end

Construct the robust confidence interval $\mathrm{CI}^{\mathrm{robust}}$ as in equation (11).

References

1. Self, W.K. and Holtzman, D.M. (2023) Emerging diagnostics and therapeutics for Alzheimer disease. *Nature medicine* **29**, 2187-2199.
2. Nichols, E. *et al.* (2022) Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health* **7**, e105-e125.
3. Van Dyck, C.H. *et al.* (2023) Lecanemab in early Alzheimer's disease. *New England Journal of Medicine* **388**, 9-21.
4. Sims, J.R. *et al.* (2023) Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. *Jama* **330**, 512-527.
5. Hardy, J.A. and Higgins, G.A. (1992) Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256**, 184-185.
6. Karran, E. and De Strooper, B. (2022) The amyloid hypothesis in Alzheimer disease: new insights from new therapeutics. *Nature reviews Drug discovery* **21**, 306-318.
7. Khoury, R., Rajamanickam, J. and Grossberg, G.T. (2018) An update on the safety of current therapies for Alzheimer's disease: focus on rivastigmine. *Therapeutic advances in drug safety* **9**, 171-178.
8. Davey Smith, G. and Ebrahim, S. (2003) ' Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1--22.
9. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N. and Davey Smith, G. (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133--1163.
10. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**, R89--R98.
11. Davey Smith, G. and Ebrahim, S. (2004) Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33**, 30--42.
12. Burgess, S., Butterworth, A. and Thompson, S.G. (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* **37**, 658--665.
13. Pierce, B.L. and Burgess, S. (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology* **178**, 1177--1184.
14. Lawlor, D.A. (2016) Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology* **45**, 908.
15. Slob, E.A.W. and Burgess, S. (2020) A comparison of robust Mendelian randomization methods using summary data. *Genetic Epidemiology* **44**, 313--329.
16. Zhao, Q., Wang, J., Hemani, G., Bowden, J. and Small, D.S. (2020) Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics* **48**, 1742--1769.
17. Morrison, J., Knoblauch, N., Marcus, J.H., Stephens, M. and He, X. (2020) Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics* **52**, 740--747.

18. Cheng, Q., Zhang, X., Chen, L.S. and Liu, J. (2022) Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nature Communications* **13**, 1--13.
19. Yao, C. *et al.* (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature communications* **9**, 3268.
20. Sun, B.B. *et al.* (2022) Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *BioRxiv*, 2022--06.
21. Didelez, V. and Sheehan, N. (2007) Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**, 309--330.
22. Sanderson, E., Richardson, T.G., Hemani, G. and Davey Smith, G. (2021) The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International Journal of Epidemiology* **50**, 1350--1361.
23. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. (2013) Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14**, 483--495.
24. Sivakumaran, S. *et al.* (2011) Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics* **89**, 607--618.
25. Parkes, M., Cortes, A., Van Heel, D.A. and Brown, M.A. (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics* **14**, 661--673.
26. Schmidt, A.F. *et al.* (2020) Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255.
27. Bowden, J. *et al.* (2017) A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine* **36**, 1783--1802.
28. Bowden, J., Davey Smith, G. and Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512--525.
29. Verbanck, M., Chen, C.-Y., Neale, B. and Do, R. (2018) Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693--698.
30. Bowden, J., Davey Smith, G., Haycock, P.C. and Burgess, S. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* **40**, 304--314.
31. Hartwig, F.P., Davey Smith, G. and Bowden, J. (2017) Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology* **46**, 1985--1998.
32. Qi, G. and Chatterjee, N. (2019) Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications* **10**, 1--10.
33. Burgess, S., Foley, C.N., Allara, E., Staley, J.R. and Howson, J.M.M. (2020) A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications* **11**, 1--11.
34. He, B., Shi, J., Wang, X., Jiang, H. and Zhu, H.-J. (2020) Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC biology* **18**, 1-16.
35. Swerdlow, D.I. *et al.* (2016) Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology* **45**, 1600--

1616.
36. Zaneveld, J.R., McMinds, R. and Vega Thurber, R. (2017) Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nature microbiology* **2**, 1-8.
37. Shriner, D. and Rotimi, C.N. (2018) Whole-genome-sequence-based haplotypes reveal single origin of the sickle allele during the holocene wet phase. *The American Journal of Human Genetics* **102**, 547-556.
38. Karki, R., Pandya, D., Elston, R.C. and Ferlini, C. (2015) Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC medical genomics* **8**, 1-7.
39. Ashley-Koch, A., Yang, Q. and Olney, R.S. (2000) Sickle hemoglobin (Hb S) allele and sickle cell disease: a HuGE review. *American journal of epidemiology* **151**, 839-845.
40. Rees, D.C., Williams, T.N. and Gladwin, M.T. (2010) Sickle-cell disease. *The Lancet* **376**, 2018-2031.
41. Kato, G.J. *et al.* (2018) Sickle cell disease. *Nature reviews Disease primers* **4**, 1-22.
42. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589.
43. Mirdita, M. *et al.* (2022) ColabFold: making protein folding accessible to all. *Nature methods* **19**, 679-682.
44. Wayment-Steele, H.K. *et al.* (2023) Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, 1-3.
45. Abramson, J. *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3.
46. Dunn, O.J. (1961) Multiple comparisons among means. *Journal of the American Statistical Association* **56**, 52--64.
47. Ouyang, Q., Kaplan, P.D., Liu, S. and Libchaber, A. (1997) DNA solution of the maximal clique problem. *Science* **278**, 446--449.
48. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289--300.
49. Guo, Z. (2023) Causal inference with invalid instruments: post-selection problems and a solution using searching and sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85**, 959-985.
50. Jansen, I.E. *et al.* (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics* **51**, 404--413.
51. Naj, A.C. *et al.* (2011) Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature genetics* **43**, 436--441.
52. Rathore, N. *et al.* (2018) Paired Immunoglobulin-like Type 2 Receptor Alpha G78R variant alters ligand binding and confers protection to Alzheimer's disease. *PLoS genetics* **14**, e1007427.
53. Griciuc, A. *et al.* (2019) TREM2 acts downstream of CD33 in modulating microglial pathology in Alzheimer's disease. *Neuron* **103**, 820--835.
54. Helgadottir, H.T. *et al.* (2019) Somatic mutation that affects transcription factor binding upstream of CD55 in the temporal cortex of a late-onset Alzheimer disease patient. *Human Molecular Genetics* **28**, 2675--2685.
55. Rimal, S. *et al.* (2023) Reverse electron transfer is activated during aging and contributes

