

Causal inference for multiple risk factors and diseases from genomics data

Received: 18 April 2025

Accepted: 7 August 2026

Published online: 25 August 2026



Nick Machnik ^{1,5}✉, Mahdi Mahmoudi ^{2,5}, Malgorzata Borczyk ³,
Ilse Krätschmer ¹, Markus J. Bauer⁴ & Matthew R. Robinson ¹✉

Mendelian randomization, the standard instrument variable causal-inference approach for genomics, can be severely biased by inadequate selection of appropriate instrument variables and horizontal pleiotropy. We introduce Causal Inference GWAS, a graphical approach that selects genetic instruments from summary statistics controlling for linkage and pleiotropy, accommodates rare variants and binary outcomes, distinguishes direct from indirect risk factors in high dimensions, and flags latent confounding. Applied to 9 risk factors, 4 metabolic disease outcomes, and 8.4M variants across 458,747 UK Biobank individuals, our method runs in 20 minutes, identifying only 696 genome-wide valid instruments. We replicate nearly all previously reported risk factor-outcome paths, but find that most are indistinguishable from unmeasured confounding. This suggests robust biobank causal inference will require longitudinal and family data to resolve temporal precedence from reverse causation. Our approach offers a principled first step toward screening genuine causal signal from phenotypic correlation in biobank data.

Distinguishing causation from correlation is fundamentally important for understanding the etiology of common complex diseases. Randomized controlled trials (RCTs) are the gold standard for causal inference, but their difficulty, expense, and ethical considerations often prohibit testing potential disease-associated risk factors across the vast array of human complex traits. As a result, the most commonly used methods for statistical causal inference in a genetic or biomedical setting center around the instrumental variable (IV) framework of Mendelian randomization (MR)^{1,2}. A wide variety of MR approaches have been developed to test whether there exists a potential causal effect of one or more mediator variable(s) onto one or more outcome variable(s) using genome-wide association study (GWAS) data^{3–16}.

MR relies on core assumptions that: (1) the genetic instrumental variables are associated with the exposure (the relevance assumption); (2) there are no measured or unmeasured confounding variables of the instrumental variable and the outcome Y (termed the independence assumption); and (3) all paths between the instrumental variable and

the outcome go through the exposure (exclusion restriction or no horizontal pleiotropy assumption). These assumptions are notoriously difficult to verify directly in empirical applications, and violations of independence and exclusion restriction in particular can substantially inflate type I error, producing spurious claims of causation where none exists¹⁷.

Theoretical developments in statistical causal inference show that graphical modeling can be used to learn the underlying causal structure of hundreds of variables, whilst controlling for potential unobserved confounding within a single model^{18–20}. As we describe below, combining graphical inference with standard MR methods has the potential to prioritize exposure-outcome pairs and to select instrumental variables in a principled manner by explicitly verifying MR assumptions. However, there are currently no algorithms capable of conducting this form of causal discovery at scale. Developing these approaches to accommodate highly correlated variables across diverse distributions and millions of highly correlated genetic markers, whilst allowing data to be combined across biobanks, would facilitate

¹Institute Of Science and Technology Austria, Klosterneuburg, Austria. ²Faculty of Medicine, Sigmund Freud University, Vienna, Austria. ³Laboratory of Pharmacogenomics, Maj Institute of Pharmacology PAS, Krakow, Poland. ⁴Boehringer Ingelheim RCV GmbH & Co KG, Vienna, Austria. ⁵These authors contributed equally: Nick Machnik, Mahdi Mahmoudi. ✉e-mail: nick.machnik@ist.ac.at; matthew.robinson@ist.ac.at

accurate, high-throughput screening of potential causal links between modifiable risk factors and disease outcomes.

Here, we propose an approach for conducting Causal Inference in Genome-Wide Association Studies (CI-GWAS) that can infer a single graph at scale describing the relationships among millions of genetic markers, tens of risk factors and multiple disease outcomes. We utilize the inferred graph for more principled detection of potential IVs, controlling for linkage disequilibrium (LD) genome-wide and providing better empirical checking of core MR assumptions, in particular the absence of horizontal pleiotropy and potential latent confounding of the marker outcome relationship. In simulations, we show that CI-GWAS selects appropriate IVs and that this improves upon the high false discovery rates of current MR approaches under violation of the INSIDE assumption (i.e., correlated pleiotropy / heritable confounding). Our approach can accommodate rare variants and rare disease outcomes by placing relationships on the liability scale to correct for differences in outcome prevalence. Applying CI-GWAS to summary statistic data of 8.4 M SNPs and 15 phenotypes requires only 20 minutes of wall-clock compute time on a standard GPU server. We find that few GWAS variants are suitable MR IVs, but that many previously reported paths among risk factors and outcomes¹¹ can be recovered within both univariate and multivariable methods. Further investigation reveals that within the UK Biobank, unmeasured latent confounding cannot be ruled out as a potential underlying explanation for the phenotypic relationships that we report. CI-GWAS can be applied to individual-level or summary-statistic data and provides a potential solution for facilitating robust one- or two-sample MR inference on large-scale genotype-phenotype data.

Results

The CI-GWAS algorithm

Here, we present CI-GWAS, an algorithm designed to learn a graphical model that describes the causal relationships among genetic and phenotypic variables. CI-GWAS constructs such a model in two steps: (1) by learning a skeleton, an undirected graph over the set of observed variables; (2) by directing the edges in the skeleton to obtain a partial ancestral graph (PAG), while taking into account potential confounding through latent variables. The inferred graph encodes which causal relations are compatible with the data's conditional independence structure. In the following, we show how the learned graph can be used in conjunction with an MR method to infer a set of potential causal relationships and their effects.

In the first step, the skeleton is learned by a modified version of the PC-stable algorithm²¹, in which the correlations between observed pairs of variables are tested for significance, conditional on sets of other observed variables (see Methods for a thorough explanation). All edges between pairs for which any test has determined the association to be insignificant are removed from the graph. Thus, for any adjacent pair of variables in the final skeleton, there exists a degree of correlation that cannot be explained away by conditioning on any subset of observed variables.

One very useful property of the learned skeleton is that it enables the identification of likely valid instrumental variables for MR. We explicitly define valid MR instruments in a finite sample as those that are associated with the exposure but not with the outcome after conditioning on the exposure. We can identify valid IVs based on the conditional independence tests conducted, which ensure that the common MR assumptions about instrumental variables hold. In particular, by inspecting the learned graph, one can verify the following assumptions: (1) that the genetic instrumental variables are associated with the exposure; (2) that there are no measured or unmeasured confounding variables of the instrumental variable and the outcome Y ; and (3) that all paths between the instrumental variable and the outcome go through the exposure. The conditions that must be met in the learned graph for these assumptions to hold are simple: for a given

exposure trait X and an outcome trait Y , any marker G , found to be adjacent to X and not adjacent to Y and for which all paths to Y contain X , is a valid instrument to test the effect of X on Y , under the assumption that the learned skeleton is faithful to the underlying distribution.

These conditions ensure that the relevance assumption holds for (G, X) , because G and X are adjacent only if there is a significant effect of G on X that cannot be explained away by conditioning on any other measured variables. The independence assumption also holds for (G, Y) , because, if G and Y are affected by unobserved latent confounding, a dependency is induced and G and Y remain adjacent in the skeleton, since no conducted test can remove the dependency. If they are, on the other hand, affected by a common confounding variable that is measured and therefore in the graph, then there is a path from G to Y that does not contain X , namely through that confounding variable. The exclusion restriction also holds for (G, Y) by similar arguments: if G has a direct effect on Y , they are adjacent in the graph. If G has an effect on Y that is mediated by a third variable L , this path can be either read off the graph (if L is measured) or again appears as an adjacency of G and Y if L is latent.

However, this approach has limitations. In particular, there exist configurations of effects and effect sizes under which these rules will produce false positive and also false negative IVs in the presence of unmeasured confounding. First, it is possible that (G, Y) are not adjacent in the skeleton even if there is a path from G to Y that does not include X . This happens when there exist multiple paths from G to Y in the data-generating structure whose total effect sizes cancel each other out^{22,23}. If not all relevant variables are measured, the latent paths will not be visible in the graph, and G will be marked as a valid IV, thereby making a false positive error. Second, in the presence of an unmeasured confounder U of X and Y , G and Y may stay adjacent in the skeleton due to collider bias, if the association through the path $G \rightarrow X \leftarrow U \rightarrow Y$ is large enough in magnitude to be significant in the conditional independence test. Our rules will produce false-negative errors in this situation, removing potentially valid instruments. In summary, this shows that by inspecting the learned skeleton, one can select variables for which the above-stated MR assumptions hold, assuming no latent confounding. However, the selection may be flawed if not all relevant variables are measured. We investigate to what extent confounding impacts the performance of our algorithm in extensive simulation experiments below.

Our procedure to learn the skeleton is based on a GPU-accelerated version of the PC-stable algorithm^{21,24}, which we additionally accelerated by iterative variable selection and by exploiting the LD-block structure of the genome. This enables us to construct a complete marker-trait graph, for millions of markers and tens of traits, despite the exponential run-time scaling in the number of variables of the PC algorithm. One challenge here is that the PC algorithm relies on partial correlations vanishing for (conditionally) independent variables, which generally holds for Gaussian variables? but not necessarily for binary or ordinal variables. We find that treating binary case-control outcomes like Gaussian variables can lead to largely elevated numbers of false positive findings, especially for independence tests between markers with very low minor allele frequencies and binary traits with very unbalanced numbers of cases and controls (Supplementary Fig. 1). As has been shown in previous work²⁵, a heterologous correlation matrix with Pearson correlations between Gaussian, polyserial between Gaussian and ordinal and polychoric correlations between ordinal variables can be used to prevent this issue. This effectively places binary variables on the liability scale and has the additional advantage of allowing comparisons among traits at the same scale, accounting for differences in outcome prevalence. At the sample size, lowest minor allele frequency and lowest trait prevalence in the UK Biobank data used in this study, using polyserial correlations between

markers and case-control traits is effective in controlling the numbers of false positives (Supplementary Fig. 1).

This first step of the algorithm can be combined with the second to prioritize exposure-outcome pairs, through selection of risk factors that are most likely to have a significant directed partial correlation with a particular outcome. The second step is based on the RFCI algorithm²⁰, which attempts to direct the edges of the learned skeleton, testing for the presence of latent confounding. We introduce a novel way to obtain order-independent separation sets that offer a conservative way of orienting trait relationships, that is, to learn which phenotypes have evidence for a potential directed relationship with each outcome. We also incorporate prior knowledge that markers precede traits temporally, which dramatically reduces computation time. We denote our version of RFCI as sRFCI. For further details, see Methods.

Taken together, we improve upon previously proposed graphical methods that were limited to few hundreds²⁶ or thousands²⁷ of genetic variables or relied on the construction of separate graphs for trait-trait and marker-trait relationships with pre-selected genetic variants by GWAS¹⁷. Combining graphical inference with standard MR methods takes advantage of the principled instrumental variable selection and the identification of relevant risk factors. This also advances previous work on graphical inference in MR¹⁷ because now marker selection is made through joint estimation, where many variables can be conditioned on at once, and pleiotropic markers can be identified prior to MR inference, promising more accurate instrumental variable selection and improved MR performance as shown below.

CI-GWAS takes as input: (1) the matrix of marker-marker correlations (an LD matrix); (2) a matrix of marker-trait correlations; (3) a matrix of trait-trait correlations. This makes CI-GWAS applicable to summary-statistic data, since LD matrices, effect sizes of marker-trait associations, minor allele frequencies, trait variances and trait-trait correlations can be shared whilst preserving privacy.

CI-GWAS controls for horizontal pleiotropy in instrumental variable selection

Here, we study how the usage of IVs selected by CI-GWAS in MR compares to standard GWAS-based IV selection. To this end, we conduct a simulation study using 654,173 imputed genetic markers recorded on 458,747 UK Biobank individuals (see Methods). We simulate data according to the graph generally considered in MR studies, shown in Fig. 1a.

We generate strong unobserved confounding, U , influencing both the exposure and outcome of interest, and focus on the three MR assumptions, given above, by simulating six scenarios where: (1) genetic effects are strong and therefore easily identifiable and there is no horizontal pleiotropy, or latent genetic confounding (labeled: baseline, high power); (2) genetic effects are weak and therefore more difficult to identify, effects are estimated with higher error and there is no horizontal pleiotropy or latent genetic confounding (labeled: baseline, low power); (3) genetic effects are easily identifiable and there is latent genetic confounding (latent pleiotropy, high power); (4) genetic effects are difficult to identify and there is latent genetic confounding (latent pleiotropy, low power); (5) genetic effects are easily identifiable and there is both latent genetic confounding and horizontal pleiotropy (horizontal pleiotropy, high power), which is also termed correlated pleiotropy or heritable confounding in the literature; (6) genetic effects are difficult to identify and there is both latent genetic confounding and horizontal pleiotropy (horizontal pleiotropy, low power). For each scenario, we simulate the exposure to be either causal for the outcome with average causal effect $ACE = 0.2$, or non-causal for the outcome $ACE = 0.0$, resulting in 12 settings for which we simulate 100 replicates. Thus, we generate 1200 data sets in total, each with a sample size of 458,747 individuals and a feature size of 654,172 SNPs.

