Causal inference for multiple risk factors and diseases from genomics data

Nick Machnik[1†], Mahdi Mahmoudi[1†], Malgorzata Borczyk[2], Ilse Krätschmer[1], Markus J. Bauer[2], Matthew R. Robinson[1*]

[1]Institute Of Science and Technology Austria, Am Campus 1, Klosterneuburg, 3400, Austria.
[2]Laboratory of Pharmacogenomics, Maj Institute of Pharmacology PAS, Smetna 12, Krakow, 31155, Poland.
[3]Boehringer Ingelheim RCV GmbH & Co KG, Dr. Boehringer-Gasse 5-11, Vienna, 1120, Austria.

*Corresponding author(s). E-mail(s): matthew.robinson@ist.ac.at;
Contributing authors: nick.machnik@ist.ac.at; seyedmahdi.mahmoudi@ist.ac.at; gosborcz@if-pan.krakow.pl; ilse.kraetschmer@ist.ac.at; markus_johann.bauer@boehringer-ingelheim.com;
†these authors contributed equally to this work.

Abstract

Statistical causal learning in genomics relies on the instrumental variable method of Mendelian Randomization (MR). Currently, an overwhelming number of MR studies purport to show causal relationships among a wide range of risk factors and outcomes. Here, we show that selecting instrument variables from genome-wide association study estimates leads to high false discovery rates for many MR approaches, which can be greatly reduced by employing a graphical inference approach which: (i) explicitly tests instrumental variable assumptions; (ii) distinguishes direct from indirect factors in very high-dimensional data; (iii) discriminates pleiotropic from trait-specific markers, controlling for LD genome-wide; (iv) accommodates rare variants and binary outcomes in a principled way; and (v) identifies potential unobserved latent confounding. For 17 traits and 8.4M variants recorded for 458,747 individuals in the UK Biobank, we show that standard MR analysis gives an abundance of findings that disappear under stringent assumption checks, with many relationships reflecting potential unmeasured confounding. This implies that mixtures of temporal precedence and potential for reverse-causality prohibit understanding the underlying nature of phenotypic and genetic correlations in biobank data. We propose that well-curated longitudinal records are likely needed and that our approach provides a first-step toward robust principled screening for potential causal links.

Keywords: Causal inference, Mendelian randomisation, Genotype-phenotype relationships, Genome-wide association studies (GWAS), Graphical models

Introduction

Distinguishing causation from correlation is fundamentally important for understanding the etiology of common complex disease. Randomized controlled trials (RCTs) are the gold standard for causal inference, but their difficulty, expense, and often also ethical considerations, prohibit examining many potentially disease-associated risk factors among the vast array of human complex traits. Currently, the most commonly used methods for statistical causal inference in a genetic or biomedical setting center around the 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, direct or indirect, of one or more mediator variable(s) onto one or more outcome variable(s) [3–15]. Multivariate MR (MVMR) considers multiple exposure variables at the same time to distinguish risk factors with potential indirect paths, from those of potential direct effect on each disease. While easy to apply in practice, there are major limitations to MR that make inference unreliable: (i) testing is conducted using a single form of fixed graphical pattern; (ii) SNPs with strongest genotype-phenotype association test statistics are used as instrumental variables (IVs), which relies on a number of key assumptions that are generally assumed to hold as opposed to being explicitly tested for. This has created the current scenario where an overwhelming number of recent MR studies purport to show causal relationships among a wide range of risk factors and outcomes.

Theoretical developments in statistical causal inference show that graphical modelling can be used to learn the partial correlations among hundreds of variables, whilst controlling for potential unobserved confounding within a single model [16–18]. However, statistical causal discovery in large-scale genome-wide association study (GWAS) data, where there are many highly correlated potential risk factors and outcomes and millions of genetic markers, make learning difficult. Currently, there are no statistical causal inference algorithms able to accommodate millions of correlated DNA variants and facilitate high-throughput screening of individual-level, or summary-statistic data for the discovery of potential causal links among modifiable risk factors and disease outcomes.

Here, we propose an approach for conducting Causal Inference in Genome-Wide Association Studies (CI-GWAS), that is capable of inferring a single graph at scale describing the relationships among millions of genetic markers, tens of risk factors and multiple disease outcomes. We utilise the inferred graph for more principled fine-mapping of potential IVs, controlling for linkage disequilibrium genome-wide and fulfilling the 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 better than using markers of strongest genotype-phenotype association test statistics. Our approach to IV selection is comparable to multivariate fine-mapping, but with performance greater than the current state-of-the-art approach. We show that conducting MVMR by explicitly utilising the inferred graph to select IVs that likely fulfil MR assumptions, improves upon the high false discovery rates of current MR approaches. We show that this holds even when using rare variants and rare disease outcomes when placing all disease relationships

on the liability-scale to correct for differences in outcome prevalence. Applying CI-GWAS to summary statistic data of 8.4M SNPs and 17 phenotypes requires only 20 minutes of compute time on a standard GPU server. In the UK Biobank, we reveal that many findings from standard MVMR disappear when following our graphical inference approach and that many relationships represent unmeasured confounding. Our approach can be used on individual-level or summary statistic data and provides a solution to facilitating robust one- or two-sample MR inference on large-scale genotype-phenotype data.

Results

The structure of CI-GWAS

CI-GWAS is a method that aims to learn a graphical model that describes the causal relationships among genetic and trait variables. It consists of three parts. Part 1 is the construction of the skeleton, an undirected graph describing all possible direct causal relationships between all observed variables in the model; part 2 directs the edges in the skeleton to obtain a partial ancestral graph (PAG), while taking into account potential confounding through latent variables; and part 3 infers potential causal relationships and their effects by taking the intersection of significant relationships obtained by *MVIVW* and i) traits that are adjacent in the skeleton (denoted as *CI-GWAS (skeleton + MVIVW)*), or ii) causal relationships in the PAG (denoted as *CI-GWAS (PAG + MVIVW)*), where the latter comes with lower power but higher precision.

As we describe in the Methods and show in simulation below, part 1 fine-maps genetic effects to those variables that are most likely to have direct effects on traits and yields valid IVs for multivariate MR. The Methods describe how our algorithm, which is based on a GPU-accelerated version of the PC-stable algorithm [19, 20], overcomes exponential runtime scaling with the number of variables, enabling a full marker-trait graph to be constructed, extending previously proposed graphical methods that were limited to few hundreds [21] or thousands [22] of genetic variables or relied on the construction of separate graphs for trait-trait and marker-trait relationships with pre-selected genetic variants by GWAS [23]. Part 2 is based on the RFCI algorithm [18], for which we introduce i) a novel way to obtain order-independent separation sets that offer a conservative way of orienting trait relationships and ii) the incorporation of the prior knowledge that markers precede traits temporally, which dramatically reduces computation time. We use only summary statistic data to do this and denote this modified version as sRFCI. Finally, part 3 combines graphical inference and MVIVW, taking advantage of the principled instrumental selection in part 1, which stands in contrast to most other current MR approaches, where IVs are selected by taking variants with the highest marginal association test statistics. This also advances previous work on graphical inference [23] as marker selection is through joint estimation, many variables can be conditioned on at once, and pleiotropic markers can be used for inference, promising more accurate IV selection and improved inference in MR as shown below.

CI-GWAS takes as input (i) the matrix of marker-maker correlations (LD matrix); (ii) a matrix of marker-trait correlations; (iii) a matrix of trait-trait correlations. This

makes CI-GWAS applicable to summary statistic data, as LD-matrices, effect sizes of marker-trait associations, minor-allele frequencies, trait variances and trait-trait correlations can be shared whilst preserving privacy.

Incorporating disease outcomes and rare variants

The PC-algorithm that is at the core of CI-GWAS relies on partial correlations vanishing for (conditionally) independent variables. This generally holds for Gaussian variables [24], but not necessarily when variables are binary or ordinal, as is the case for case-control outcomes. We find that treating case-control outcomes like Gaussian variables can lead to largely elevated numbers of false positive findings (Figure S1). This is most pronounced for independence tests between markers with very low minor-allele-frequencies and binary traits with very unbalanced numbers of cases and controls (Figure S1). As has been shown in previous work [25], 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. At the sample size, lowest minor-allele-frequency and lowest trait prevalence in the UK Biobank data used in this study, we find that using polyserial correlations between markers and case-control traits is effective in controlling the numbers of false positives (Figure S1).

CI-GWAS selects candidate instrumental variables at scale

We then show that the discovery of valid IVs can be framed as a multivariate fine-mapping task, as we wish to identify genetic marker associations that represent conditional dependencies, having associations that are conditional on other traits and genetic variants. Graphical inference allows identification of marker-trait associations that fulfil core MR assumptions for IVs, such as the absence of horizontal pleiotropy and the independence of markers of confounders of the exposure-outcome pair. This is done by conducting conditional-independence (CI) tests that correspond to the MR assumptions: testing if a marker is marginally or conditionally independent of the outcome, which checks both of the above mentioned assumptions (see Methods for a thorough explanation). Conducting CI-tests of each marker-trait combination conditioning on other markers and traits in this way, fine-maps genetic associations across traits genome-wide and identifies genetic variants that are pleiotropic from those that are trait-specific, allowing for principled IV selection.