to aging and age-related disease. *EMBO reports* **24**, e55548.
56. Winfree, R.L. *et al.* (2023) TREM2 gene expression associations with Alzheimer's disease neuropathology are region-specific: implications for cortical versus subcortical microglia. *Acta Neuropathologica* **145**, 733-747.
57. Yang, X. *et al.* (2023) Functional characterization of Alzheimer's disease genetic variants in microglia. *Nature Genetics*, 1-10.
58. (2023) UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research* **51**, D523-D531.
59. Zhou, Y. *et al.* (2022) Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Research* **50**, D1398-D1407.
60. Wishart, D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* **46**, D1074-D1082.
61. Bross, P.F. *et al.* (2001) Approval summary: gemtuzumab ozogamicin in relapsed acute myeloid leukemia. *Clinical cancer research* **7**, 1490-1496.
62. Norsworthy, K.J. *et al.* (2018) FDA approval summary: mylotarg for treatment of patients with relapsed or refractory CD33-positive acute myeloid leukemia. *The oncologist* **23**, 1103-1108.
63. Kim, J. *et al.* (2021) FDA approval summary: pralsetinib for the treatment of lung and thyroid cancers with RET gene mutations or fusions. *Clinical Cancer Research* **27**, 5452-5456.
64. Bradford, D. *et al.* (2021) FDA approval summary: selpercatinib for the treatment of lung and thyroid cancers with RET gene mutations or fusions. *Clinical Cancer Research* **27**, 2130-2135.
65. Griciuc, A. *et al.* (2013) Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. *Neuron* **78**, 631-643.
66. Tortora, F. *et al.* (2022) CD33 rs2455069 SNP: correlation with alzheimer's disease and hypothesis of functional role. *International Journal of Molecular Sciences* **23**, 3629.
67. Kamat, M.A. *et al.* (2019) PhenoScanner V2: an expanded tool for searching human genotype--phenotype associations. *Bioinformatics* **35**, 4851--4853.
68. Faux, N.G. *et al.* (2014) An anemia of Alzheimer's disease. *Molecular Psychiatry* **19**, 1227--1234.
69. Winchester, L.M., Powell, J., Lovestone, S. and Nevado-Holgado, A.J. (2018) Red blood cell indices and anaemia as causative factors for cognitive function deficits and for Alzheimer's disease. *Genome Medicine* **10**, 1--12.
70. Raudvere, U. *et al.* (2019) g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191--W198.
71. Rijpma, A., van der Graaf, M., Meulenbroek, O., Rikkert, M.G.O. and Heerschap, A. (2018) Altered brain high-energy phosphate metabolism in mild Alzheimer's disease: A 3-dimensional 31P MR spectroscopic imaging study. *NeuroImage: Clinical* **18**, 254-261.
72. Parasoglou, P. *et al.* (2022) Phosphorus metabolism in the brain of cognitively normal midlife individuals at risk for Alzheimer's disease. *Neuroimage: Reports* **2**, 100121.
73. Lazarczyk, M.J. *et al.* (2016) Major Histocompatibility Complex class I proteins are critical for maintaining neuronal structural complexity in the aging brain. *Scientific reports* **6**,

26199.
74. Kim, M.-S. *et al.* (2023) Neuronal MHC-I complex is destabilized by amyloid-β and its implications in Alzheimer's disease. *Cell & Bioscience* **13**, 181.
75. Le Guen, Y. *et al.* (2023) Multiancestry analysis of the HLA locus in Alzheimer's and Parkinson's diseases uncovers a shared adaptive immune response mediated by HLA-DRB1* 04 subtypes. *Proceedings of the National Academy of Sciences* **120**, e2302720120.
76. Giambartolomei, C. *et al.* (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383.
77. Barfield, R. *et al.* (2018) Transcriptome - wide association studies accounting for colocalization using Egger regression. *Genetic epidemiology* **42**, 418-433.
78. Hukku, A. *et al.* (2021) Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *The American Journal of Human Genetics* **108**, 25-35.
79. Zuber, V. *et al.* (2022) Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics* **109**, 767-782.
80. Yavorska, O.O. and Burgess, S. (2017) MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International journal of epidemiology* **46**, 1734-1739.
81. Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559--575.
82. Bulik-Sullivan, B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291--295.
83. Guo, Z., Kang, H., Tony Cai, T. and Small, D.S. (2018) Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 793--815.
84. Kang, H., Zhang, A., Cai, T.T. and Small, D.S. (2016) Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association* **111**, 132-144.
85. Burgess, S., Dudbridge, F. and Thompson, S.G. (2016) Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine* **35**, 1880-1906.