For each of these data sets, we select IVs in two ways. First, we run standard marginal linear regression for X and Y at each of the 654,172 SNPs and then select IVs as an independent set of SNPs (within 1Mb, with $LD R^2 < 0.001$) associated with X with $p\text{-value} \leq 5 \times 10^{-8}$, as is commonly done in MR studies. Second, we use CI-GWAS to select the IVs as described above: For a given exposure trait X and an outcome trait Y , any marker G , found to be adjacent to X and not adjacent to Y and for which all paths to Y contain X , is discovered as a valid instrument to test the effect of X on Y . We then use these IVs alongside their marginal GWAS regression coefficients in two commonly used MR approaches of inverse-variance weighting (IVW) and MR-Egger¹² to estimate the causal effect of the exposure X and outcome Y .

We find that selecting genetic IVs from CI-GWAS allows unbiased estimation of the ACE across all scenarios considered here, when using both IVW and MR-Egger, as shown Fig. 1b. This accurate ACE estimation is further reflected in the false positive rate (FPR) of zero for each of the twelve settings (Fig. 1c). In contrast to that, we see a considerable effect of horizontal pleiotropy for both MR-methods used with GWAS selected IVs. Note here, FPR measures the proportion of simulation replicates where $ACE = 0$, but MR methods incorrectly identify a causal effect of the exposure X and outcome Y (at $p\text{-value} \leq 0.05/800$), out of all actual negatives (100 replicates per setting where $ACE = 0$). While MR-Egger shows zero FPR and little bias in the baseline setting, this changes to clear bias and a FPR of 0.6 and close to 1 in the presence of horizontal pleiotropy. Similarly, bias and FPR increase markedly for IVW in that setting. This suggests that the usage of IVs selected by CI-GWAS can help to avoid biases and false positives in MR inference that are introduced through either latent or horizontal pleiotropy.

We then compare the IVs detected by CI-GWAS to those identified by the standard approach from GWAS summary statistics (associated with X with $p\text{-value} \leq 5 \times 10^{-8}$ with $LD R^2 < 0.001$ within 1Mb). True valid IVs are those that are simulated to be associated with X , but not pleiotropic with Y or U . We find that CI-GWAS has a true positive rate (TPR, defined as the proportion of IVs correctly identified) comparable to GWAS for the selection of valid IVs under the horizontal pleiotropy settings considered here, but with a slight loss of power (Fig. 1d). Additionally, we calculate the proportion of selected IVs for X that are also causal for Y through a path that is not through X or pleiotropic with Y (labeled PHP: proportion of horizontally pleiotropic IVs in Fig. 1d). We find that CI-GWAS selected IVs are less likely to be causal for Y through a path that is not through X or pleiotropic with Y than IVs selected through GWAS, which explains why CI-GWAS improves causal effect estimates under both latent and horizontal pleiotropy.

Graph-based filtering of MR test results

As mentioned above, the graphical model learned by CI-GWAS indicates which potential causal relationships are compatible with the data's conditional independence structure. Consequently, one can use it to filter out hypothetical causal relationships that are not compatible with that conditional independence structure, with the aim of reducing the number of false positive results and exploring the nature of the trait relationships. Here, we explore two ways of doing that: (1) we inspect the skeleton and filter for pairs of variables that are adjacent. While non-adjacent pairs of variables may generally be in a causal relationship with each other, that relationship is mediated at least by one other observed variable, and the effect is therefore not direct. (2) We inspect the PAG and filter for pairs between which there is a directed or semi-directed edge, as for these, the conditional independence structure of the data does not provide any evidence against the existence of a causal effect (see Methods for more details).

We conduct an additional simulation experiment to investigate the effect of these filters. For that, we repeat the data generation for the four horizontal pleiotropy scenarios described above (high/low power and $ACE=0.0/ACE=0.2$ for $X1$ to Y), but modified the graph to include an additional exposure variable ($X2$), which does not have a

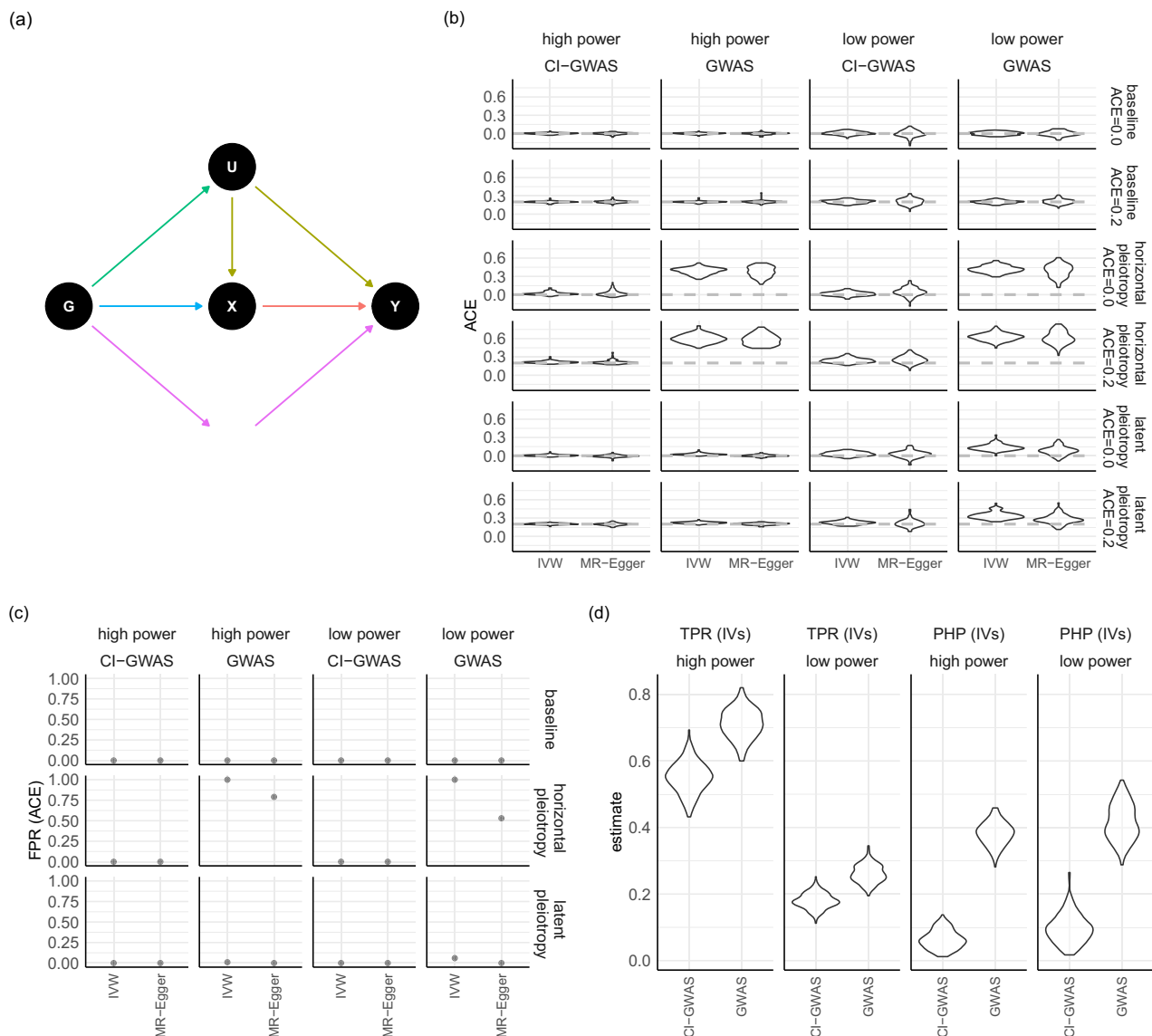


Fig. 1 | Causal inference performance of two Mendelian randomization (MR) approaches using CI-GWAS instrumental variable selection. a Simulation graph: unobserved confounding (U) affects both exposure (X) and outcome (Y) in all settings (light green); instrument G – X association strength is varied to create low- and high-power settings (blue); the average causal effect (ACE) of X on Y is set to 0.0 or 0.2 (red), giving a baseline scenario with four settings. Adding weak genetic overlap between U and X/Y (dark green) creates four latent pleiotropy settings; adding direct paths from G to Y (purple) creates four horizontal pleiotropy settings. Each of the twelve scenarios is replicated 100 times (1200 datasets total, $n = 458,747$). **b** ACE estimates from inverse-variance weighting (IVW) and MR-Egger,

using instruments selected either by standard GWAS criteria (independent SNPs associated with X at $p \leq 5 \times 10^{-8}$) or by CI-GWAS. **c** False positive rate (FPR) for each method across all twelve scenarios: the proportion of replicates where the true ACE = 0 but a causal effect is incorrectly detected ($p \leq 0.05/800$), out of 100 null replicates per setting. **d** True positive rate (TPR) – the proportion of instruments correctly identified by CI-GWAS versus standard GWAS – shown for the horizontal pleiotropy settings, alongside the proportion of selected instruments that are also horizontally pleiotropic (PHP), i.e., causal for Y through a path not passing through X .

causal effect on Y (ACE=0.0 for X_2 to Y , see Fig. 2a and Methods). Thus, here, true causal relationships exist only from X_1 onto Y , and horizontal pleiotropy occurs either through unmeasured variables or through X_2 . To perform the MR tests, we use IVW and multivariable inverse-variance weighting (MIVW). We then assess the performance with and without the above described filtering strategies, which yields three distinct scenarios: MR tests are run on all exposure-outcome pairs (labeled s1 in Fig. 2); MR tests are run only on exposure-outcome pairs that are adjacent in the skeleton (labeled s2); and lastly, MR tests are run only on exposure outcome pairs for which there is either a directed or semi-directed edge from exposure to outcome in the PAG (labeled

s3). For edges that were filtered out, we set the inferred ACE = 0 for the calculation of the performance metrics. We then calculate the FPR, TPR and true discovery rate (TDR) of the inferred causal relationships, where we calculate the rates as proportions of simulation replicates, since in each replicate, there is at most one causal link between an exposure and the outcome. For example, the FPR measures the proportion of simulation replicates where the true ACE = 0, but we incorrectly identify a causal effect of one of the exposures, X_1 or X_2 , on the outcome Y .

We find that the proposed filters are effective in reducing the FPR when using GWAS-selected IVs, but have only a little room left to