We assess the ability of CI-GWAS to identify causal markers and correctly place edges between markers and traits in the causal graph in a simulation study, where we benchmark against the current state-of-the-art mvSuSiE [26], a multivariate fine-mapping approach shown to outperform other fine-mapping methods. We simulate from a structural equation model consisting of two continuous traits A and B each with a set of 200 causal markers X_A and X_B chosen at random from 654,172 imputed SNP markers on chromosome 1 recorded for 458,746 individuals in the UKB, allowing for some shared fraction between X_A and X_B of size $\gamma|X_A|$. We simulate data for A and B using causal effect sizes $\beta \in \{0, 0.25\}$ for $A \to B$, with marker effects chosen such that the heritability of each trait was fixed at 0.2.

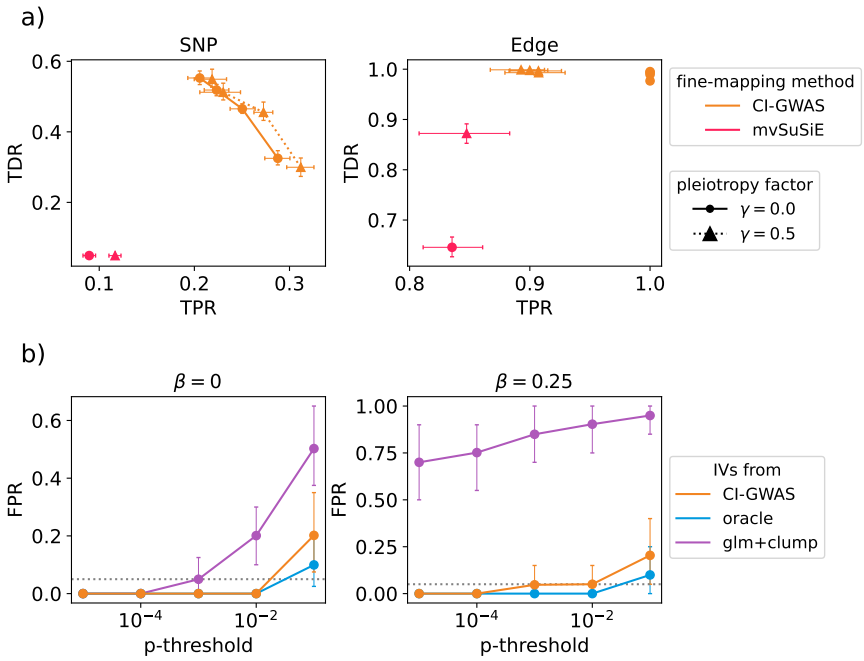

Fig. 1: Fine-mapping and causal inference performance of CI-GWAS, mvSuSiE and MVIVW in simulation with strong LD and two traits. (a) shows the fine mapping performance of CI-GWAS and mvSuSiE as given by the true discovery rate (TDR) against the true positive rate (TPR) of globally selected SNPs (CI-GWAS) or credible sets (mvSuSiE) in the left panel, and of direct SNP-trait effects in the right panel. For mvSuSiE, a credible set was counted as true positive if it contained at least one true causal SNP. Error bars indicate the 95% confidence interval of the mean, bootstrapped. For CI-GWAS we present results at alpha values $10^{-8}, 10^{-6}, 10^{-4}$ and 10^{-2}. For mvSuSiE we apply a p-value threshold of 10^{-2}. (b) shows the false positive rate of the inferred causal relationships between the two simulated traits y_1 and y_2, for two different true causal effects β of y_1 into y_2, with different p-value thresholds on the x-axis. We compare the performance if MVIVW supplied with the true markers (oracle), parental markers of traits in the CI-GWAS inferred skeleton or the baseline MR procedure of clumped results of linear models at a p-value threshold for association of 10^{-8} (glm + clump). Gray dotted lines indicate FPR=0.05. A detailed description of all calculations and MR methods used is given in the Methods section.

We find that CI-GWAS selects markers with higher true positive rate (TPR) and higher true discovery rate (TDR) than mvSuSiE across a wide range of p-value thresholds (Figure 1a left panel, where markers are considered as true positives if causal for either A or B). For identification of the correct trait-specific causal variants, CI-GWAS selects the target trait with higher TPR and TDR than mvSuSiE, and does so

almost perfectly in the scenario with $\gamma = 0$ (Figure 1a right panel). With pleiotropy, the edge-TPR is slightly decreased for CI-GWAS, while the edge-TDR stays constant, which is caused by the correct SNP being identified, but a failure to attribute the effect to both traits as opposed to just one, effectively underestimating pleiotropy, but still with very much improved performance over mvSuSiE. In contrast, in a scenario with low linkage-disequilibrium (LD), where we removed highly correlated variants with LD $R^2 > 0.8$ and true causal variants are all present in the data, then both CI-GWAS and mvSuSiE show no difference in performance (Figure S2). Thus, our findings highlight the utility of CI-GWAS to find causal variants for specific conditions with precision and to separate pleiotropy from linkage in scenarios of realistic LD.

CI-GWAS instrumental variables reduce false positives in MR

Next, we demonstrate that our selection of appropriate IVs is key to robust MR, using results from the same simulation described above. We ran MVIVW with three different sets of IVs: (i) the parent markers selected with CI-GWAS; (ii) the true causal markers selected for simulating the traits (oracle); and (iii) markers with marginally significant effects at the genome wide level ($\alpha < 10^{-8}$), identified in a GWAS (glm + clump). We find that in the absence of a true causal effect between A and B, the approach with GWAS selected IVs tends to find spurious associations at a frequency that increases with the p-value threshold, while MVIVW run with oracle or CI-GWAS selected markers returns almost no false positive associations irrespective of the cutoff (Figure 1b). Only at a large cutoff of $\alpha < 10^{-1}$, do both the oracle and CI-GWAS results show a few false positives. In the presence of a true causal effect between A and B, both the CI-GWAS, as well as the oracle marker sets, result in causal link FPRs below or at 0.05 up to a p-value threshold of $\alpha < 10^{-2}$. In complete contrast, the GWAS selected IVs give a much higher FPR at around 0.75 even at low cutoffs, indicating that links are found in both directions. In contrast to the difference in false positives, the true relationship $A \rightarrow B$ was identified with all three IV sets, irrespective of the p-threshold. In summary, this simple simulation scenario shows that using marker IVs from CI-GWAS within an MR framework like MVIVW gives robust results, in stark contrast to standard GWAS selected markers.

In the following, we show that CI-GWAS offers more robust causal inference in more complex scenarios with many correlated traits and latent confounding, when compared to many popular MR methods. We conduct a simulation of a structural equation model consisting of twelve continuous traits, two of which are latent confounders and are not used in the inference, each with a set of 50 causal markers chosen at random from 169,904 imputed genetic markers recorded on 458,747 UK Biobank individuals. We simulated marker-trait and trait-trait adjacencies, causal effects, and trait data, but we do not consider the graph among genotypes, rather we focus on the recovery of the randomly drawn underlying causal variants in the presence real-data LD, and the estimation of the causal effects among traits. We compare CI-GWAS performance to commonly used MR methods: IVW implemented in the Mendelian-Randomization R package [27], MR-PRESSO [8], their multivariable extensions [9], CAUSE [11] and MR-BMA [10].

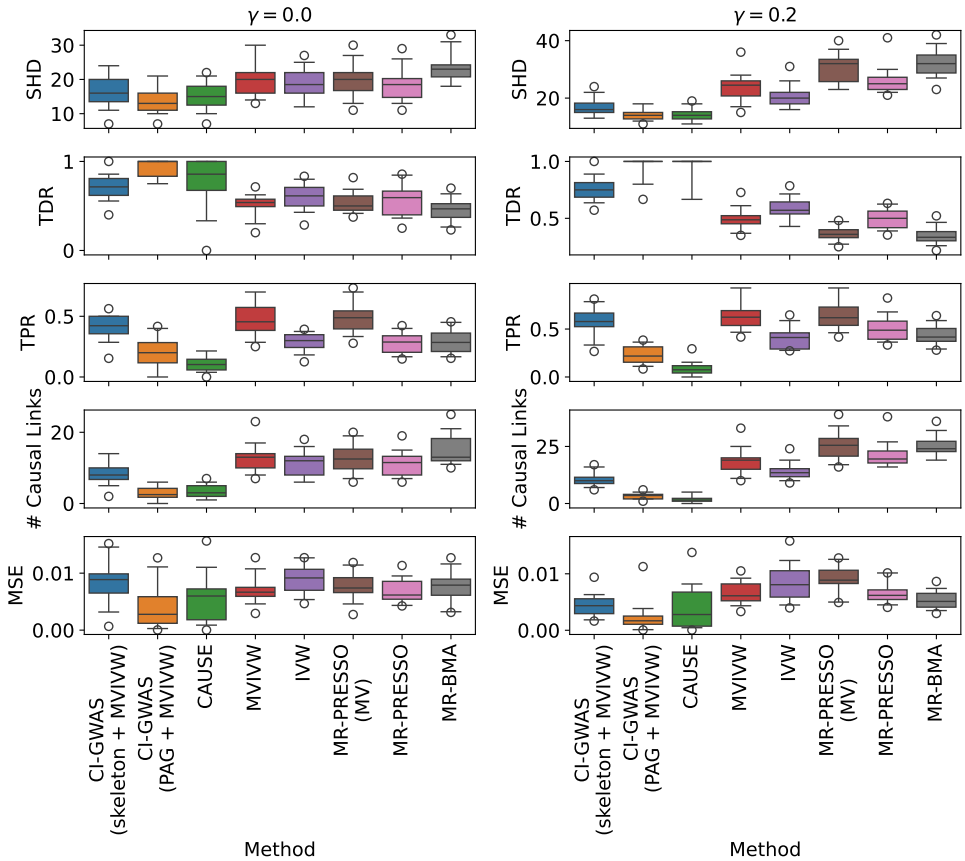

Fig. 2: **Causal inference performance of CI-GWAS and other MR methods in simulation with strong LD and ten strongly correlated traits.** We run CI-GWAS in a two of different ways: (skeleton + MVIVW): links are accepted if significant in MVIVW and an edge exists in the CI-GWAS inferred skeleton. (PAG + MVIVW): links are accepted if significant in MVIVW and an unambigously directed edge exists in the sRFCI inferred PAG. SHD: number of edge insertions, deletions or flips needed to transform the inferred graph among traits to the true graph; TDR: True discovery rate of trait-trait causal relationships; TPR: True positive rate of trait-trait causal relationships; # Causal Links: Number of trait-trait causal relationships inferred; MSE: mean squared error of inferred effect sizes of inferred causal links. We applied a p-value threshold of 0.05 to select significant causal links. Error bars indicate the 5% and 95% percentiles across 40 replicates. γ indicates the probability with which any additional affected trait is assigned to a marker after the first one and indicates the strength of pleiotropy. A detailed description of all calculations and MR methods used is given in the Methods section.