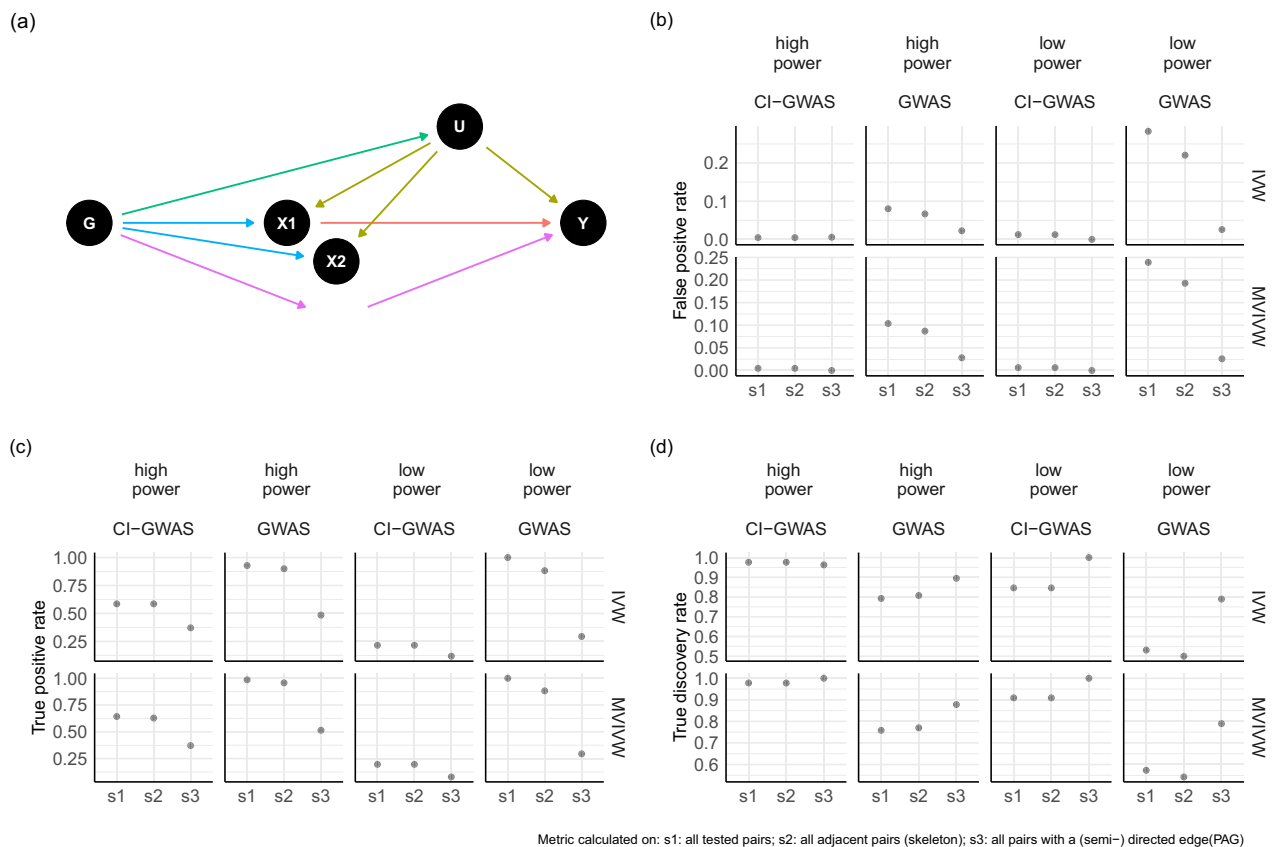


Fig. 2 | Causal inference performance of two Mendelian randomization (MR) methods using CI-GWAS instrumental variable selection, graphical adjacency, and partial ancestral graph (PAG) inference. **a** Simulation graph: unobserved confounding (U) affects both exposures ($X1$, $X2$) and outcome (Y) in all settings (light green); instrument G – $X1/X2$ association strength is varied to create low- and high-power settings (blue); the average causal effect (ACE) of $X1$ on Y is set to 0.0 or 0.2 (red); horizontal pleiotropy is introduced through unmeasured latent variables (purple); and U has a genetic basis weakly overlapping with that of $X1$ and $X2$ (dark green). Each of the four scenarios (low/high power $\times$ ACE = 0.0/0.2) is replicated 100 times (400 datasets total, $n = 458,747$). Instruments were selected either by standard GWAS criteria (independent SNPs associated with $X1$ and $X2$ at $p \leq 5 \times 10^{-8}$) or by CI-GWAS. **b** False positive rate (FPR) for inverse-variance weighting (IVW) and

multivariable IVW (MVIVW): the proportion of replicates where the true ACE = 0 but a significant causal effect of $X1$ or $X2$ on Y is incorrectly detected ($p \leq 0.05/800$), out of 600 null replicates total (400 where ACE = 0 for $X2 \rightarrow Y$, 200 where ACE = 0 for $X1 \rightarrow Y$). **c** True positive rate (TPR): the proportion of true causal relationships correctly identified. **d** True discovery rate (TDR): the proportion of identified causal relationships that are genuinely true. Performance is assessed with and without three filtering strategies: s1, MR tests run on all exposure-outcome pairs; s2, tests restricted to pairs adjacent in the learned skeleton; and s3, tests restricted to pairs with a directed or semi-directed edge from exposure to outcome in the PAG. For edges filtered out under s2 or s3, the inferred ACE is set to 0 when calculating performance metrics.

reduce the FPR when using CI-GWAS selected IVs (Fig. 2b), since that in itself brings the FPR down to almost 0, as already shown above (Fig. 1c). Analyzing the effect on the TPR (Fig. 2b), we see that the filters are not perfectly specific to false positives, but remove true positives as well, showing a classical recall-precision tradeoff, which is already evident when focusing only on the effect of GWAS vs CI-GWAS selected markers (s1). This is because, particularly in low-powered settings, the causal relationship between $X1$ and Y cannot always be directed due to simulated unobserved confounding and horizontal pleiotropy. Looking at the TDR, i.e., the proportion of identified causal effects that are actually true, we see that CI-GWAS-selected IVs yield extremely precise results across all settings with TDRs at or above 0.85. When looking only at relationships that can be directed, the TDR is further improved with the filters up to a TDR close to 1, even in the low power scenario. When using GWAS-selected IVs, we see that the PAG filter (s3) can bring a strong TDR improvement of about 0.2. The effects of the filters on the performance are consistent across IVW and MVIVW, with MVIVW generally yielding higher TDR at similar TPR, compared to IVW. Overall, we see that by trading recall for precision, CI-GWAS provides ways to refine causal inference in these simulation settings.

Causal inference in a set of correlated traits and in the presence of strong LD

In the settings above, we have explored CI-GWAS in a very classical setting where it is pre-determined which traits are potential exposures and which are outcomes. Here, we move to a more complex scenario where we have a set of traits of interests and make no such assumptions about the causal structure, but instead allow that every trait might be causal for any other trait. Starting there, our aim is to infer the structure of the causal graph that relates these traits.

To study how well such a task can be solved by using CI-GWAS in conjunction with MR, we create an expanded simulation study in which we simulate many correlated traits and multiple latent confounders. We use a structural equation model consisting of twelve continuous traits, two of which are latent confounders and are not used in the inference. Each trait is influenced by a set of 50 causal markers randomly selected from 654,173 imputed genetic markers recorded for 458,747 UK Biobank individuals. Here, we include two additional MR approaches in the comparison, namely CAUSE¹¹ and MV-PRESSO⁸, in addition to MVIVW. We compare their performance when used with GWAS IV selection (MR (IV_G)) as input, and in conjunction with CI-GWAS IVs (MR (IV_{CI})). We then repeat the analyses using the CI-GWAS

filters of the learned adjacencies (+ adj: using all adjacent pairs in the skeleton) or the PAG (using all pairs with evidence of a directed relationship). Since we assume that there are no pre-determined exposures and outcomes, we run CAUSE, which is a single-outcome, single-exposure model, for each simulated trait pair. For the multivariable approaches, we test each trait as an outcome, with all other traits treated as risk factors. When using the CI-GWAS inferred graph as a filter: we only conduct MR testing with CAUSE for trait pairs that have an adjacency, or a corresponding directed edge in the PAG; while for the multivariable approaches, we only select traits to include as exposures that have an adjacency, or a corresponding directed edge in the PAG. We simulate only a well-powered setting with a limited degree of horizontal pleiotropy as this allows us to compare approaches in an idealistic setting, where the results are not obscured by the lack of implicit control for horizontal pleiotropy of GWAS IV selection (see Methods). We run 20 replicates as this number was sufficient to support the conclusions of the simulations that we have already presented above.

We make the following observations: First, as shown above in Fig. 2, filtering standard MR results with the CI-GWAS-learned adjacencies increases the TDR for all three tested MR methods (MR (IV_C) vs. MR (IV_C) + adj in Fig. 3). Second, the TDR is further increased, at the expense of power (TPR), by using the PAG instead of the skeleton for filtering (MR (IV_C) + adj vs. MR (IV_C) + PAG in Fig. 3). Third, for methods such as MV-PRESSO and MVIVW, only by using the PAG for filtering is an FDR of 0.05 achieved at a p -value threshold of 0.05 for the testing of the average causal effect (Fig. 3b). Fourth, as CAUSE conducts variable selection on the marginal GWAS associations to select IVs, we find no evidence for improved performance when restricting CAUSE to using only CI-GWAS selected IVs. However, inference is more precise, with higher power when using the CI-GWAS-learned adjacencies. Finally, MVIVW with the PAG learned by CI-GWAS and CI-GWAS-selected IVs has a comparable FDR of 0.05 to CAUSE, but has significantly improved TPR. We find that these results remain largely consistent with different selection thresholds for GWAS (p -value $\leq 10^{-4}$), when more markers are selected (Supplementary Figs. 2, 3, 5) and when pleiotropic effects are removed from the simulation (Supplementary Figs. 3, 5). Thus, we show that learning a graphical model, as in CI-GWAS, can be used in conjunction with traditional MR methods to improve precision.

Application of CI-GWAS to the UK Biobank

We apply CI-GWAS to 15 traits and 8.4 million single-nucleotide polymorphism (SNP) markers in UK Biobank data, as described in the Methods and Supplementary Table 1. To compare our results to previous findings, we first analyze 36 pairs of risk factors and outcomes (single r.f.), focusing on four outcomes: asthma (AT), cardiovascular disease (CAD), stroke (ST), and type-2 diabetes (T2D). We then analyze all risk factors jointly with each outcome within an MVIVW framework (joint r.f.).

The number of SNP markers discovered in part 1 of our algorithm is given in Fig. 4. From the single r.f. Analyses, we find overall 1052 unique markers, which decline to 696 in the joint r.f. analyses that are suitable IVs. We find that the composition of the SNP sets varies between the single r.f. and the joint r.f. analyses (Supplementary Fig. 6), as expected due to the varying correlation between the traits considered. Interestingly, we find that the majority of the IVs identified are located in protein-coding genes (Table 1).

In the right panel of Fig. 4, we compare our single r.f. results to those obtained by CAUSE in the UK Biobank presented in ref. 11. IVW with CI-GWAS adjacencies and CI-GWAS identified IVs (CI-GWAS single r.f.) recover all of the suggested statistically causal paths found by CAUSE, with the exception of height (HT) to CAD and diastolic blood pressure (DBP) to ST. In addition to those discovered by CAUSE, we find paths from blood glucose levels (GLU) to both CAD and ST, from

triglycerides (TRIG) to CAD and T2D, from low-density lipoproteins (LDL) to stroke (ST) and waist-hip ratio (WHR) to CAD (Fig. 4c). This supports our simulation results, which showed that our algorithm performs similarly to CAUSE, with moderately improved power.

We then move from pair-wise single r.f. testing to using MVIVW with CI-GWAS adjacencies and CI-GWAS identified IVs (CI-GWAS joint r.f.), again focusing on the four outcomes of AT, CAD, ST and T2D as the outcome variables in MVIVW. Here, CI-GWAS joint r.f. finds GLU and WHR as potential causal risk factors for both CAD and T2D. HDL and LDL are also potential causal risk factors for CAD, and we find that LDL remains a potential causal risk factor for ST.

We then filter the exposure-outcome relationships more conservatively using the inferred directed edges in the PAG. We find that only the relationship between GLU and CAD persists (Supplementary Fig. 7). In addition, we find a directed edge from HDL onto AT, for which there is support in neither the previous CAUSE nor in our MVIVW results (Supplementary Fig. 7), but support in the literature²⁸. All other previous potential causal relationships that are significant in CAUSE, the CI-GWAS single r.f. and the CI-GWAS joint r.f. analyses, are assigned a bi-directed edge in the inferred PAG returned by part 2 of our algorithm. Here, a bi-directed edge is defined as two observed variables that directly influence each other simultaneously (a feedback loop or mutual causation occurring too quickly to observe directionality in the data), or are governed by unobserved selection variables, reflecting the presence of a latent confounder. Thus, within the UK Biobank, our model suggests that all of the other potential causal relationships inferred (e.g., by CAUSE presented in ref. 11) may be explained by feedback loops, mutual causation, or unmeasured latent confounding. Latent confounding may reflect processes like population structure, study ascertainment, linkage, shared parental effects, shared environment and/or assortative mating.

Discussion

Here, we present a method for statistical causal inference in a genomic setting to learn a graph by inferring the adjacencies from partial correlations, orienting the edges and then estimating the strength of the causal effects, whilst allowing for the presence of multiple confounders and multiple correlated phenotypes. The algorithms we present have strong theoretical support and have been heavily studied in the causal inference literature^{20,21,29}. We demonstrate that extending statistical causal discovery from MR methods to multivariable graphical inference for large GWAS or biobank data sets with millions of variables and hundreds of thousands of records provides: (1) appropriate selection of IVs that are likely to fulfill MR assumptions, and (2) robust inference of potential causal links with greatly reduced false discovery rates and fast run time for both continuous and binary measurements. Current MR methods generally give consistent results, they all build upon each other, assume a similar fixed directed acyclic graph and can have high type-I error. Here, we show empirical evidence for this and how this can be addressed by combining existing MR approaches with graphical inference.

High-dimensional non-sparse graphical causal inference is an active area of research, and there have been a number of proposed alternatives to the approach presented here, such as various alternative forms of multivariable MR^{8,30,31}. Here, we show that MR is ineffective without being paired with graphical inference. Structural equation models have also been suggested³², but these require a correct causal order of variables to be known (or learned) and are cumbersome to run on large-scale high-dimensional data, even in summary statistic form. A number of artificial intelligence approaches have recently been presented for causal inference in life science settings³³. However, these lack rigorous theory and an understanding of the conditions required for these algorithms to recover meaningful causal structures and estimates. As a result, while estimates are obtained, one has little assurance that reliable inference is obtained, much like

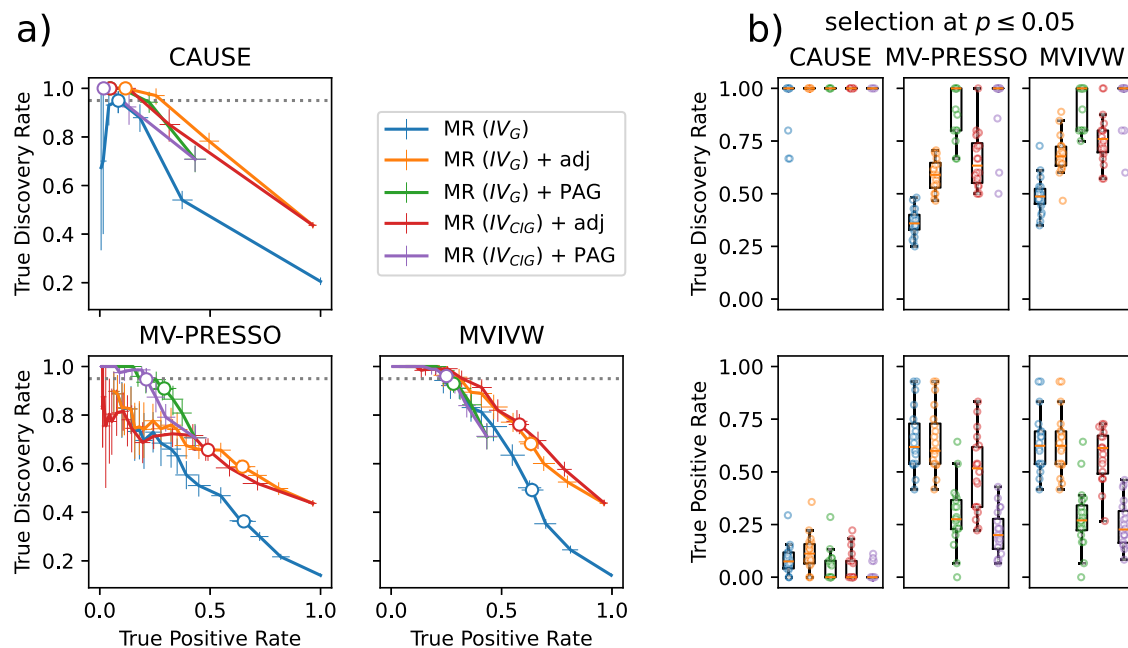


Fig. 3 | Causal inference performance of CI-GWAS with commonly used MR methods in simulation with strong LD and ten strongly correlated traits. **a** True discovery rate as function of true positive rate of the trait-trait causal link detection for a number of commonly used MR methods (CAUSE, MV-PRESSO, MVIVW) when used with GWAS selected instrumental variables (IV_G), CI-GWAS selected instrumental variables (IV_{CIG}), and with and without using the CI-GWAS skeleton (adj) or PAG as an additional filter for causal relations. Lines show bootstrapped means, error bars indicate the bootstrapped 95% confidence interval of the mean. We

generate and analyze 20 replicates per scenario, each with sample size 458,747. The dotted line indicates $TDR = 0.95$. Empty circles indicate the mean at a p -value threshold of 0.05 and in **b** values from **(a)** are directly compared when selecting trait-trait causal links at p -value ≤ 0.05 . Orange line shows median (interpolated if between two values). Box goes from first quartile to the third quartile of the data. Whiskers go from box edges to the farthest point that is within $1.5 \times IQR$ (inter-quartile-range) from the box. Outliers are not shown.

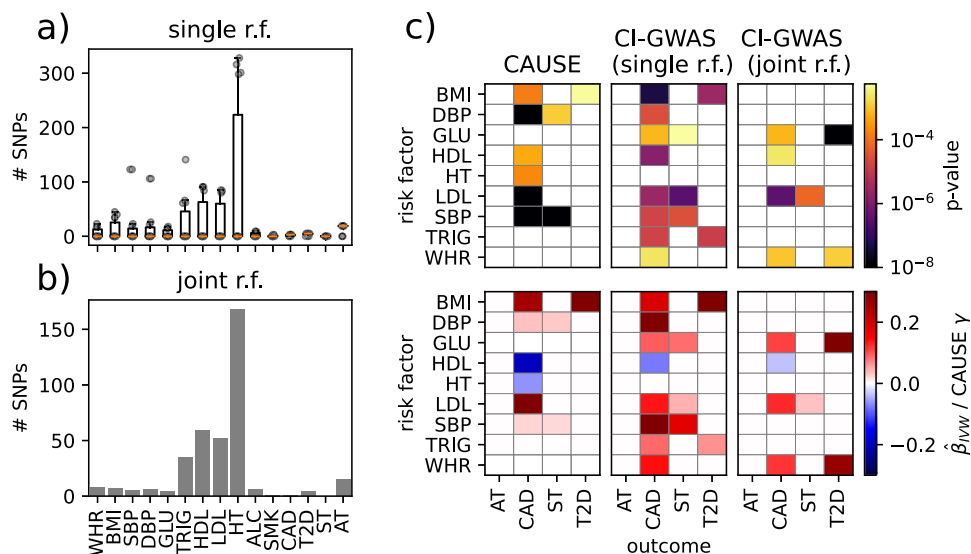


Fig. 4 | Causal inference in the UK Biobank with CI-GWAS. We analyze 36 risk factor (r.f.) - outcome relationships in the UK Biobank. We run either pairs of a single risk factor and outcome (single r.f.), or all risk factors with a single outcome jointly (joint r.f.). **a** Numbers of identified causal SNPs per trait, where orange line shows median (interpolated if between two values). Box goes from first quartile to the third quartile of the data. Whiskers go from box edges to the farthest point that is within $1.5 \times IQR$ (inter-quartile-range) from the box. Outliers are not shown.

b Numbers of identified causal SNPs per trait in the joint r.f. analyses. **c** p -values (top row) and causal effect sizes (bottom row) for the analyzed risk factor to outcome relations. CAUSE results are taken from ref. 11. We discard results with p -values above 0.005 (0.05 adjusted for the number of tests) as insignificant. The corresponding squares are left white. We use the MVIVW (IV_{CIG}) + adj approach, selecting only links as causal if the MVIVW result is significant and the traits are adjacent in the inferred skeleton.

Table 1 | Number of causal SNPs identified by CI-GWAS in the UK Biobank classified as biotypes of various kinds

Biotype	Count
None	343
Protein-coding gene (exon)	175
Protein-coding gene (intron)	434
lincRNA (intron)	53
lincRNA (exon)	4
Antisense of protein-coding gene intron	36
Antisense of protein-coding gene exon	4
snoRNA	1
Sense overlapping	1
miRNA	1

standard MR approaches. Thus, we focus here on applying and developing a principled, theoretically supported method that provides a crucial step toward high-dimensional statistical causal learning for genomics.

A number of limitations remain. Our approach requires either individual-level data or summary information on the minor allele frequency, the LD correlations, and the phenotypic correlations, alongside standard GWAS regression coefficient estimates. This level of summary data is not always publicly available, but could easily be shared, inferred across biobanks, or approximated. Data sets have a structure that can inhibit unbiased causal inference, such as selection biases like the ascertainment of healthy older people into biobank studies, which can bias causal inference³⁴. Additionally, there is also a general tendency in the UK Biobank sampled individuals for almost everyone to be under some form of medical intervention. In particular, natural variation in lipid levels and blood pressure may be impossible to observe within the UK Biobank as the majority of people are being treated with some form of drug. Here, we adjust for the effects of drug taking prior to analysis, but if the majority have interventions, this is likely ineffective, and a larger-scale model investigating prescriptions, observations and outcomes through repeated measures may be more effective. By utilizing longitudinal records, where discovered adjacencies in the skeleton can then be orientated longitudinally and combined with therapeutic information, statistical causal effects may be more effectively examined as there is certainty about the directions of many time ordered edges. This will then reduce the number of bi-directed edges learned in the PAG. Our approach easily facilitates this and we are actively working on this topic. Finally, genotype-phenotype associations likely reflect a complex combination of direct, parental genetic and epigenetic effects, and this may also be why we infer a large number of relationships to occur due to latent confounding. Parental genotypes and their relationship to offspring traits could also be accommodated within our framework, but at this time there is a lack of large-scale family data.

In summary, our approach is applicable to screen for potential causal relationships within life-science data, from genomics to the microbiome to single-cell sequencing and phenotyping datasets. This may be very fruitful as trait-omics associations are often stronger than marker-trait correlations, giving improved recall rates. By combining methods of genetic epidemiology and statistical causal inference, within algorithms that require only summary statistic correlations from the data, we expect that our work will facilitate extensive investigation of potential causal hypotheses. We show that inferring causal relationships from simple observational biobank data is limited, and our results call for caution when applying current MR methods to biobank data. With any dataset, we feel that it is important to understand which relationships hold under which assumptions, and our approach provides a way to screen associations under explicit checks.

This way, more can be learned about the likely underlying nature of phenotypic and genetic correlations within the human population.

Methods

Ethical approval declaration

This project uses UK Biobank data under project 35520. UK Biobank genotypic and phenotypic data is available through a formal request at (<http://www.ukbiobank.ac.uk>). The UK Biobank has ethics approval from the North West Multi-center Research Ethics Committee (MREC). Methods were carried out in accordance with the relevant guidelines and regulations, with informed consent obtained from all participants.

CI-GWAS

CI-GWAS takes either individual or summary-level data, learns conditional dependencies between genetic variants and traits and determines whether conditional dependencies among traits can be attributed to a directed causal relationship. It can then estimate, in conjunction with a Mendelian Randomization method, the magnitude and direction of the potential causal effect. A full analysis with CI-GWAS consists of three computational steps: (1) a GPU-accelerated algorithm that conducts genetic variable selection and returns an undirected graph that describes the potential causal relationships among markers and traits (called cuda-skeleton); (2) an approach called sRFCI that directs the skeleton edges and thereby infers causal relationships, allowing for arbitrarily many latent variables; and (3) estimation of the magnitude of the potential causal relationships among the discovered exposures and outcomes with a MR method like multi-variable inverse weighted Mendelian Randomization (MVIVW)⁹. This final part uses the sets of directly associated markers for each trait as instrument variables, which conform to the assumptions of MVIVW. Additionally, we propose only estimating exposure-outcome effects for trait-trait conditional dependencies discovered by either cuda-skeleton or sRFCI that are also significant in MVIVW. We provide detailed information on each step of CI-GWAS in the following sections, along with algorithms for their implementation.

CI-GWAS operates on correlation matrices M^C , M^{GT} , M^T representing correlations between markers, between markers and traits, and between traits, respectively. As described below, we run the first step of the CI-GWAS (cuda-skeleton) in parallel over LD-blocks, each of which has a distinct M^C matrix. Thus, we first block the genomic input data with the parameter set as shown in Table 2. Then we prepare the correlation matrices as follows:

- M^C : for each input block B_i we obtain the LD-matrix using PLINK 1.9 with the options `-r triangle bin 4` and `-allow-no-sex`, selecting the rSDs of the markers in the block with the `-from` and `-to` options³⁵.
- M^{GT} : we compute the marker-trait correlations using linear regression models. We first obtain linear regression coefficients β_{ij} for each marker i and trait j by running PLINK 2.0 on the bed file with the complete marker data and the individual-level trait-data with `-glm omit-ref -variance-standardize` options³⁶. We then compute the Pearson correlation via $\rho_{ij} = \beta_{ij} \sqrt{2f_i(1-f_i)}$, where f_i is the minor allele frequency of marker i .
- M^T : Here, we simply compute the Pearson correlations between the individual-level data trait columns.

Note that any choice of approximations of (i) to (iii) is accepted within our algorithm. For example, recent proposals for summary statistic prediction and heritability models are to whiten, shrink, or block the LD matrix, and to use approximations from other publicly available samples, if the original sample matrices are unavailable³⁷ can all be accommodated.

In order to accommodate binary traits in our framework, we calculate all correlations with the binary traits placed on the liability scale, as suggested by ref. 25. For this, we use polyserial correlations³⁸

Table 2 | Algorithm parameters and their values in all applications on real data

Algorithm	Parameter	Value
cuda-skeleton	$l_{\max}^{(1)}$	3
cuda-skeleton	$l_{\max}^{(2)}$	14
cuda-skeleton	α	10^{-3}
blocking	$s_{\max}$	10^4
blocking	ϵ	10^2
blocking	corr-width	2×10^3

between continuous variables and binary variables, tetrachoric correlations³⁹ between binary variables, and Pearson correlations between continuous variables.

We provide open source code and scripts to run our full analysis (see Code Availability). We now describe each of the component parts of our approach in full below.