We find that all commonly applied MR methods show true discovery rates where around half the reported findings represent false positives, with the exception of CAUSE, which has a controlled false discovery rate in most simulated scenarios (Figure 2). CI-GWAS (skeleton + MVIVW) shows a similar TPR, but elevated TDR compared to the other MR methods. CI-GWAS (PAG + MVIVW) performs very similarly to CAUSE, with slightly higher TPR. While the MSE of the estimated effect sizes is comparable among all methods in the absence of pleiotropy, in the presence of pleiotropy CI-GWAS (PAG + MVIVW) yields more reliable effect size estimates than most MR methods.

As we simulated most marker effects to be small, limiting the number of discovered SNP-trait associations, we also explored a higher-power simulation setting where marker-trait effects were increased by a factor of 10 (see Methods). We find that the advantages of CI-GWAS over standard MVIVW and IVW becomes even more apparent. In addition, we see that the TPR-TDR tradeoff is broken by CI-GWAS (skeleton + MVIVW) in the absence of pleiotropy, where it yields very high TPR and TDR. However, in the presence of very strong pleiotropy, latent confounding and large effect sizes, CI-GWAS (PAG + MVIVW) recovers the latent confounding with many ambiguously directed edges, but finds no evidence to separate causation from confounding and pleiotropy and returns no positive findings (Figure S3). This shows that the IV filters and MR assumption checks that are facilitated by inferring a full graphical skeleton, and the orientation procedure of sRFCI, yield a conservative approach with improved precision over many MR methods. We highlight that CAUSE only estimates causal paths (i.e. there is no distinction of direct or indirect links) and requires greater run time than CI-GWAS. In comparison, with improved precision performance and faster run time, CI-GWAS uniquely distinguishes which risk factors have direct effects versus those that have indirect effects on each outcome, whilst also distinguishing markers that have pleiotropic effects from those that are trait-specific, controlling for LD.

Application to the UK Biobank

We applied CI-GWAS to 17 traits and 8.4 million single nucleotide polymorphism (SNP) markers in UK Biobank data, as described in the Methods and Table S1. We find that applying a standard MR approach where IVs are selected from standard GWAS summary statistics and used within MVIVW results in very many statistically causal relationships being discovered everywhere, most of them in both directions (Figure 3a). In contrast, when we use parental markers in the CI-GWAS inferred skeleton as IVs and accept only significant relationships for which an edge exists in the skeleton, we find only few causal links, most of which are also present in the base-MVIVW results (CI-GWAS skeleton + MVIVW in Figure 3b).

We also explore learning the causal graph within in our sRFCI approach that allows for latent confounding. We find that many relationships that are significant in MVIVW are assigned a bidirected edge, which is commonly defined as the absence of a causal relationship and the presence of a latent confounder. This implies that results obtained by standard MVIVW are not robust to the possibility of latent confounding (Figure 3c). From sRFCI, we find very few unambigously directed relationships, some of which replicate known potential biological relationships (*e.g.* SMK → CAD, ALC

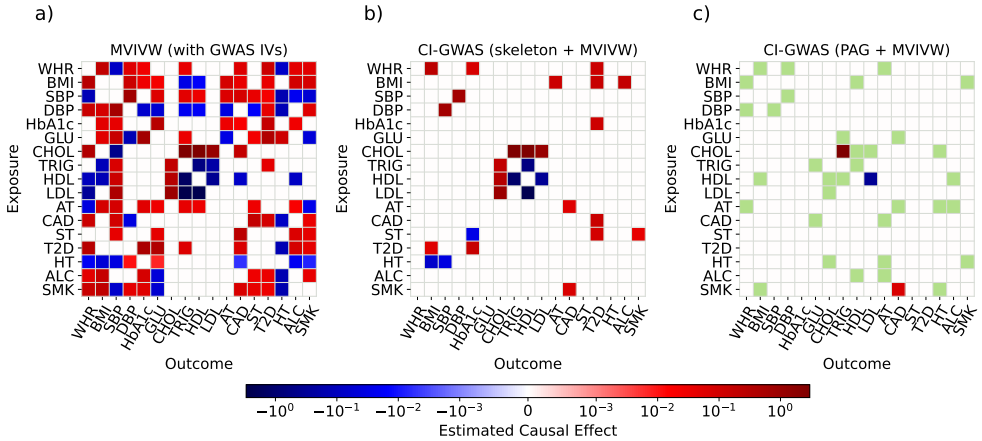

Fig. 3: **Causal inference on 17 traits in the UK Biobank.** (a): Markers selected from PLINK linear models at a genome-wide significance threshold of $p < 5 \times 10^{-8}$ and clumped with the clump approach of PLINK, then supplied to MVIVW as instrumental variables (IVs). (b): Parent-markers of traits in the CI-GWAS inferred skeleton are selected as IVs and supplied to MVIVW. Causal relationships are accepted when significant in MVIVW and variables are adjacent in the skeleton. (c): MVIVW run as in (b), but causal relationships accepted when significant in MVIVW and invariantly directed in the RFCI inferred PAG. Green squares indicate relationships where a bidirected edge was inferred, indicating the absence of a causal relationship and a dependency induced through confounding. We use a p-value threshold of 10^{-2} for asserting significance of MVIVW results. The skeleton inference in CI-GWAS was run at a significance threshold for conditional independence tests of 10^{-3}.

→ CAD, SMK → AT, ALC → BMI), while two are likely erroneous (*e.g.* SBP → HT and LDL → HT, also present in the base-MVIVW results) reflecting either more complex underlying mechanisms creating the appearance of a causal link in the data, or missing information from the inferred sRFCI PAG (we show the inferred PAG in Figure S4).

When we conservatively filter the exposure-outcome relationships using the PAG, as well as the IVs used to test them, according to MR assumptions, we find that only three consistent relationships remain: SMK → CAD, CHOL → TRIG, and HDL → LDL (Figure 3c). Thus, our results imply that it is premature to assign any but three relationships as potentially causal within the UK Biobank data and that many significant results may be explained by violation of MR assumptions.

The parental markers discovered in our skeleton represent joint marker-trait associations that are conditional on other traits and markers, with the conditioning on other markers controlling for LD (as we have shown above, essentially they can be viewed as multi-trait genome-wide "fine-mapping" results). We annotate these 998 significant

causal marker discoveries in the UK Biobank and find that almost all are protein coding variants (Table S2 and Supplementary Data Table 1 and 2 for full results). Below, we highlight a number of key findings.

For stroke, we fine-map 250 variants, 212 of which are associated with stroke and no other phenotype, including protein coding variants in *C17orf99*, circadian gene *PER1*, and *ABCC1* (a stroke drug target) which replicate known biology. We find that rs185353598, an exonic variant in *CTU2*, is pleiotropic with HbA1c and rs145802670 in *PGAP2* which is pleiotropic with T2D, both of which are novel. Circadian gene *RGL3* SNP rs167479 is pleiotropic with both systolic and diastolic blood pressure and cardiovascular disease.

For asthma, we find 24 associated markers, 9 of which are trait-specific links including in *TNFRSF8* and asthma GWAS genes also with known roles in inflammation, such as STAT6 [28]. The 24 SNPs are in genes that are predominantly in interleukin signalling pathways (p=$4.22e^{-5}$ PANTHER [29] database and p=$1.22e^{-5}$ and p=$1.18e^{-3}$ in the Reactome [30] database). Levels of serum interleukin-7 were significantly elevated in patients with allergic asthma and interleukin 5 is a major cytokine responsible for eosinophil proliferation, migration and degranulation, which play a role in the manifestation of type 2 asthma.