Part 1: variable selection and inference of adjacencies

Algorithm 1. Genome blocking procedure for summary statistics cuda-skeleton

Require: LD matrix C for n markers; max block size $s_{\max}$; tolerance ϵ ; bandwidth b

- 1: $C^b \leftarrow$ make banded LD matrix from C given b
- 2: $S_i \leftarrow \sum_{j=i+1}^n C_{ij}^b$
- 3: $h \leftarrow n$
- 4: $l \leftarrow 3$
- 5: $w \leftarrow (h + l)/2$
- 6: $s \leftarrow n$
- 7: **while** $(|s - s_{\max}| > \epsilon) \vee (s > s_{\max})$ **do**
- 8: **if** $s > s_{\max}$ **then**
- 9: $h \leftarrow \min(h, w)$
- 10: **else**
- 11: $l \leftarrow \max(l, w)$
- 12: **end if**
- 13: $w' \leftarrow (h + l)/2$
- 14: **if** $w' = w$ **then break**
- 15: **end if**
- 16: $w \leftarrow w'$
- 17: $S' \leftarrow$ Hanning-smoothen S with window w
- 18: $M \leftarrow$ positions of local minima in S'
- 19: $s \leftarrow \max\{|M_{i+1} - M_i| \mid 1 \leq i \leq (n - 1)\}$
- 20: **end while**
- 21: **return** M

The first step of our algorithm: (i) selects markers that have a statistically significant effect on one or more of the traits; and (ii) builds the *skeleton*, the undirected graph describing all possible direct relationships between all variables (i.e., all markers and traits). This step is based on the skeleton inference procedure of the recently proposed cuPC algorithm²⁴, a GPU-accelerated implementation of the PC-stable algorithm²¹.

Given all pairwise correlations between variables, the first part of the PC-stable algorithm infers the skeleton by testing pairs of variables V_i and V_j for independence conditional on an exponential number of *separation sets*, S , consisting of other variables in the graph. In particular, $V_i \perp\!\!\!\perp V_j | S$ is tested by first estimating the partial correlation of V_i and V_j given a separation set S consisting of neighbors of V_i . We denote this by $\hat{\rho}(V_i, V_j | S)$. The Fisher's z-transform of this value, i.e.

$$Z(\hat{\rho}(V_i, V_j | S)) = \left| \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}(V_i, V_j | S)}{1 - \hat{\rho}(V_i, V_j | S)} \right) \right| \quad (1)$$

is then compared to a threshold τ , given by

$$\tau = \frac{\Phi^{-1}(1 - \frac{\alpha}{2})}{\sqrt{m - |S| - 3}} \quad (2)$$

where m is the sample size, α is the significance level of the test and Φ is the CDF of the standard normal distribution. $Z(\hat{\rho}(V_i, V_j | S)) \leq \tau$ is taken to imply $V_i \perp\!\!\!\perp V_j | S$. These tests are run for consecutively increasing level $l = |S|$, starting with $l = 0$. A $l_{\max}$ can be set to limit the run time of the algorithm. Edges between independent variables are removed. The algorithm terminates when $(l > n) \vee (l > l_{\max})$, where n is the number of variables. The full parameter setting used in this work is given in Table 2.

The run time scales exponentially with the number of variables, since the number of possible S of a given size l increases exponentially. cuPC performs the conditional independence tests in a highly parallelized fashion, utilizing modern GPU capabilities, enabling skeleton inference for many variables in a short time. In the genomic setting, with order 10^9 markers and order 10^2 traits, both storage of the input correlation matrix as well as computation time required prohibit the application of cuPC on all variables at once. Therefore, we block the marker LD matrix in a fashion inspired by ref. 40 and as shown in Algorithm 1. When relaxing the α threshold, there may be circumstances where the memory requirement at larger l exceeds that on standard GPU architecture. If this issue arises, we propose to simply run the algorithm twice, first with a low $l = 3$, to select parental markers, and then run all levels from $l = 0$ to $l = 14$ using only the parental markers and all traits. We apply the skeleton inference component of the cuPC algorithm to each block of markers and to all traits individually. We performed a sensitivity analysis for this choice in our empirical UK Biobank application and concluded that this strategy results in approximately the same marker selection that one would obtain from running on full chromosomes, as can be seen in Supplementary Fig. 8.

Having inferred a set of skeletons with adjacency matrices $A^B = A_1, A_2, \dots, A_n$ and separation sets $S^B = S_1, S_2, \dots, S_n$ blockwise, we first create a merged skeleton with adjacency matrix A' following four rules: (1) a pair of traits are adjacent in A' if and only if they are adjacent in all $A' \in A^B$, since the presence of a single separation is sufficient to imply conditional independence; (2) if two markers are in different blocks, they are not adjacent in A' ; (3) if two markers are in the same block i , they are adjacent in A' if and only if they are adjacent in A_i ; and (4) if a marker and a trait are adjacent in any block, they are adjacent in A' . This gives a merged skeleton, representing a reduced set of markers that have a direct edge to one or more traits in the graph. This reduces the set of variables to the ones that are most likely to have direct causal effects on traits. To prevent errors arising from correlated markers being placed in separate blocks, we repeat the CUDA-skeleton step on a single artificial block consisting only of the markers selected in the first run. This ensures that each marker-trait association is always tested conditional on every other potentially correlated marker.

Part 2: orientating edges

Having inferred a skeleton, we propose to infer the direction of causal relationships by orientating the graph edges. For this, we employ two procedures: (1) inference of a partial ancestral graph (PAG) by a modified RFCl²⁰ algorithm (sRFCl); and (2) multi-variable inverse variance weighted Mendelian randomization (MVIW)⁹, or a comparable method, as described in the following section.

Inferring a PAG. Unobserved confounding variables (i.e., latent variables) are a common problem for the inference of causal relationships. They indicate violation of causal sufficiency (i.e., knowledge of all causally important variables in the graph), which is a common assumption in causal inference. In consequence, a variable may be

inferred as causal for another even though there is no causal relationship between the two. To allow for an additional check for unobserved variables, we devise an approach based on the RFCI algorithm²⁰ for the inference of causal graphs, which allows for arbitrarily many latent variables in the true underlying directed acyclic graph (DAG). True separation sets, i.e., the full set of variables that render two non-adjacent variables independent when conditioned upon, are difficult to obtain when marker-trait effects are weak. We observed that weak marker-trait effects result in incomplete separation sets, causing traits to be missing from the separation sets of markers and other traits, thereby giving incorrect edge orientation. To solve this issue, we present an order-independent algorithm that uses the RFCI procedure to orient edges between adjacent variables, whilst considering a locally maximal number of variables in the separation sets used to infer the orientations. Specifically, we infer two types of separation sets: (1) S^{max} , a maximal separation set, meaning that the addition of any other variable results in $Z > \tau$, and (2) S^{mpc} a subset of the maximal separation set that locally minimizes the partial correlation between variables V_i and V_j .

S^{max} is conservative as to what is not included in the separation set. Thus, everything that is not within S^{max} can be considered a collider between the variables. This separation set is used within sRFCI to orient unshielded triples, as described below. A triple of nodes V_i , V_k and V_j is unshielded if all of the following are true: (1) V_i and V_k are connected; (2) V_k and V_j are connected; (3) V_i and V_j are not connected. We only want to orientate unshielded triples, when V_k is not an element of S^{max}_{ij} , and thus is an unlikely collider. The set S^{mpc} is more conservative with regard to the inclusion of variables into the separation set, and we use this in the later orientation steps of sRFCI, as described below.

Algorithm 2 describes the algorithm we created to obtain S^{max} and S^{mpc} using the skeleton obtained from cuda-skeleton and a matrix containing the marker-marker correlations, marginal marker-phenotype correlations and phenotype-phenotype correlations. We made S order independent and attempted to overcome missing non-collider variables by maximizing separation sets as much as possible. We note that the CPC/MPC versions of RFCI²¹ provide a solution to the same issue by performing a maximal search over all potential separation sets. Unfortunately, this solution is infeasible in a high-dimensional genomic setting as it requires checking 2^n sets where n is the degree of V_i . Therefore, we propose the heuristic Algorithm 2, where we do not exhaustively test all possible separation sets. Rather, we incrementally increase their size, always adding the element that yields the lowest partial correlation among the tested elements at each step. Thereby, we reduce the search space to n^2 sets. Furthermore, we reduce the set of unshielded triples used for orientation to relevant unshielded triples, which we define as those that contain at most one marker and in which the middle variable is not a marker. The separation sets for all other triples are not needed since we assume that markers temporally precede traits and therefore force marker-trait edges to be directed towards traits. This reduces the number of triples for which separation sets need to be selected greatly, especially when the number of identified markers is large. One limitation of our modifications is that S^{max} may not be the largest separation set. Likewise, S^{mpc} may not be the separation set that yields the partial correlation among all possible sets that is closest to zero. However, we found that selecting these separation sets improves the orientation within sRFCI as compared to using the minimal subset S^* .

Algorithm 2. Separation set selection

Require: α, n, A^*, C

- 1: $ut \leftarrow$ relevant unshielded triples of A^*
- 2: $op \leftarrow$ outer variables in ut
- 3: **for** V_i, V_j in op **do**
- 4: $N \leftarrow \text{neighbors}(V_i, A^*)$
- 5: $N' \leftarrow N$

- 6: $sep \leftarrow \{\}$
- 7: $foundSepset \leftarrow \text{false}$
- 8: $foundMinimum \leftarrow \text{false}$
- 9: **for** l in $1 \dots |N|$ **do**
- 10: $newElem \leftarrow \{\}$
- 11: $z_{ref} \leftarrow \infty$
- 12: **for** V_k in N' **do**
- 13: $S \leftarrow sep \cup V_k$
- 14: $z \leftarrow Z(\hat{\rho}(V_i, V_j | S))$
- 15: **if** $z \leq z_{ref}$ **then**
- 16: $z_{ref} \leftarrow z$
- 17: $newElem \leftarrow V_k$
- 18: **end if**
- 19: **end for**
- 20: **if** $z_{ref} > z_{ref, l-1} \vee foundSepset \vee \neg foundMinimum$ **then**
- 21: $foundMinimum \leftarrow \text{true}$
- 22: $S^{mpc}_{i,j} \leftarrow sep$
- 23: **end if**
- 24: **if** $z_{ref} > \tau(\alpha, n, l) \wedge foundSepset$ **then**
- 25: **break**
- 26: **end if**
- 27: **if** $z_{ref} \leq \tau(\alpha, n, l)$ **then**
- 28: $foundSepset \leftarrow \text{true}$
- 29: **end if**
- 30: $sep \leftarrow sep \cup newElem$
- 31: $N' \leftarrow N' \setminus \{newElem\}$
- 32: **end for**
- 33: $S^{max}_{i,j} \leftarrow sep$
- 34: **end for**
- 35: **return** M

We now pass these separation sets to parts of the RFCI algorithms as described below:

- Step 1. Create a list of unshielded triples using Algorithm 4.4 of ref. 20.
- Step 2. Orientate the v-structures using S^{max} and Algorithm 4.4 of ref. 20.
- Step 3. Make a list of ambiguous unshielded triples V_i, V_k, V_j , where the V_k is in $S^{max}_{i,j}$, but not in $S^{mpc}_{i,j}$. These ambiguous triples do not contribute to the orientation in later stages, although they themselves can later be oriented if there is evidence for this.
- Step 4. Further edges are oriented as much as possible following the guidelines provided in Algorithm 4.5 of ref. 20. This step involves repeated applications of orientation rules (R1) to (R10) of ref. 41. We implemented the suggested modifications from ref. 21 to rules R1-R10 that make them order-independent. In this step, we provide the option to limit the application to only the phenotypic edges in order to reduce the runtime.
- Step 5. All marker-phenotype edges are oriented to ensure that phenotypes are not ancestral to markers. This conforms to the assumption that marker states temporally precede phenotypes.

The output of this algorithm is a Partial Ancestral Graph (PAG) representing a set of possible Maximal Ancestral Graphs (MAG, see ref. 20 for definition). Adjacencies are inferred to be composed of arrows ($<$ or $>$), circles ($\circ$), and tails ($-$). An arrow at a variable implies that the variable is not ancestral to the variable at the other end of the edge. The tail implies that the variable is ancestral to the variable at the other end of the edge. Circles represent arrowheads in some MAGs in the Markov equivalence class represented by the PAG and tails in others. RFCI allows inference of causal information in the presence of latent variables and hidden factors or confounders²⁰, which has been shown to be of importance for Mendelian Randomization methods applied to biobank data^{6,11,42–44}. A situation can arise when neither variable is ancestral to the other, which results in the adjacency being inferred as $\leftrightarrow$. This implies that neither variable is ancestral to the

other, with no direct causal relationship inferred, suggesting the existence of a latent confounder. Numerically, the algorithm outputs: 0 = no edge; 1 = circle; 2 = tail; 3 = arrow. Throughout this work, we use the following interpretation, circle-circle: undirected edge, uncertain orientation; circle-arrow: partial edge, potential causal link; tail-arrow: directed edge, unambiguous potential causal link; arrow-arrow: bidirected edge, indicating hidden latent confounding; circle-tail: semi-directed edge, potential causal link; tail-tail: undirected edge, insufficient data available. To select a set of relationships with clear causal direction from the output of this procedure, we select all inferred edges with one arrow and one tail, since these are invariant across the MAGs represented by the PAG and therefore are inferred to have a definite orientation. We note that a PAG represents a set of graphs that all share certain structures (i.e., node adjacency causal relations), and it is possible that the information on conditional independence in the data is insufficient to uniquely determine the exact type of an edge. Thus, it is also possible to relax the selection to relationships that include both directed and semi-directed edges, as for these, the conditional independence structure of the data does not provide any evidence against the existence of such an effect.