For type-2 diabetes (T2D), we fine-map 133 variants, and identify *CCND2* (SNP rs76895963 and rs3217792 whose rare alleles are associated with reduced susceptibility in GWAS studies [31]), which is a provisional therapeutic target for beta cell proliferation and is pleiotropic for HbA1c levels in blood. Additionally, we fine-map 69 trait-specific T2D variants independent of BMI, waist-hip ratio, blood glucose levels and all other included traits, all of which have been implicated in insulin resistance and/or T2D GWASs, except *OTOP1* which has only been implicated in mouse models. We find that lincRNA AC105393.1 variant rs62106258 is pleiotropic for BMI, smoking, and T2D. Generally, the genes found are enriched for RNA metabolic process (p=$5.07e^{-6}$ in the PANTHER database) and regulation of gene expression in late stage pancreatic bud precursor cells and regulation of beta-cell development (p=$3.48e^{-5}$ and p=$3.18e^{-3}$ in the Reactome database). For HbA1c, the majority of our findings are either pleiotropic with T2D or are SNPs within the same genes of *JAZF1*, *ANK1*, *ANKH*, *CCND2*, *HK1*, *KCNQ1*, *PAM*, which are enriched in carbohydrate phosphorylation (p=$1.30e^{-5}$ in the PANTHER database) and glycolysis and glucose metabolism (p=$6.72e^{-4}$ and p=$1.23e^{-3}$ in the Reactome database).

For height, we fine-map 268 loci, of which 173 loci are specific to height, the majority of which are protein coding and enriched in steroid hormone processes, growth plate chondrocytes (p=$2.95e^{-2}$ and p=$3.42e^{-3}$ in the PANTHER database) and fibroblast growth factor receptor 3 (FGFR3) pathways (p=$3.07e^{-4}$ in the Reactome database). However for BMI, we find only 11 loci, 5 of which are specifically associated with BMI, which are enriched in RUNX1 and Wnt signaling and obesity (p=$2.58e^{-5}$ in the Reactome database) and extracellular matrix (ECM) remodelling (p=$2.10e^{-5}$ in the PANTHER database).

For lipids, we fine-map 16 parental markers for triglyceride levels supporting known lipid pathways including exonic SNPs in *APOE*, *ApoA-V*, lipoprotein lipase, Delta-6 desaturase (*Fads2*), as well as the general *APOA4-APOA5-ZNF259-BUD13* gene

cluster on chromosome 11. The genes for LDL and triglyceride levels are enriched in lipoprotein particle clearance (p=$2.97e^{-6}$ in the PANTHER database) and plasma lipoprotein remodelling (p=$6.14e^{-6}$ in the Reactome database) among many others related to lipase, cholesterol and sterol activity and transport. For LDL we fine-map 14 protein coding loci and for HDL we fine-map 20 loci with similar enrichment and the addition of NR1H2 and NR1H3-mediated signaling (p=$1e^{-5}$ in the Reactome database).

These results are only a subset of those presented in Supplementary Data Tables 1 and 2 and demonstrate that our approach can aid in identifying shared and trait-specific sets of genes. Notable additional examples of pleiotropy are a coding variant in alcohol dehydrogenase 1B (ADH1B) which catalyzes the oxidation of alcohol to acetaldehyde and is pleiotropic for alcohol consumption, LDL levels, systolic blood pressure and stroke. Similarly, we find an exonic variant in *DAG1* that is pleiotropic for HbA1c, smoking, CAD and BMI. rs28601761 in lincRNA *RP11-136O12.2* is pleiotropic for cardiovascular disease, HbA1c and blood triglyceride levels. *SNRNP70* SNP rs73046792 is pleiotropic for smoking and systolic blood pressure, as is rs13107325 in *SLC39A3* which links stroke, asthma, diastolic blood pressure and smoking. *snoU13* SNP rs139263177 links smoking to stroke and *snoZ6* links smoking to blood lipid levels. A variant in *FAM101A* is pleiotropic for stroke, T2D and waist-hip ratio. SNP rs13107325 in *SLC39A8* is pleiotropic for diastolic blood pressure, HDL levels and asthma. Finally, exonic variants in *APOE* are pleiotropic for blood cholesterol, HDL, LDL and trigylercide levels, waist-hip ratio and smoking, supporting the link between *APOE* and BMI acting through lipid transport into cells. Taken together, our graphical inference approach fine-maps associations to predominantly exonic variants, that are unique and shared across traits, with clear pathways to each outcome, and represents a follow-up step to support association study testing.

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, orientate the edges and then estimate 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 [18, 19, 32]. 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: *(i)* fine-mapped associations that separate pleiotropy from linkage; *(ii)* appropriate selection of IVs that fulfil MR assumptions, *(iii)* 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, however they all build upon each other, assume a similar fixed directed acyclic graph and have very high type-I error. Here, we demonstrate this empirically by showing that standard MVMR results obtained in the UK Biobank generally disappear with selection of IVs from 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. Multivariable MR [8, 33, 34] has been proposed for this problem, but here we show that it is ineffective without being paired with graphical inference. Structural equation models have also been suggested [35], but these require a correct causal order of variables to be known (or learnt) 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 settings [36]. 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 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: (i) individual-level data, or (ii) 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 are not always publicly available, but could easily be shared, inferred across biobanks, or approximated. Data sets have 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 inference [37]. Additionally, there is also a general tendency in 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 utilising longitudinal records, where discovered adjacencies in cuda-skeleton can then be orientated longitudinally and combined with therapeutic information, statistical causal effects may then be more effectively examined as there is certainty about the directions of many time ordered edges. 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 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. In this way, more can be learned about the likely underlying nature of phenotypic and genetic correlations within the human population.

Methods

CI-GWAS

CI-GWAS takes either individual or summary-level data, learns conditional dependencies or links between genetic variants and traits, determines whether conditional dependencies among traits can be attributable to a directed causal relationship, and then estimates the magnitude and direction of the potential causal effect. There are three parts to our approach: i) 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); ii) an approach called sRFCI to direct the skeleton edges and thereby infer causal relationships, allowing for arbitrarily many latent variables; and iii) estimation of the magnitude of the potential causal relationships among the discovered exposures and outcomes with multi-variable inverse weighted Mendelian randomisation (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 only report exposure-outcome effect estimates 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, alongside algorithms for their implementation.

In this work, we calculate correlation matrices M^G, 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 algorithm in parallel over LD-blocks, each of which has a distinct M^G matrix. Thus, we first block the genomic input data with parameter set as shown in Table 1. Then we prepare the correlation matrices as follows:

- M^G: 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 rsIDs of the markers in the block with the `--from` and `--to` options [38, 39].
- M^{GT}: we compute the marker-trait correlations via 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` options [40, 41]. We then compute the Pearson correlation via $\rho_{ij} = \beta_{ij}\sqrt{2 * f_i * (1 - f_i)}$, 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) are 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 another publicly available samples, if the original sample matrices are unavailable [42] can all be accommodated within our software.

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 [25]. For this we use polyserial correlations between continuous variables and binary variables, and polychoric correlations between binary variables, and Pearson correlations between continuous variables. For the calculation of polyserial and polychoric correlations we use the 'hetcor' function in the 'polycor' R package [43].

We provide open source code and scripts to run our full analysis (see Code Availability).

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

$C^b \leftarrow$ make banded LD matrix from C given b
$S_i \leftarrow \sum_{j=i+1}^{n} C^b_{ij}$
$h \leftarrow n$
$l \leftarrow 3$
$w \leftarrow (h+l)/2$
$s \leftarrow n$
while $(|s - s_{max}| > \epsilon) \vee (s > s_{max})$ **do**
    if $s > s_{max}$ **then**
        $h \leftarrow min(h, w)$
    else
        $l \leftarrow max(l, w)$
    end if
    $w' \leftarrow (h+l)/2$
    if $w' == w$ **then** break
    end if
    $w \leftarrow w'$
    $S' \leftarrow$ Hanning-smoothen S with window w
    $M \leftarrow$ positions of local minima in S'
    $s \leftarrow max(\{M_{i+1} - M_i | 1 \leq i \leq (n-1)\})$
end while
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 algorithm [20], a GPU accelerated implementation of the PC-stable algorithm [19].

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.*

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

is then compared to a threshold τ, given by

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

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 runtime 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 1.

The runtime 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 [44] 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 part of the cuPC algorithm to each block of markers together with 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 Figure S5.

Having inferred a set of skeletons with adjacency matrices $A^B = A_1, A_2, ..., A_n$ and separation sets $S^B = S_1, S_2, ..., S_n$ blockwise, we first create a merged skeleton with adjacency matrix A^* following four rules: (i) 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; (ii) if two markers are in different blocks, they are not adjacent in A^*; (iii) 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 (iv) 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 which are most likely to have direct causal effects on traits. In order to prevent fine-mapping errors due to 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.

Table 1: 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	$2 * 10^3$

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: (i) inference of a partial ancestral graph (PAG) by a modified RFCI [18] algorithm; and (ii) multi-variable inverse variance weighted Mendelian randomization (MVIVW) [9] 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 algorithm [18] 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, which cause traits to be missing from the separation sets of markers and other traits, which then gives incorrect edge orientation. To solve this issue, we present an order-independent algorithm that uses the RFCI procedure to orientate edges between adjacent variable, 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: (i) S^{max}, a maximal separation set, meaning that the addition of any other variable results in $Z > \tau$, and (ii) 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, and thus everything that is not within S^{max} can be considered a collider between the variables. This separation set is used within sRFCI to orientate 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: (i) V_i and V_k are connected; (ii) V_k and V_j are connected; (iii) V_i and V_j are not connected. We only want to orientate unshielded triples, when V_k is not an element of S_{ij}^{max}, and thus is an unlikely collider. The set S^{mpc} is more conservative with regards 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 SNP-SNP correlations, marginal marker-phenotype correlations and phenotype-phenotype correlations. We made S order independent and we attempted to overcome missing non-collider variables, by maximising separation sets as much as possible. We note that the CPC/MPC versions of RFCI [19] provide a solution for the same issue by maximally searching all potential separation sets, but this unfortunately 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 that are used for orientation to *relevant unshielded triples*, which we define as those that contain at most one marker, and where 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 among all possible sets. 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^*.