Part 3: learning the magnitude of the causal effects

Having learned the conditional dependencies in the data with cuda-skeleton and tested for latent confounding among traits with sRFCI, we propose using the MVIVW framework, or a comparable method, to assess the significance and magnitude of causal effects.

To identify potential causal relationships in an MR framework, such as MVIVW, sets of IVs and exposures are needed for each outcome. Specifically, MVIVW assumes that IVs associated with at least one exposure are not associated with any latent confounders of the exposures and outcome, and are conditionally independent of the outcome given the exposures, i.e., that all causal paths from the IV to the outcome go through the exposures, and that there is no latent confounder of IV and outcome^{9,45}. MVIVW can be employed using the SNPs inferred from cuda-skeleton because the conditional-independence tests conducted in cuda-skeleton explicitly check for several common MR assumptions on IVs (see ref. 45 for a discussion of those assumptions). In consequence, valid IVs can be directly read off the skeleton. In particular, for a given exposure trait *A* and an outcome trait *B*, any marker *X* adjacent to *A* that is not adjacent to *B* is a valid instrument to test the effect of *A* on *B* (assuming the skeleton is faithful to the underlying distribution) for the following reasons:

1. If *X* had a direct effect on *B*, the dependency should not have been removed by any conducted test. Thus, the exclusion restriction holds.
2. If marker *X* and the outcome *B* relationship was confounded by a latent variable, that confounding would induce a dependency, and again, *X* and *B* should remain adjacent in the skeleton, since no conducted test should be able to remove the dependency. Thus, the effect of *X* on *B* should not be confounded.

Note that the above only holds as long as all traits in the model, except for the outcome, are used as exposures in a multivariable MR analysis. This is necessary because *X* could also be associated with *B* through the omitted trait, which could either have in turn an effect on *B* and thereby invalidate the exclusion restriction. Thus, for single-variable MR like IVW, IV candidates need to be further verified by checking that the inferred graphical model does not contain a path from the IV to the outcome that does not include the exposure.

Principled instrumental selection of this kind contrasts with most other current MR approaches, in which IVs are selected by selecting variants with the highest marginal association test statistics. A similar approach, based on the association of markers with a single trait, while having multiple traits in the model, has been proven to be valid and successfully applied in previous work¹⁷. Here, we advance this idea

within a framework in which no markers need to be pre-selected, many variables can be conditioned on simultaneously, and pleiotropic markers can be used for inference, thereby promising more accurate IV selection and powerful analysis.

For each outcome, we assess the significance and estimate the effect size of each potential exposure using MVIVW with unfiltered parental markers (conforming to the criteria above) as instrumental variables. We use all traits in the model, aside from the outcome, as exposures. From here, we select only the links that are selected as significant in cuda-skeleton or are directed in the same way as inferred by MVIVW in the sRFCI inferred PAG. This provides sets of potential causal links with evidence of directionality learned from the data.

We use the selected instrumental variable and exposure sets as input to the MVIVW implementation in the *MendelianRandomization* R package⁴⁶, with the `robust` option.

Application of CI-GWAS to UK Biobank data

UK Biobank has approval from the North-West Multicenter Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants (<https://www.ukbiobank.ac.uk/ethics/>). These ethical regulations cover the work in this study. Written informed consent was obtained from all participants. From the measurements, tests, and electronic health record data available in the UK Biobank data⁴⁷, we selected six blood-based biomarkers, four of the most common heritable complex diseases, and seven quantitative measures. The full list of the 15 traits, the UK Biobank coding of the data used, and the covariates adjusted for are given in Supplementary Table 1. For the quantitative measures and blood-based biomarkers, we adjusted the values by the covariates, removed any individuals with a phenotype greater or less than 7 SD from the mean (assuming these are measurement errors), and standardized the values to have a mean of 0 and a variance of 1.

For the common complex diseases, we determined disease status using a combination of available information. For asthma (AT), we used self-reported asthma diagnosis by a doctor (3786-0.0), date of asthma report (42014-0.0), and self-reported recent asthma medication (22167-0.0). For cardiovascular disease (CAD), we used self-report information of whether a heart attack was diagnosed by a doctor (3894-0.0), the age angina was diagnosed (3627-0.0), whether the individual reported a heart problem diagnosed by a doctor (6150-0.0), and the date of myocardial infarction (42,000-0.0). For stroke (ST), we used self-report stroke diagnosed by a doctor (6150-0.0), the age at which stroke was diagnosed (4056-0.0). For type-2 diabetes (T2D), we used self-report information of whether diabetes was diagnosed by a doctor (2443-0.0), the age diabetes was diagnosed (2976-0.0), and whether the individual reported taking diabetes medication (6153-0.0, 6177-0.0). For each disease, we then combined this with primary death ICD10 codes (40001-0.0), causes of operative procedures (41201-0.0), and the main (41202-0.0), secondary (41204-0.0) and inpatient ICD10 codes (41270-0.0). For AT, we selected all ICD10 J45 codes, for CAD we selected ICD10 codes I20-I29, for ST we selected all ICD10 I6 codes, and for T2D we selected ICD10 codes E11 to E14 and excluded from the analysis individuals with E10 (type-1 diabetes). Thus, for the purposes of this analysis, we define these diseases broadly simply to maximize the number of cases available for analysis. Individuals with neither a self-report indication nor a relevant ICD10 diagnosis were assigned a zero value as a control.

We first restricted our analysis to a sample of UK Biobank individuals of European ancestry. To infer ancestry, we used both self-reported ethnic background (UK Biobank field 21000-0), selecting coding 1, and genetic ethnicity (UK Biobank field 22006-0), selecting coding 1. We projected the 488,377 genotyped participants onto the first two genotypic principal components (PC) calculated from 2504 individuals of the 1000 Genomes project. Using the obtained PC loadings, we then assigned each participant to the closest 1000

Genomes project population, selecting individuals with PC1 projection $\leq$ absolute value 4 and PC2 projection $\leq$ absolute value 3. We applied this ancestry restriction to provide the first application of our approach and to replicate our results in a sample as genetically homogeneous as possible. Note, however, that our approach can be applied within different human groups (by age, genetic sex, ethnicity, etc.). Our future work will focus on exploring differences in causal inference across a diverse range of human populations.

Samples were also excluded based on UK Biobank quality control procedures with individuals removed of (1) extreme heterozygosity and missing genotype outliers; (2) a genetically inferred gender that did not match the self-reported gender; (3) putative sex chromosome aneuploidy; (4) exclusion from kinship inference; (5) withdrawn consent. We used genotype probabilities from version 3 of the imputed autosomal genotype data provided by the UK Biobank to hard-call the single-nucleotide polymorphism (SNP) genotypes for variants with an imputation quality score above 0.3. The hard-call threshold was 0.1, setting genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p -value for Hardy-Weinberg test larger than 10^{-6} , as determined in the set of unrelated Europeans), we selected those with minor allele frequency (MAF) ≥ 0.0002 and rs identifier, in the set of European-ancestry participants. This resulted in a dataset with 458,747 individuals and 8,430,446 markers. We apply CI-GWAS to these data as described above and we use these data for our simulation study as described below.

Simulation study

We provide code to simulate data of the kind presented in either of the simulation scenarios (see Code Availability).

Baseline scenario. Using the observed imputed SNP data from 458,747 UK Biobank individuals on chromosome 1, we first simulate a graph standard to MR as follows:

$$\begin{aligned} X &= W_{s_X} \beta_X + \lambda_{U_X} U + \epsilon_X \\ Y &= W_{s_Y} \beta_Y + \lambda_X X + \lambda_{U_Y} U + \epsilon_Y \end{aligned} \quad (3)$$

where U is a latent unobserved random confounding variable with drawn from a Gaussian distribution with mean zero and variance of one. U has a causal effect on both X and Y , with $\lambda_{U_X} = \lambda_{U_Y} = 0.2$. We then simulate four different baseline settings, where the causal effect of X on Y (the average causal effect, termed ACE) is either $\lambda_X = 0$, or $\lambda_X = 0.2$ and the power to detect the genotype effects on X and Y are either high or low. To simulate high power, $\beta_X \sim N(0, h_X^2/m_c)$, $\beta_Y \sim N(0, h_Y^2/m_c)$, with $h_X^2 = h_Y^2 = 0.05$, $m_c = 100$, $\epsilon_X \sim N(0, \sigma_{\epsilon_X})$, $\epsilon_Y \sim N(0, \sigma_{\epsilon_Y})$, $\sigma_{\epsilon_X} = \sigma_{\epsilon_Y} = 0.95$. Note here that SNPs are randomly selected for X and Y independently. The expected variance contributed by a single SNP variant is $0.05/100 = 5 \cdot 10^{-4}$, meaning that for a trait with SNP heritability 0.5, there would be 1,000 causal variants, 100 of which are on chromosome 1. The GWAS power for these effects is 1 for the sample size and the frequency distribution of the SNPs used here. These effects are much larger on average than would typically be observed in real data for common complex traits. However, this choice is important to be able to explore the TPR and FDR of MR approaches in a “best case” scenario.

For the low power scenario, $h_X^2 = h_Y^2 = 0.01$, $m_c = 200$, $\sigma_{\epsilon_X} = \sigma_{\epsilon_Y} = 0.99$. The expected variance contributed by a single SNP variant is $0.01/200 = 5 \cdot 10^{-5}$, meaning that for a trait with SNP heritability 0.5, there would be 10,000 causal variants genome-wide, 200 of which are on chromosome 1. The GWAS power for these effects is 0.25 for the sample size and the frequency distribution of the SNPs used here. This genetic architecture is closer to that typically observed in real data for common complex traits and returns a smaller but non-zero set of instrument variables. For each of the four scenarios, we generate 100 data sets, each with a sample size of 458,747 individuals.

Having generated the data, we analyze each of the 400 data sets, using all 654,172 SNPs on chromosome 1. We use part 1 of the CI-GWAS algorithm to select instrument variables from the full set of chromosome 1 SNPs. We do this by selecting the SNPs that CI-GWAS returns as adjacent to X in the graph, excluding those adjacent to both X and Y . We then use these SNPs in either inverse-variance-weighted (IVW) MR, or MR-EGGER as implemented in the R package *MendelianRandomization*. This requires the marginal least squares GWAS regression coefficients for both X and Y for the CI-GWAS-selected SNPs. This returns an estimate of the ACE for the causal effect of X on Y , $X \rightarrow Y$. We compare this to the ACE estimates obtained when selecting instrument variables from the marginal least squares GWAS regression coefficients for X , applying a standard genome-wide significance threshold of $p\text{-values} \leq 5 \times 10^{-8}$ to select an independent (within 1MB, with LD $R^2 < 0.001$) set of instrument variables from the full set of chromosome 1 SNPs, as is commonly conducted in MR studies. Valid instrument variables for MR are the columns of W_{s_X} that are not within W_{s_Y} .

Latent pleiotropy scenario. We then simulate the following:

$$\begin{aligned} U &= W_{s_U} \beta_U + \epsilon_U \\ X &= W_{s_X} \beta_X + \lambda_{U_X} U + \epsilon_X \\ Y &= W_{s_Y} \beta_Y + \lambda_X X + \lambda_{U_Y} U + \epsilon_Y \end{aligned} \quad (4)$$

where U is a latent unobserved random confounding variable with W_{s_U} a matrix containing the observed genotypes, with one row for each of 458,747 individuals and one column for each of $m_c = 400$ randomly selected SNPs that are causal for U . Each column of W_{s_U} is scaled and centered to a z-score value. The effects and residuals are drawn from a normal distribution $\beta_U \sim N(0, h_U^2/m_c)$ and $\epsilon_U \sim N(0, \sigma_{\epsilon_U})$, where the SNP heritability of the latent confounder $h_U^2 = 0.1$ and $\sigma_{\epsilon_U} = 0.9$. U has a causal effect on both X and Y , with $\lambda_{U_X} = \lambda_{U_Y} = 0.2$. We then simulate the same four settings as described above for the baseline scenario, where the average causal effect of X on Y is either $\lambda_X = 0$, or $\lambda_X = 0.2$ and the power to detect the genotype effects on X and Y are either high or low. For each of these four scenarios, we generate 100 data sets, each with a sample size of 458,747 individuals, giving another 400 data sets. We repeat the analysis of this data in the same way as described for the baseline scenarios.