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 [18].

Step 2. Orientate the v-structures using S^{max} and Algorithm 4.4 of [18].

Algorithm 2 Separation set selection

Require: α, n, A^*, C

$ut \leftarrow$ relevant unshielded triples of A^*
$op \leftarrow$ outer variables in ut
for V_i, V_j in op **do**
  $N \leftarrow$ neighbors(V_i, A^*)
  $N' \leftarrow N$
  $sep \leftarrow \{\}$
  $foundSepset \leftarrow$ false
  $foundMinimum \leftarrow$ false
  for l in $1 \ldots |N|$ **do**
    $newElem \leftarrow \{\}$
    $z_{ref} \leftarrow \infty$
    for V_k in N' **do**
      $S \leftarrow sep \cup V_k$
      $z \leftarrow Z(\hat{\rho}(V_i, V_j|S))$
      if $z \leq z_{ref}$ **then**
        $z_{ref} \leftarrow z$
        $newElem \leftarrow V_k$
      end if
    end for
    if $z_{ref} > z_{ref,l-1} \wedge foundSepset \wedge \neg foundMinimum$ **then**
      $foundMinimum \leftarrow$ true
      $S^{mpc}_{i,j} \leftarrow sep$
    end if
    if $z_{ref} > \tau(\alpha, n, l) \wedge foundSepset$ **then**
      break
    end if
    if $z_{ref} \leq \tau(\alpha, n, l)$ **then**
      $foundSepset \leftarrow$ true
    end if
    $sep \leftarrow sep \cup newElem$
    $N' \leftarrow N' \setminus \{newElem\}$
  end for
  $S^{max}_{i,j} \leftarrow sep$
end for
return M

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 then contribute to the orientation in later stages, although they themselves can later be orientated 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 [18]. This step involves repeated applications of orientation rules (R1) to (R10) of [45]. We implemented the suggested modifications from [19] 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 orientated 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 the set of possible Maximal Ancestral Graphs (MAG, see Ref. [18] 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 confounders [18], which has been shown to be of importance for Mendelian Randomisation methods applied to biobank data [6, 11, 46–48]. A situation can arise when neither variable is ancestral to the other, which results in the adjacency being inferred as $\longleftrightarrow$. This implies that neither of the two variables is ancestral to the other, with no direct causal relationship inferred, implying the existence of a latent confounder.

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.

Part 3: Learning the magnitude of the causal effects

Having learnt the conditional dependencies in the data with cuda-skeleton and tested for the presence of latent confounding among traits with sRFCI we then propose to use the framework of MVIVW to asses the significance and magnitude of causal effects.

In order to identify potential causal relationships in an MR framework like MVIVW, sets of IVs and exposures are needed for each outcome. Specifically, MVIVW assumes that IVs are 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, 49]. MVIVW can be employed using the fine-mapped SNPs inferred from cuda-skeleton because the conditional-independence tests conducted in *cuda-skeleton* explicitly check for several common MR assumptions on IVs (see [49] 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), since

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 multivariate 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. Principled instrumental selection of this kind stands contrast to most other current MR approaches, where IVs are selected by taking variants with the highest marginal association test statistics. A similar approach, relying 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 work [23]. Here, we advance by utilizing this idea in a framework where no markers need to be pre-selected, many variables can be conditioned on at once, and pleiotropic markers can be used for inference, 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 i) are selected as significant in cuda-skeleton or ii) are directed in the same way as inferred by MVIVW in the sRFCI inferred PAG. This gives sets of potential causal links with evidence for a direction that have been learned from the data.

We use the selected instrumental variable and exposure sets as input to the MVIVW implementation in the *MendelianRandomization* R package [50], with the `robust` option. We select significant causal relationships by applying a p-value threshold of 0.05 to the output.

Application of CI-GWAS to the 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 data[51], we selected six blood based biomarkers, four of the most common heritable complex diseases, and seven quantitative measures. The full list of the 17 traits, the UK Biobank coding of the data used, and the covariates adjusted for are given in Table S1. 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 mean 0 and variance 1.

For the common complex diseases, we determined disease status using a combination of information available. For asthma (AT), we used self-report asthma diagnosed by a doctor (3786-0.0), date of asthma report (42014-0.0), and self-report recent medication for asthma (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 (42000-0.0). For stroke (ST), we used self-report stroke diagnosed by doctor (6150-0.0), the age 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 maximise the number of cases available for analysis. Individuals with neither a self-report indication or a relevant ICD10 diagnosis, were then assigned a zero value as a control.

We first restricted our analysis to a sample of European-ancestry UK Biobank individuals. 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 2,504 individuals of the 1,000 Genomes project. Using the obtained PC loadings, we then assigned each participant to the closest 1,000 Genomes project population, selecting individuals with PC1 projection $\leq$ absolute value 4 and PC2 projection $\leq$ absolute value 3. We applied this ancestry restriction as we wished to provide the first application of our approach, and to replicate our results, within a sample that was 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 (i) extreme heterozygosity and missing genotype outliers; (ii) a genetically inferred gender that did not match the self-reported gender; (iii) putative sex chromosome aneuploidy; (iv) exclusion from kinship inference; (v) 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 the 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 data set 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).

Simple scenario

We set up a structural equation model (SEM) of two continuous trait variables A and B with 200 causal markers X_A, X_B each, chosen randomly from the 654,172 markers on chromosome 1 in the 8,430,446 marker UK Biobank data set. To model pleiotropy, we allow allow an intersection of X_A and X_B of size $200 \times \gamma$, where $\gamma \in \{0.0, 0.5\}$. We simulate two scenarios: one where there is no causal effect between traits A and B ($\beta = 0$) and one where $\beta = 0.25$. The effects of the markers in X_A, X_B on the traits are sampled such that the heritability of each trait is equal to 0.2. We repeat this SEM generation for 20 distinct replicates for each combination of values of γ, β. Given a sampled SEM and the marker data from UKB we then sample values for traits A and B for 458,747 individuals.

We apply CI-GWAS to these data as described above. To run mvSuSiE [26], we employed its summary data interface, which integrates z-scores from association tests across multiple SNPs. For setting up the prior in this context, we configured the parameter w = c(0.2, 0.2, rep(0.6/5, 5)). We calculate TPR and TDR of the selected markers. We distinguish snp-TPR/TDR and edge-TPR/TDR: the first regards a marker as true positive if it is in the set of causal markers for any of the simulated traits. The latter uses only true positive markers categorizes each edge / association as true or false positive, depending on whether a marker has a direct effect on the respective trait in the SEM. As mvSuSiE returns credible sets that typically contain multiple markers, we compute these statistics for it by counting a credible set as true positive if it contains at least one true causal marker, and false positive if it does not contain at least one true causal marker.

We also compare the TDR of the causal relationships among the two traits between MVIVW [9] (from the MendelianRandomization R package (version 0.4.2) [27]) when run with GWAS selected markers (as described above), CI-GWAS selected markers (as described above) or oracle markers, *i.e.* the true causal markers from the SEM. We apply range of p-value thresholds (10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}) to select the positive test outcomes from MVIVW.

For all metrics calculated for this simulation we present the bootstrapped mean and its 95% confidence interval, calculated in 1000 draws with replacement from the 20 replicates.

Complex scenario

To simulate a more complex setting, we follow a similar strategy as described above for the simple setting. For twelve continuous traits, we select 50 causal markers each using imputed SNP data from chromosome 1. We selected causal markers that were not highly correlated by removing variants with LD $R^2 \leq 0.8$ within a 1-Mb to leave a set of 169,904 markers from which the causal variants were then drawn at random. We then simulated a marker-trait and trait-trait adjacency matrix G, causal effects A, and phenotypic data as described below.

First, we simulate a randomly directed acyclic graph $\mathcal{G}$ 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 $\mathbf{G}$ of dimensionality $d \times d$, where $d = m + l + p$. Set all elements of $\mathbf{G}$ to zero. Here we chose $l = 2;\ p = 10$.

Step 2. Generate random adjacencies by setting values of $\mathbf{G}$ to 1 with probability $\frac{deg}{d}$. Here we chose 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 $\mathbf{G}$ accordingly. Here, we selected $probPleio = 0.2$ 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 $\mathbf{A}$ as a true causal effect matrix. Here, we used 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 allowed us to simulate a wide range of causal effects, and to have separate strengths for marker-trait and trait-trait relationships, which reflected the pattern we observed in the UK Biobank data.

Step 5. Simulate data based on a structural equation model. For each variable V_j in $\mathbf{G}$, check if V_j has parents ($Pa(V_j)$) in $\mathbf{G}$. 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:

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

where b_0 is the intercept (set to zero) and coefficients for the regression model are the appropriate b_i values selected from $\mathbf{A}$ for the ith parent of V_j. The sampling error ϵ_x is then sampled from a standard normal with variance $1 - var(\sum_i b_i Pa_i(V_j))$. 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 analysed using CI-GWAS. For the MR analysis comparison, we used three commonly employed methods: IVW, MR-PRESSO, and CAUSE and their multivariable extensions where appropriate. IVW were implemented using the MendelianRandomization R package (version 0.4.2) [27]. 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. We employed MR-BMA [10] using code from the publications repository, following the README (code from https://github.com/verena-zuber/demo_AMD). We ran `summarymvMR_SSS` with settings `kmin=1, kmax=9, prior_prob=0.3, max_iter=10000, sigma=0.25` and `create.permutations` with `nrepeat=1000`. Unless stated we used default options given by each software. We applied a significance threshold of $p < 0.05$ to the trait-trait causal relations.

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 2.0 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` 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, the structural-hamming-distance (SHD) between the inferred causal graph and the true causal graph (where we count the number of operations needed to convert the inferred directed adjacency matrix into the true one), the number of discovered causal links and the mean squared error (MSE) of the causal effect estimates. To calculate TPR, TDR and SHD, 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 = 1 - (FP / (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.

To calculate the MSE we exclude all missing or 0 entries in the inferred effect matrix from the calculation. Thus the presented MSE represents the error inferred non-zero estimates, because both CI-GWAS and MR methods miss many causal relationships, which gives large weight to zero entries. Additionally, we have already calculated TDR and TPR for adjacency recovery above, and we wish to assess the error of positive discoveries. Since CAUSE, IVW and PRESSO estimate path effects instead of direct effects, for those methods we compare the estimated effect to the total effect in the true DAG over all path lengths. 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.

Complex scenario with very strong marker-trait effects

We repeated the simulation described for the complex scenario with enhanced marker-trait effect sizes, drawn from a uniform distribution $U[(-0.1, -0.001) \cup (0.001, 0.1)]$ in Step 4 of the simulation procedure. Due to the strongly increased number of significant markers selected from a standard GWAS conducted on these data (in the same way as described above), we were not able to obtain inference results from MR-PRESSO, multivariate MR-PRESSO, nor MR-BMA because these cannot accommodate larger numbers of IVs as either the p-values are never obtained, models continue to run over 10-day periods with no indication of completion time, and there is no way to parallelize the approaches.

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 (Figure S1). We conducted two small simulations: (i) independently drawn samples of $x \sim Binom(2, 2 \times 10^{-3})$ and $y \sim Binom(1, 0.028)$. We then tested $CI(x, y)$ as described above with $\alpha = 0.05$. (ii) we simulated from the DAG: $x \rightarrow y_{cont} \rightarrow y_{bin}$ where $x \sim Binom(2, 2 \times 10^{-3})$, $y_{cont} \sim \mathcal{N}(\beta x, 1)$, $y^l_{bin} \sim \mathcal{N}(\gamma y_{cont}, 1)$, and y_{bin} is 1 wherever $y^l_{bin} > \Phi^{-1}(K)$ 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 variables, in 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 package [43].

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 Ensembl [52], 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 BEDtools [53] version 2.30.0 with

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

We used the identified genes for each outcome and tested them for enrichment in the Reactome [30] (https://reactome.org) and Panther [29, 54, 55] databases (https://pantherdb.org).

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-centre Research Ethics Committee (MREC). Methods were carried out in accordance with the relevant guidelines and regulations, with informed consent obtained from all participants.

Supplementary information. Supplementary Information accompanies this work.

Acknowledgments. We thank Zoltan Kutalik and members of the Robinson group at ISTA for their comments, which improved this manuscript. 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. 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).

Declarations

Author contributions

MRR, MM and NM conceived and designed the study, conducted the analysis and wrote the paper. MJB and MRR provided study oversight. IK contributed to study

design and analysis. MB contributed to the analyses. All authors approved the final manuscript prior to submission.

Author 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.

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 summary statistic estimates are released publicly on Dryad: https://doi.org/xx.xxxx/dryad.xxxxxxxxx.

Code availability

The CI-GWAS code is fully open source and available at https://github.com/medical-genomics-group/ci-gwas. 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

[1] Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research* **26**, 2333–2355 (2017).

[2] E, S. Multivariable mendelian randomization and mediation. *Cold Spring Harb Perspect Med* **11** (2021).

[3] Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature Communications* **9**, 1–12 (2018).

[4] Hemani, G. *et al.* The mr-base platform supports systematic causal inference across the human phenome. *elife* **7**, e34408 (2018).

[5] Pingault, J.-B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19**, 566–580 (2018).

[6] Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**, R89–R98 (2014).

[7] O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics* **50**, 1728–1734 (2018).

[8] Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693–698 (2018).

[9] Burgess, S. & Thompson, S. G. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology* **181**, 251–260 (2015).

[10] Zuber, C. J. K. C. B. S., V. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nat Commun* **11** (2020).

[11] 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. *Nature Genetics* **52**, 740–747 (2020).

[12] Bowden J, B. S., Davey Smith G. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *M. Int J Epidemiol* **44** (2015).

[13] Burgess S, T. S., Butterworth A. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37** (2013).

[14] Zuber V, L. M. H. A. B.-A. S. E. C. D. S. B. S. G. D. B. L., Lewin A. Multi-response mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. *Am J Hum Genet* **110** (2023).

[15] Lin Z, P. W., Xue H. Combining mendelian randomization and network deconvolution for inference of causal networks with gwas summary data. *PLoS Genet* **19** (2023).

[16] Pearl, J. *Causality* (Cambridge university press, 2009).

[17] Spirtes, P. *et al. Causation, prediction, and search* (MIT press, 2000).

[18] Colombo, D., Maathuis, M. H., Kalisch, M., Richardson, T. S. *et al.* Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics* **40**, 294–321 (2012).

[19] Colombo, D. & Maathuis, M. H. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15**, 3741–3782 (2014).

[20] Zarebavani, B., Jafarinejad, F., Hashemi, M. & Salehkaleybar, S. cupc: Cuda-based parallel pc algorithm for causal structure learning on gpu. *IEEE Transactions on Parallel and Distributed Systems* **31**, 530–542 (2020). Conference Name: IEEE Transactions on Parallel and Distributed Systems.

[21] Neto EC, A. A. Y. B., Keller MP. 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).

[22] Mckeigue, P., Krohn, J., Storkey, A. J. & Agakov, F. Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. & Culotta, A. (eds) *Sparse instrumental variables (spiv) for genome-wide studies.* (eds Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. & Culotta, A.) *Advances in Neural Information Processing Systems*, Vol. 23 (Curran Associates, Inc., 2010). URL https://proceedings.neurips.cc/paper_files/paper/2010/file/4476b929e30dd0c4e8bdbcc82c6ba23a-Paper.pdf.

[23] Amar, D., Sinnott-Armstrong, N., Ashley, E. A. & Rivas, M. A. Graphical analysis for phenome-wide causal discovery in genotyped population-scale biobanks. *Nature Communications* **12**, 350 (2021).

[24] Lauritzen, S. L. *Graphical Models* (Oxford University Press, 1996).

[25] Cui, G. P. . H. T. L., R. Learning causal structure from mixed data with missing values using gaussian copula models. *Stat Comput* **29**, 311–333 (2019).

[26] Wang, G., Zou, Y., Carbonetto, P. & Stephens, M. *mvsusieR: Multivariate Sum of Single Effects Regression* (2024). URL https://github.com/stephenslab/mvsusieR. R package version 0.1.7, commit 103eabadb3de827191915814526168b181a90218.

[27] Yavorska, O. O. & Burgess, S. Mendelianrandomization: an r package for performing mendelian randomization analyses using summarized data. *International journal of epidemiology* **46**, 1734–1739 (2017).

[28] Stikker, B. S., Hendriks, R. W. & Stadhouders, R. Decoding the genetic and epigenetic basis of asthma. *Allergy* **78**, 940–956 (2023). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/all.15666.

[29] Paul D. Thomas, A. M. T. M.-L.-P. A., Dustin Ebert & Mi, H. Panther: Making genome-scale phylogenetics accessible to all. *Protein Society* **31** (2022).

[30] Milacic M, C. P. G. C. G. M.-G. J. H. R. J. B. M. L. M. B. P. R. R. E. R. K. S. C. S. V. S. R. T. K. V. T. W. J. W. A. W. G. S. L. H. H. D. P., Beavers D. The reactome pathway knowledgebase 2024. *Nucleic Acids Research* (2024).

[31] Osonoi, S., Sasaki, T., Takeuchi, Y. & Mizukami, H. 1732-P: Correlation of Cyclin D2 with Cell Proliferation in Human Pancreatic Beta Cells and Possible Regulation by CCND2-AS1. *Diabetes* **72**, 1732–P (2023).

[32] Cheng, D. *et al.* Toward unique and unbiased causal effect estimation from data with hidden variables. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[33] 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. *International Journal of Epidemiology* **48**, 713–727 (2019).

[34] Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* **40**, 304–314 (2016).

[35] Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour* **3**, 513–525 (2019).

[36] Lagemann, K., Lagemann, C., Taschler, B. & Mukherjee, S. Deep learning of causal structures in high dimensions under data limitations. *Nature Machine Intelligence* (2023).

[37] Gkatzionis, A. & Burgess, S. Contextualizing selection bias in mendelian randomization: how bad is it likely to be? *Int. J. Epidemiol.* **48**, 691–701 (2019).

[38] Christopher C Chang, L. C. T. S. V. S. M. P. J. J. L., Carson C Chow. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience* **4** (2015).

[39] Purcell, S. & Chang, C. *PLINK [1.9]*. URL www.cog-genomics.org/plink/1.9/.

[40] Purcell, S. & Chang, C. *PLINK [2.0]*. URL www.cog-genomics.org/plink/2.0/.

[41] Purcell S, T.-B. K. T. L. F. M. B. D. M. J. S. P. d. B. P. D. M. S. P., Neale B. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* **81**, 559–575 (2007).