Here, we generate an unobserved confounder U that influences both the exposure X and the outcome Y , but we assume no explicit confounding of the instrument variables and the outcome Y (the independence assumption in MR). However, it is possible that the 400 randomly selected SNPs that are causal for U overlap those selected for either X and Y . Genotypes that are causal for U , W_{s_U} , will be correlated with both X and Y and thus may be detected in standard GWAS association testing. Valid instrument variables for MR are the columns of W_{s_X} that are not within W_{s_Y} or W_{s_U} .

Horizontal pleiotropy scenario. We then repeat the data generation for the latent pleiotropy scenarios, but we modify the graph to create additional horizontal pleiotropy (heritable confounding). Horizontal pleiotropy violates the exclusion restriction assumption of MR approaches, and occurs when genetic variants influence the outcome (directly or indirectly, through other traits) independently of the hypothesized exposure. To simulate an independent pathway between the instrument variables for X and the outcome Y , that acts other than through the exposure X , we simulate half of the causal SNPs in the set W_{s_X} and W_{s_Y} to be the same, by randomly selecting this set jointly for X and Y . For the other half of the causal SNPs, we randomly select SNPs for X and Y independently. For the SNPs shared by X and Y , we draw their effects from a bivariate normal distribution $\beta_{X,Y} \sim N(0, V/m_c)$, with $V = \begin{pmatrix} 0.5h_X^2 & 0.8 \cdot 0.5h_X^2 \\ 0.8 \cdot 0.5h_X^2 & 0.5h_Y^2 \end{pmatrix}$, with $h_X^2 = h_Y^2$. This means that half of

the variance attributable to the SNPs is allocated to half of the causal variants, which are shared by X and Y with a genetic correlation of 0.8. For the other half of the causal variants, which are selected independently, the effects are drawn from independent normal distributions $\beta_X \sim N(0, 0.5h_X^2/m_c)$, $\beta_Y \sim N(0, 0.5h_Y^2/m_c)$. We use the same power and ACE settings as described above to simulate four scenarios with horizontal pleiotropy. For each of these four scenarios, we generate 100 data sets, each with a sample size of 458,747 individuals, giving another 400 data sets. We repeat the analysis of this data in the same way as described for the baseline scenarios.

In this setting, valid instrument variables for MR are the columns of W_{s_X} that are not within W_{s_Y} or W_{s_U} . Here, selecting IVs for which the independence assumption holds is the same as selecting SNPs for which the exclusion restriction holds. This is because potential confounding creates an association between SNPs and Y . Thus, selecting SNPs that CI-GWAS returns as adjacent to X in the graph, excluding those adjacent to both X and Y , controls for this form of confounding.

Horizontal pleiotropy scenario with multiple mediators. We repeat the data generation for the four scenarios with horizontal pleiotropy, but we modify the graph to have two mediator variables $X1$ and $X2$. For the half of the causal SNPs shared by $X1$, $X2$ and Y , we draw their effects from a trivariate normal distribution $\beta_{X1,X2,Y} \sim N(0, V/m_c)$, with

$$V = \begin{pmatrix} 0.5h_{X1}^2 & 0.5 \cdot 0.5h_{X1}^2 & 0.5 \cdot 0.5h_{X1}^2 \\ 0.5 \cdot 0.5h_{X1}^2 & 0.5h_{X2}^2 & 0.5 \cdot 0.5h_{X2}^2 \\ 0.5 \cdot 0.5h_{X1}^2 & 0.5 \cdot 0.5h_{X2}^2 & 0.5h_Y^2 \end{pmatrix}, \text{ with } h_{X1}^2 = h_{X2}^2 = h_Y^2.$$

The effects of the other half of the causal SNPs are drawn from independent normal distributions. We use the same power settings as described above. We simulate the average causal effect of $X1$ to Y to be either 0 or 0.2, while the causal effect of $X2$ to Y is always set to 0. For each of these four scenarios with horizontal pleiotropy, we generate 100 data sets, each with sample size of 458,747 individuals, giving an additional 400 data sets. We analyze these data with IVW and multivariable IVW (MVIW) as implemented in the R package *MendelianRandomization*. Here, as above, we provide the MR methods with either independent GWAS selected IVs, or CI-GWAS selected IVs. Valid instrument variables for MR are the columns of $W_{s_{X1}}$ that are not within $W_{s_{X2}}$, W_{s_Y} or W_{s_U} .

Simulation metrics. In total, we analyze 1600 data sets, 100 each for 16 different settings, each with a sample size of 458,747 individuals and 654,172 SNPs. To compare the causal inference performance, we calculate the false positive rate, FPR = false positives / (false positives + true negatives), for each MR approach for each of the settings. FPR measures the proportion of simulation replicates where $ACE = 0$, but MR methods incorrectly identify a causal effect of the exposure X and outcome Y (at p -value $\leq 0.05/800$), out of all actual negatives (100 replicates per setting where $ACE = 0$). For the horizontal pleiotropy settings, the true positive rate (power), TPR = true positives / total positives, is calculated for the identified IVs, giving the proportion of IVs correctly identified by CI-GWAS, or standard GWAS across settings. We also calculate the proportion of selected IV SNPs for X that are also causal for Y through a path that is not through X (labeled PHP: proportion of horizontally pleiotropic IVs).

In the settings with two mediator variables, we calculate the false positive rate, FPR = false positives / (false positives + true negatives), of two commonly used MR approaches of inverse-variance weighting (IVW) and multivariable inverse-variance weighting (MVIW). FPR measures the proportion of simulation replicates where $ACE = 0$, but MR methods incorrectly identify a causal effect of one of the exposures $X1$ or $X2$ and the outcome Y (at p -value $\leq 0.05/800$), out of all actual negatives (600 occurrences in total, comprised of all of the 400 simulation replicates where $ACE = 0$ for $X2$ to Y , and the 200

replicates where $ACE = 0$ for $X1$ to Y). We also compare the TPR (power), TPR = true positives / total positives, the proportion of causal relationships correctly identified by CI-GWAS, or standard GWAS, across settings. We give the true discovery rate, TDR = true positives / total discoveries, the proportion of identified causal relationships by CI-GWAS, or standard GWAS, that are true across settings. First, TPR and TDR are calculated for all tested relationships. We then calculate TPR and TDR, setting all non-adjacent edges in the graph to $ACE = 0.0$, as the undirected graph does not find a causal relationship. Finally, FPR, TPR and TDR are calculated after all edges in the partial ancestral graph (PAG) for which no potential directed relationship is found are set to $ACE = 0.0$.

Complex scenario. To simulate a more complex setting, we follow a similar strategy as described above for the simple setting but expand it to twelve continuous traits. We select 50 causal markers for each trait using the same imputed SNP data from chromosome 1. We simulate a marker-trait and trait-trait adjacency matrix G , causal effects A , and phenotypic data:

First, we simulate a randomly directed acyclic graph G and generate synthetic data based on the specified graph structure, with the following parameters: the number of observations in the dataset (n), the number of SNP variables (m), the number of treatment variables (p), the number of latent variables (l), the average node degree (deg), the probability of pleiotropy (a SNP affecting multiple traits) (*probPleio*), and bounds for the causal effect sizes (*lo* and *hi*). The following steps are used to generate the data:

- Step 1. Initialize a graph structure represented by a sparse adjacency matrix G of dimensionality $d \times d$, where $d = m + l + p$. Set all elements of G to zero. Here, we choose $l = 2$; $p = 10$.
- Step 2. Generate random adjacencies by setting values of G to 1 with probability $\frac{deg}{d}$. Here, we choose average degree $deg = 3$.
- Step 3. Create additional marker-trait links, by iterating over each SNP variable and randomly connecting it to latent and phenotype variables with a probability of pleiotropy (*probPleio*). The connections are added to the adjacency matrix G accordingly. Here, we select *probPleio* = 0.2 or 0.0 for 20 replicates each.
- Step 4. Given the established adjacency matrix, sample the causal effects for the edges from a uniform distribution within specified bounds (*lo* and *hi*), and store them in matrix A as a true causal effect matrix. Here, we use a uniform distribution $U[(-0.01, -0.0001) \cup (0.0001, 0.01)]$ to set the causal effect sizes between markers and traits, and a uniform distribution $U[(-0.2, -0.001) \cup (0.001, 0.2)]$ between traits. This approach allows us to simulate a wide range of causal effects, and to have separate strengths for marker-trait and trait-trait relationships, which reflects the pattern we observe in the UK Biobank data.
- Step 5. Simulate data based on a structural equation model. For each variable V_j in G , check if V_j has parents ($Pa(V_j)$) in G . Consider only variables with an index $i < j$ as parents. If V_j has no parents, assign it random values drawn from a standard Normal distribution. If V_j has parents, a linear regression model is constructed using the parental variables' values. In other words, apply the following structural equation model to generate V_j :

$$V_j = f(Pa(V_j)) = b_0 + \sum_i b_i Pa_i(V_j) + \epsilon_x, \quad (5)$$

where b_0 is the intercept (set to zero) and coefficients for the regression model are the appropriate b_i values selected from A for the i^{th} parent of V_j . The sampling error ϵ_x is then sampled from a standard normal with variance $1 - \text{var}(\sum_i b_i Pa_i(V_j))$. This results in all variables having unit variance and zero mean, up to a sampling error.

We simulated data for all 458,747 UK Biobank individuals and conducted 40 replicates (20 for each pleiotropy regime), which we analyzed using CI-GWAS. For the MR analysis comparison, we used IVW and MVIVW as well as two additional methods, MR-PRESSO, and CAUSE and their multivariable extensions where appropriate. MR-PRESSO was implemented using code from the original publication (version 1.0)⁸ with outlier correction. We employed the CAUSE R package (version 1.2.0.0331)¹¹ to perform the CAUSE analysis. Unless stated we used default options given by each software. When comparing against MR methods, we supply the methods where appropriate with IVs selected by the standard procedure of running marginal association tests. To perform the marker selection, we run PLINK 2.0³⁶ with `-glm omit-ref -variance-standardize` on the marker data and the trait data. Secondly, we clump the results by running PLINK 2.0 with the options `-clump -clump-field P -clump-p1 10e-8` (or `-clump-p1 10e-4`) and `-clump-r2 0.1 -clump-snp-field ID` on the results. We then select the focal markers across all clumps for downstream analysis.

To compare the causal inference performances, we calculate TPR and TDR of the returned trait-trait causal relationships. To calculate TPR and TDR, we compare to the directed adjacency matrix of the true DAG. We calculate the sum of all true edges P , which is the total number of possible true edges. We compute TP by logical AND between the true matrix and the inferred causal relationship matrix. We compute FP, which is the number of inferred, but not true, edges. We then calculate $TPR = TP/P$ and $TDR = TP / (TP + FP)$. For the TDR calculation for single variable MR, the true matrix differs: it is the matrix of all causal paths in the true graph over the traits. We calculate this because MR estimates causal effects between variables, without regard (a distinction) between edges and paths, or in other words between direct or mediated (through other variables) causality. Multivariable MR (MV-MR) differs as it seeks to infer direct effects, and thus, for MV-MR, we calculate TP, FP, TPR and TDR in the same way as for CI-GWAS.

Since CAUSE estimates path effects instead of direct effects, we compare the estimated effect to the total effect in the true DAG over all path lengths when assessing its performance. The matrix of total effects M_{tot} is calculated as $M_{tot} = \sum_{l=1}^n M^l$, where M is the matrix of the true direct causal effects of the DAG and n is the number of traits in the DAG.

Assessment of the error rate when including binary outcomes in the analysis. We also conducted a simulation to illustrate that treating binary outcomes as continuous can lead to increased error rates when conducting conditional testing of the sort that is done in the PC-algorithm (Supplementary Fig. 1). We conducted two small simulations: (1) We drew independent samples of $x \sim \text{Binom}(2, 2 \times 10^{-3})$ and $y \sim \text{Binom}(1, 0.028)$ and tested $CI(x, y)$ as described above with $\alpha = 0.05$; (2) We simulated from the DAG: $x \rightarrow y_{cont} \rightarrow y_{bin}$ where $x \sim \text{Binom}(2, 2 \times 10^{-3})$, $y_{cont} \sim \mathcal{N}(\beta x, 1)$, $y_{bin}^l \sim \mathcal{N}(y y_{cont}, 1)$, and y_{bin} is 1 wherever $y_{bin}^l > \Phi^{-1}(K)$ and 0 otherwise. We used $\beta = (2 \times MAF \times (1-MAF))^{-0.03}$ and $\gamma = 0.4$. We drew $n = 458746$ samples for each variable and 10,000 replicates for each scenario. We calculated correlations between the variables in three different ways: pearson - Pearson correlations between all variables; polyserial: polyserial between (x, y_{bin}) and (y_{cont}, y_{bin}) ; polychoric: polychoric between (x, y_{bin}) , polyserial between (y_{cont}, x) and (y_{cont}, y_{bin}) . We calculated the correlations using the *hetcor* function in the *polycor* R package⁴⁸.