[42] Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics* **11**, 1561 – 1592 (2017).

[43] Fox, J. *polycor: Polychoric and Polyserial Correlations* (2022). URL https://CRAN.R-project.org/package=polycor. R package version 0.8-1.

[44] Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

[45] Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* **172**, 1873–1896 (2008).

[46] Davey Smith, G. & Ebrahim, S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease?

International Journal of Epidemiology **32**, 1–22 (2003).

[47] Boef, A. G., Dekkers, O. M. & Le Cessie, S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *International Journal of Epidemiology* **44**, 496–511 (2015).

[48] Zheng, J. *et al.* Recent developments in mendelian randomization studies. *Current Epidemiology Reports* **4**, 330–345 (2017).

[49] Hernán MA, R. J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology.* **4**, 360–372 (2006).

[50] Burgess, S. & Yavorska, O. *MendelianRandomization: Mendelian Randomization Package* (2024). URL https://CRAN.R-project.org/package=MendelianRandomization. R package version 0.10.0.

[51] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[52] Martin, F. J. & et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).

[53] Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 814–842 (2010).

[54] Huaiyu Mi, J. X. H. D. E. C. M. X. G., Anushya Muruganujan & Thomas, P. D. Protocol update for large-scale genome and gene function analysis with the panther classification system (v.14.0). *Nat Protoc* **14** (2019).

[55] Mi, H. & Thomas, P. Panther pathway: An ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol* **563** (2009).

Supplementary Information

Causal inference for multiple risk factors and diseases from summary statistics data

Nick Machnik, Mahdi Mahmoudi, Malgorzata Borczyk, Ilse Krätschmer, Markus J. Bauer, Matthew R. Robinson

Type	Phenotype	Code	Biobank codes	adjusted covariate codes	sample size
Biomarker	Cholesterol	CHOL	30690-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	395,025
Biomarker	Glucose	GLU	30740-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	360,010
Biomarker	Glycated haemoglobin	HbA1c	30750-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	394,912
Biomarker	HDL cholesterol	HDL	30760-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	360,286
Biomarker	LDL direct	LDL	30780-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	394,264
Biomarker	Triglycerides	TRIG	30870-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	394,264
Disease	Asthma	AT	3786-0.0, 42014-0.0, 22167-0.0, 42000-0.0, 6150-0.0, 40001-0.0, 41202-0.0, 41204-0.0, 41270-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	399,761 (93,096)
Disease	Cardiovascular disease	CAD	3894-0.0, 3627-0.0, 42000-0.0, 6150-0.0, 40001-0.0, 41202-0.0, 41204-0.0, 41270-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	412,915 (79,937)
Disease	Stroke	ST	4056-0.0, 42006-0.0, 6150-0.0, 40001-0.0, 41202-0.0, 41204-0.0, 41270-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	445,737 (47,044)
Disease	Type-2 diabetes	T2D	2443-0.0, 2976-0.0, 6153-0.0, 6177-0.0, 40001-0.0, 41201-0.0, 41202-0.0, 41204-0.0, 41270-0.0, 41262-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	432,102 (63,585)
Measures	Alcohol consumption	ALC	1558.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	414,055
Measures	Body mass index	BMI	21001-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	413,595
Measures	Diastolic blood pressure	DBP	4079-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	377,358
Measures	Smoking behaviour	SMK	20161-0.0, 20116-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	414,055
Measures	Standing height	HT	50-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	414,055
Measures	Systolic blood pressure	SBP	4080-0.1	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	377,347
Measures	Waist-hip ratio	WHR	48-0.0, 49-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	413,595

Table S1: **UK Biobank phenotypes used within the study.** Columns show the type of trait, the name, the trait code used for the figures, the UK Biobank codes used to construct the trait values, the covariates adjusted for within the analysis, and the sample size (case numbers given in brackets).

Traits	Number of parental markers
AT	9
HT	173
ST	212
ALC	1
BMI	5
CAD	2
DBP	8
HDL	9
LDL	6
SBP	5
SMK	29
T2D	69
WHR	2
CHOL	1
TRIG	16
HbA1c	18
AT, HT	2
AT, ST	2
HT, ST	10
ALC, HT	1
ALC, ST	1
AT, DBP	1
AT, SMK	1
AT, T2D	1
CAD, HT	4
DBP, HT	5
DBP, ST	1
GLU, HT	1
HDL, HT	1
HT, SBP	9
HT, SMK	18
HT, T2D	14
HT, WHR	3
SBP, ST	2
SMK, ST	5
ST, T2D	16
ALC, SMK	1
BMI, GLU	1
BMI, SBP	1
BMI, SMK	1
CAD, DBP	2
CAD, SMK	1
CAD, T2D	1
DBP, SBP	3

DBP, SMK	1
HDL, LDL	1
HDL, WHR	1
HT, TRIG	1
LDL, SMK	1
SBP, SMK	2
SMK, T2D	4
ST, TRIG	1
CHOL, LDL	1
HDL, TRIG	4
HT, HbA1c	5
HbA1c, ST	4
SMK, TRIG	2
CHOL, TRIG	1
DBP, HbA1c	1
HbA1c, SMK	2
HbA1c, T2D	8
AT, HT, T2D	1
HT, SBP, ST	1
HT, ST, T2D	2
AT, DBP, T2D	1
AT, SMK, T2D	1
DBP, HT, SMK	1
HT, SBP, SMK	2
HT, SMK, T2D	4
ST, T2D, WHR	1
ALC, HT, TRIG	1
AT, HT, HbA1c	1
BMI, SMK, T2D	1
CHOL, HT, LDL	1
DBP, HT, TRIG	1
HDL, SMK, WHR	1
AT, HbA1c, T2D	1
DBP, HT, HbA1c	2
GLU, HT, HbA1c	1
HDL, LDL, TRIG	1
HDL, SMK, TRIG	1
HT, HbA1c, T2D	3
CHOL, LDL, TRIG	1
HbA1c, SMK, T2D	2
CAD, HbA1c, TRIG	1
ALC, LDL, SBP, ST	1
AT, DBP, HDL, SMK	1
BMI, HT, SMK, WHR	1
CAD, DBP, SBP, ST	1

AT, HT, HbA1c, T2D	2
DBP, HT, HbA1c, T2D	1
CHOL, LDL, SMK, TRIG	1
BMI, CAD, HT, HbA1c, SMK	1

Table S2: Number of parental markers for each trait combination in CI-GWAS UK Biobank application. For any not listed traits / trait combinations no parental markers were found.

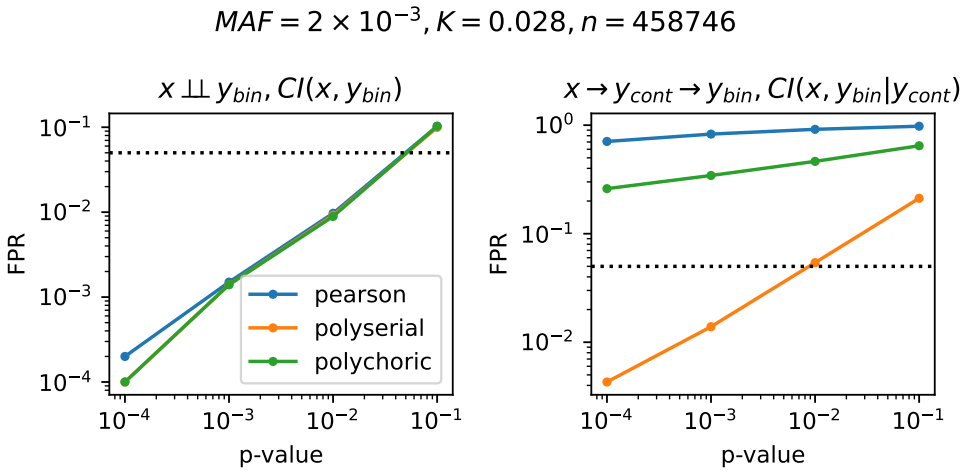

Fig. S1: **False positive rates when conducting independence tests on ordinal traits based on partial correlations.** On the left: independently drawn $x \sim Binom(2, 2 \times 10^{-3})$ and $y \sim Binom(1, K)$. On the right: $x \to y_{cont} \to y_{bin}$ where $x \sim Binom(2, 2 \times 10^{-3})$, $y_{cont} \sim \mathcal{N}(\beta x, 1)$, $y^l_{bin} \sim \mathcal{N}(\gamma y_{cont}, 1)$, and y_{bin} is 1 wherever $y^l_{bin} > \Phi^{-1}(K)$ and 0 otherwise. Here we used $\beta = (2 \times MAF \times (1 - MAF))^{-0.03}$ and $\gamma = 0.4$. We drew $n = 458746$ samples for each variables, in 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 tested for conditional independence using partial correlations as described in the methods.

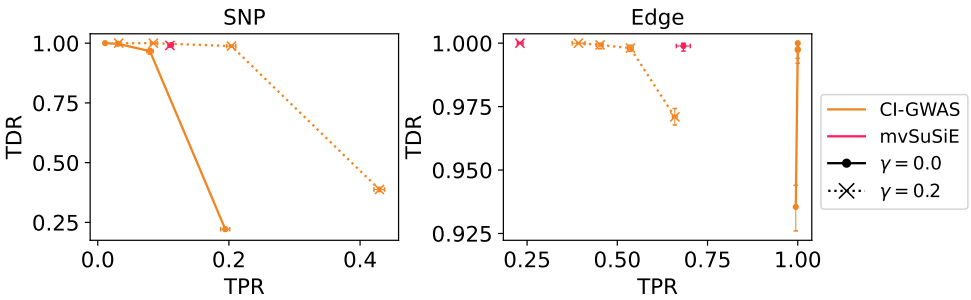

Fig. S2: **Fine-mapping performance of CI-GWAS and mvSuSiE and MVIVW in simulation with weak LD and ten traits.** Across different α values, (a) shows the fine mapping performance of CI-GWAS and mvSuSiE. Plots show true discovery rate (TDR) against the true positive rate (TPR) of globally selected SNPs (CI-GWAS) or credible sets (mvSuSiE) in the left panel, and of direct SNP-trait effects in the right panel. For mvSuSiE, a credible set was counted as true positive if it contained at least one true causal SNP. Error bars indicate the 95% confidence interval of the mean, bootstrapped. A detailed description of all calculations and MR methods used is given in the Methods section.

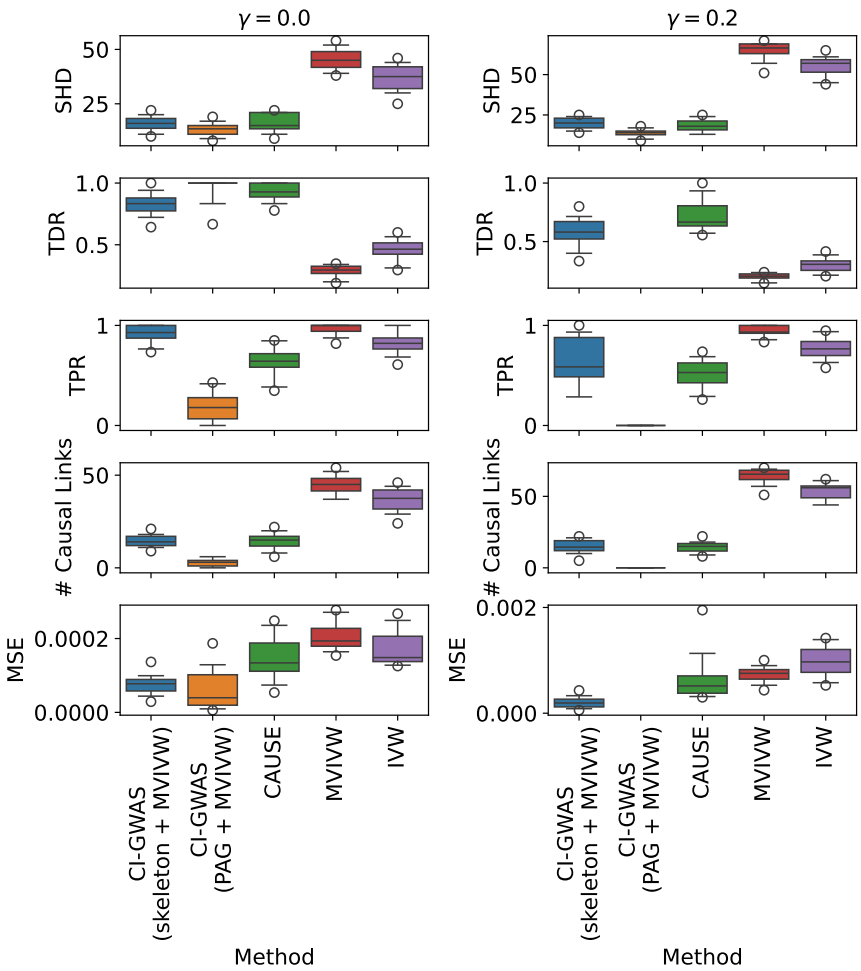

Fig. S3: **Causal inference performance of CI-GWAS and other MR methods in simulation with strong LD and ten strongly correlated traits.** We run CI-GWAS in a two of different ways: (skeleton + MVIVW): links are accepted if significant in MVIVW and an edge exists in the CI-GWAS inferred skeleton. (PAG + MVIVW): links are accepted if significant in MVIVW and an unambigously directed edge exists in the sRFCI inferred PAG. SHD: number of edge insertions, deletions or flips needed to transform the inferred graph among traits to the true graph; TDR: True discovery rate of trait-trait causal relationships; TPR: True positive rate of trait-trait causal relationships; # Causal Links: Number of trait-trait causal relationships inferred; MSE: mean squared error of inferred effect sizes of inferred causal links. We applied a p-value threshold of 0.05 to select significant causal links. Error bars indicate the 5% and 95% percentiles across 40 replicates. γ indicates the probability with which any additional affected trait is assigned to a marker after the first one and indicates the strength of pleiotropy. A detailed description of all calculations and MR methods used is given in the Methods section.

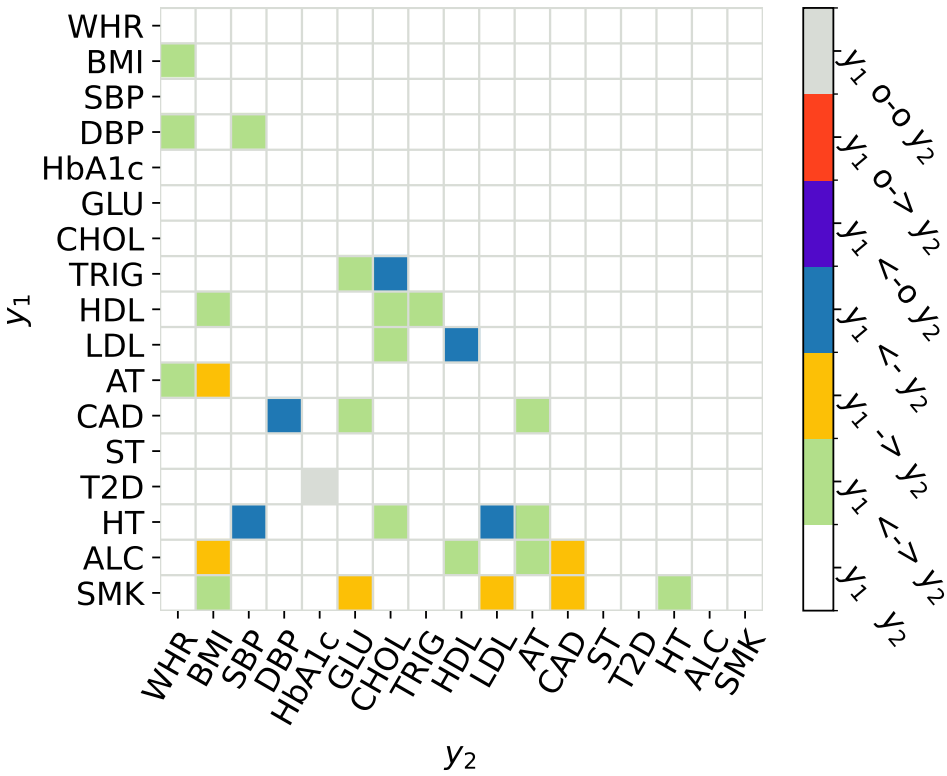

Fig. S4: **Partial ancestral graph inferred for 17 traits in the UK Biobank from sRFCI.**

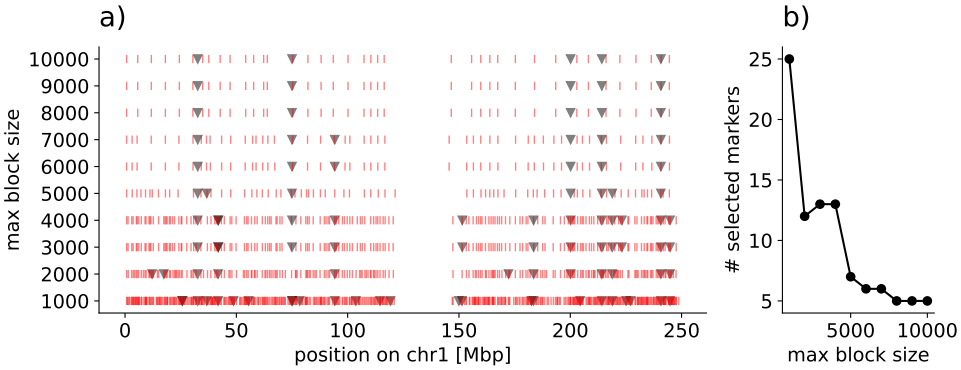

Fig. S5: **An example of the effect of maximum marker block size on the cuda-skeleton results.** The results of the cuda-skeleton are shown across various sets of block definitions generated with model parameter max block size s_{max} ranging from 1000 to 10000 markers. (a) gives the maximum block size used plotted against the positions of markers on chromosome 1. The position of markers selected as having a significant partial correlation with a trait is given by a triangle, and the marker block edges are given by red lines. (b) shows the total number of markers selected on chromosome 1, given the maximum block size used. These results were obtained from the cuda-skeleton with $\alpha = 10^{-4}$, using genomic data of chromosome 1 in the UK Biobank data and 17 traits (see Methods). Changing the marker block size either joins or separates markers that are correlated (those in linkage disequilibrium, LD). If two markers in the same block are in high LD, splitting them into two blocks can increase the partial correlation between markers within those blocks and traits, resulting in both markers being selected. Joining two neighbouring blocks, where markers previously separated are in LD, reduces marker associations. With increasing block size, fewer, more consistent markers are selected and thus we used the largest block size feasible for our GPU architecture.