Power. We present theoretical limits of the percentage of variance in a trait explained by a marker, below which our algorithm labels associations as insignificant. We calculate these limits based on the conditional independence test, which tests the significance of the sample correlation between two variables V_i and V_j , $r_{V_i V_j}$, and rejects the null

hypothesis $H_0 : \rho_{V_i V_j} = 0$ ($\rho_{V_i V_j}$ being the unobserved population correlation coefficient) if

$$|z(r_{V_i V_j})| \geq \frac{\Phi^{-1}(1 - \frac{\alpha}{2})}{\sqrt{n-3}}, \quad (6)$$

where z is Fisher's z-transform of the correlation, n is the sample size, α is the threshold applied to the test and Φ is the cumulative density function of the standard normal distribution. n is calculated from the standard errors of the provided correlations, s_r , using

$$s_r = \frac{1-r^2}{\sqrt{n-3}} \iff n = \left(\frac{1-r^2}{s_r} \right)^2 + 3 \quad (7)$$

To calculate detection limits, we also use the standard errors of the correlations provided in the input, but make the approximation $n = \frac{1}{s_r^2}$, since the correlations considered are very small, such that $1 - r^2 \approx 1$, and $n - 3 \approx n$ for large n . We then calculate the $r_{V_i V_j}^2$ that satisfies the equality in the above equation, and interpret it as the variance explained in V_j by V_i at the detection limit.

To calculate the standard errors for biserial correlations, r_b , (used for correlations between markers and binary traits), we used Hunter's and Schmidt's approximation⁴⁹:

$$s_{r_b} = \frac{\sqrt{pq}(1-r^2)}{\phi(\Phi^{-1}(p))\sqrt{n-3}}, \quad (8)$$

where ϕ is the probability density function of the standard normal distribution, p is the prevalence of the binary trait and $q = 1 - p$.

As the sample size is calculated from the standard error as stated above (using the standard error to sample size relation of the product-moment correlation), the resulting estimated sample size $\hat{n}$ is strictly smaller than n in Equation (8). This reduction in the effective sample size is a consequence of the information loss resulting from the dichotomization of the liability-scale distribution of the binary trait, and has been shown to be effective in controlling the number of false positive test results caused by treating a binary variable with low prevalence as Gaussian²⁵.

To calculate the n needed to get an effective sample size $\hat{n}$ (and the corresponding detection limit) for a binary trait with prevalence p , we set $s_{r_b} = s_r$ using Equations (7) and (8) and the approximations $1 - r^2 \approx 1$ and $n - 3 \approx n$, and solve for n :

$$\frac{\sqrt{pq}}{\phi(\Phi^{-1}(p))\sqrt{n}} = \frac{1}{\sqrt{\hat{n}}} \iff n = \hat{n} \frac{pq}{\phi(\Phi^{-1}(p))^2} \quad (9)$$

At a sample size of $n = 450,000$, our procedure has a theoretical lower limit of detection for a variant at $(\Phi^{-1}(1 - \frac{\alpha}{2})(n-3)^{-\frac{1}{2}})^2 \approx 0.0025\%$ variance explained, where Φ is the cumulative density function of the standard normal distribution and α is a sensitivity parameter for the CI tests (set at $\alpha = 1e-3$). For GWAS, taking a variant with the same explanatory power $r^2 = 0.0025\%$, with approximate standard error $= \frac{1-r^2}{\sqrt{n}}$, calculating the non-centrality parameter of a chi-square

distribution of $NCP = (\frac{\sqrt{0.000025}}{SE})^2$ and using the chi-square distribution function to compare the NCP to the quantile of the distribution at the point corresponding to a testing correction threshold of $p\text{-value} \leq 5 \times 10^{-8}$ give the expected power. At $r^2 = 0.0025\%$ the power for a standard genome-wide association study at a sample size of

$n = 450,000$, would be 2% at the genome-wide significance level of $5 \cdot 10^{-8}$, which is also at the lower bound of GWAS detection power. Thus, we explicitly set α for our approach to have similar power as standard association testing. An example power curve is given in Supplementary Fig. 9 where we simulate a graph of two variables with a single edge and we vary the strength of the ACE and thus the correlation between the variables at sample size of $n = 450,000$. For each simulated correlation value, we generate 200 datasets and estimate power via Monte Carlo simulation, comparing it with the theoretical power of GWAS for a trait-variant relationship with the same correlation.

Postprocessing of CI-GWAS-selected parental markers

Regarding the markers identified as parental to traits in the UKB analysis, we were interested in their potential biological significance, as well as their concordance with previous findings or their novelty.

To address the former, we downloaded ‘Homo_sapiens.GRCh37.87.chr.gff3’ from Ensembl⁵⁰, release 110. We then extracted all genes (except for pseudogenes) and their exon positions and sorted the file using

```
$ sort -s -k1,1 -k4,4n \
Homo_sapiens.GRCh37.87.chr.filtered.gff3 \
> Homo_sapiens.GRCh37.87.chr.filtered.sorted.gff3
```

We then searched for the nearest feature to each identified marker using BEDtools⁵¹ version 2.30.0 with

```
$ bedtools closest -d \
-a ci_gwas_identified_markers.bed \
-b Homo_sapiens.GRCh37.87.chr.filtered.sorted.gff3
```

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

This project uses the UK Biobank data under project number 35520. UK Biobank genotypic and phenotypic data is available through a formal request at (<http://www.ukbiobank.ac.uk>). All of the data generated in this study are provided in the Source Data file. Source data are provided with this paper.

Code availability

The CI-GWAS code is fully open source and available at <https://github.com/medical-genomics-group/ci-gwas> and has been archived on Zenodo⁵². The scripts used to execute the model are available at <https://github.com/medical-genomics-group/ci-gwas>. R version 4.2.1 is available at <https://www.r-project.org/>. Plink version 1.9 is available at <https://www.cog-genomics.org/plink/1.9/>.

References

- Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for mendelian randomization. *Stat. methods Med. Res.* **26**, 2333–2355 (2017).
- Sanderson, E. Multivariable Mendelian randomization and mediation. *Cold Spring Harb Perspect. Med.* **11**, a038984 (2021).
- Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 1–12 (2018).
- Hemani, G. et al. The mr-base platform supports systematic causal inference across the human phenome. *elife* **7**, e34408 (2018).
- Pingault, J.-B. et al. Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).
- Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
- O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
- Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of wide-spread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
- Burgess, S. & Thompson, S. G. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
- Zuber, V., Colijn, J. M., Klaver, C. & Burgess, S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nat. Commun.* **11**, 29 (2020).
- Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **52**, 740–747 (2020).
- Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *M. Int. J. Epidemiol.* **44**, 512–525 (2015).
- Burgess, S., Butterworth, A. & Thompson, S. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
- Yavorska, O. O. & Burgess, S. Mendelian randomization: an R package for performing mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
- Zuber, V. et al. Multi-response Mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. *Am. J. Hum. Genet.* **110**, 1177–1199 (2023).
- Lin, Z., Xue, H. & Pan, W. Combining Mendelian randomization and network deconvolution for inference of causal networks with GWAS summary data. *PLoS Genet.* **19**, e1010762 (2023).
- Amar, D., Sinnott-Armstrong, N., Ashley, E. A. & Rivas, M. A. Graphical analysis for phenome-wide causal discovery in genotyped population-scale biobanks. *Nat. Commun.* **12**, 350 (2021).
- Pearl, J. *Causality* (Cambridge University Press, 2009).
- Spirites, P. et al. *Causation, Prediction, and Search* (MIT Press, 2000).
- Colombo, D., Maathuis, M. H., Kalisch, M., Richardson, T. S. et al. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* **40**, 294–321 (2012).
- Colombo, D. & Maathuis, M. H. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15**, 3741–3782 (2014).
- Hesslow, G. Two notes on the probabilistic approach to causality. *Philosophy Sci.* **43**, 290–292 (1976).
- Zhang, J. & Spirtes, P. Detection of unfaithfulness and robust causal inference. *Minds Mach.* **18**, 239–271 (2008).
- Zarebavani, B., Jafarnejad, F., Hashemi, M. & Salehkaleybar, S. cupc: Cuda-based parallel pc algorithm for causal structure learning on GPU. *IEEE Trans. Parallel Distrib. Syst.* **31**, 530–542 (2020).
- Cui, R., Groot, P. & Heskes, T. L. Learning causal structure from mixed data with missing values using Gaussian copula models. *Stat. Comput.* **29**, 311–333 (2019).
- Neto, E., Keller, M., Attie, A. & Yandell, B. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann. Appl. Stat.* **4**, 320–339 (2010).

27. Mckeigue, P., Krohn, J., Storkey, A. J. & Agakov, F. Sparse instrumental variables (spiv) for genome-wide studies. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. & Culotta, A. (eds.) *Advances in Neural Information Processing Systems*, vol. 23. https://proceedings.neurips.cc/paper_files/paper/2010/file/4476b929e30dd0c4e8bdbcc82c6ba23a-Paper.pdf (Curran Associates, Inc., 2010).
28. Liu, Y.-S., Lin, Y.-C., Lin, M.-C., Wu, C.-C. & Wang, T.-N. Association of blood lipid profiles and asthma: A bidirectional two-sample mendelian randomization study. *Ann. Hum. Genet.* **88**, 307–319 (2024).
29. Cheng, D. et al. Toward unique and unbiased causal effect estimation from data with hidden variables. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 6108–6120 (2022).
30. Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.* **48**, 713–727 (2019).
31. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
32. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
33. Lagemann, K., Lagemann, C., Taschler, B. & Mukherjee, S. Deep learning of causal structures in high dimensions under data limitations. *Nat. Mach. Intell.* **5**, 1306–1316 (2023).
34. Gkatzionis, A. & Burgess, S. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *Int. J. Epidemiol.* **48**, 691–701 (2019).
35. Chang, C. C. et al. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 13742 (2015).
36. Purcell, S. et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
37. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.* **11**, 1561–1592 (2017).
38. Pearson, K. On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B Exceeds (or Falls Short of) a given intensity is recorded for each grade of A. *Biometrika* **7**, 96–105 (1909).
39. Pearson, K. Mathematical contributions to the theory of evolution. VIII. On the correlation of characters not quantitatively measurable. [Abstract]. *Proc. R. Soc. Lond.* **66**, 241–244 (1899).
40. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
41. Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172**, 1873–1896 (2008).
42. Davey Smith, G. & Ebrahim, S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
43. Boef, A. G., Dekkers, O. M. & Le Cessie, S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int. J. Epidemiol.* **44**, 496–511 (2015).
44. Zheng, J. et al. Recent developments in Mendelian randomization studies. *Curr. Epidemiol. Rep.* **4**, 330–345 (2017).
45. Hernán, M. & Robins, J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* **4**, 360–372 (2006).
46. Burgess, S. & Yavorska, O. *MendelianRandomization: Mendelian Randomization Package*. <https://CRAN.R-project.org/package=MendelianRandomization>. R package version 0.10.0 (2024).
47. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
48. Fox, J. *Polycor: Polychoric and Polyserial Correlations*. <https://CRAN.R-project.org/package=polycor>. R package version 0.8-1 (2022).
49. Hunter, J. E. & Schmidt, F. L. Dichotomization of continuous variables: the implications for meta-analysis. *J. Appl. Psychol.* **75**, 334–349 (1990).
50. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
51. Quinlan, A. R. & Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 814–842 (2010).
52. Robinson, M. R. Ci-gwas. *Zenodo* 10.5281/zenodo.21791037 (2026).

Acknowledgements

We thank Zoltan Kutalik and members of the Robinson group at ISTA for their comments, which improved this manuscript. We would like to acknowledge the participants and investigators of the UK Biobank study. High-performance computing was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Author contributions

M.R.R., M.M., and N.M. conceived and designed the study, conducted the analysis and wrote the paper. M.J.B. and M.R.R. provided study oversight. I.K. contributed to study design and analysis. M.B. contributed to the analyses. All authors approved the final manuscript prior to submission.

Funding

This work was funded by a research collaboration agreement between Boehringer Ingelheim and the research group of MRR at the Institute of Science and Technology Austria. Additional funding was also provided by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. MB was funded by the National Science Center Poland grant SONATINA UMO-2021/40/C/NZ2/00218.

Competing interests

This study received research funding from Boehringer Ingelheim through a research collaboration agreement with MRR at the Institute of Science and Technology Austria.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-76877-7>.

Correspondence and requests for materials should be addressed to Nick Machnik or Matthew R. Robinson.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026