Causal inference for multiple risk factors and diseases from genomics data

Nick Machnik[1†], Mahdi Mahmoudi[1†], 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]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; ilse.kraetschmer@ist.ac.at; markus_johann.bauer@boehringer-ingelheim.com;
†these authors contributed equally to this work.

Abstract

In high dimensional observational genotype-phenotype data, complex relationships and confounders make causal learning difficult. Here, we bridge a gap between genetic epidemiology and statistical causal inference, to demonstrate that graphical inference can fine-map trait-specific causal DNA variants and identify risk factors that are most likely to have a direct causal effect on a disease outcome. Our CI-GWAS approach learns a single graph representing the causal relationships among millions of DNA variants and 17 traits in less than 10 minutes on standard GPU architecture. We find over 100 trait-specific DNA variants that are exclusively exonic, with clear pathways from trait-specific "core genes" to each outcome. We separate pleiotropy from linkage to find evidence that PCSK9, LPA, and RP1-81D8.3 are pleiotropic for cardiovascular disease (CAD) with blood cholesterol, triglycerides, and low-density lipoprotiens respectively. CI-GWAS accounts for pleiotropy and selects waist-hip ratio, alcohol consumption and smoking as adjacent to CAD, with many other variables having complex paths linked through these risk factors. Our work facilitates extensive investigation of potential causal hypotheses across a wide-range of data.

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 and expense prohibit identifying all disease-associated risk factors among the vast array of human complex traits. Theoretical developments in statistical causal modelling show that graphical inference can be used to conduct causal learning among hundreds of phenotypic measures, whilst controlling for potential unobserved confounding within a single model [1, 2]. Scaling and modifying these algorithms to accommodate millions of correlated DNA variants and many hundreds of phenotypic measures would facilitate high-throughput screening of large-scale genome-wide association study (GWAS) data for the discovery of potential causal links among modifiable risk factors and disease outcomes.

Data with many highly correlated risk factors and outcomes make causal learning difficult. Currently, the most commonly used methods for causal inference in a genetic or biomedical setting center around the framework of Mendelian randomization (MR) [3–9]. MR always considers a single outcome variable (Y) and one or more exposures or mediator (M) variables, e.g. multivariate-MR (MV-MR) which considers multiple exposure variables at the same time, and tests for causal relationships among these, assuming a fixed graph. The question MR aims to address is whether there exists a causal effect, direct or indirect, of the mediator variable(s) onto the outcome. In principle, MV-MR distinguishes risk factors with direct and indirect effects on disease, but cannot infer a graph at scale describing the causal relationships among millions of genetic markers, tens of risk factors and multiple disease outcomes.

Here, our contribution is to demonstrate that graphical inference is an alternative to the MR framework that learns a single graph representing the causal relationships among all selected DNA variants and traits. We present a Causal Inference GWAS

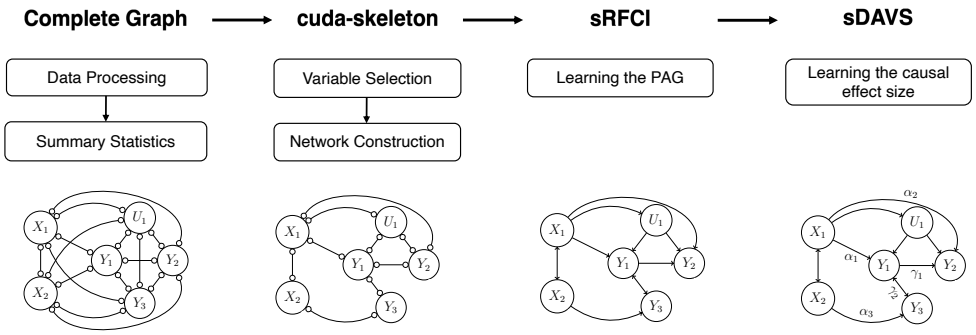

Fig. 1: **An overview of CI-GWAS.** CI-GWAS begins with a complete graph with multiple markers (X), confounders (U) and phenotypic (Y) measures. From summary statistic data, our cuda-skeleton step learns a skeleton graph of pairwise dependencies. Next, we orientate edges to obtain a partial ancestral graph (PAG) in the sRFCI step. Finally, we estimate the size of the average causal effect (ACE) in the sDAVS step. See text for a detailed description.

(CI-GWAS) approach using genome-wide summary statistic data. CI-GWAS is a set of causal discovery algorithms that are based on learning a graphical model of genetic variants and phenotypes in three steps (Figure 1c). CI-GWAS takes either individual or summary-level data, learns a graph representing the conditional dependencies between genetic variants and phenotypes (Step 1), determines whether these conditional dependencies among traits can be attributed to a causal relationship (Step 2), and then estimates the magnitude and sign of the potential causal effect (Step 3). Step 1 is a GPU-based variable selection algorithm, called cuda-skeleton, which learns a sparse undirected graph describing all possible direct relationships between all variables. Step 2 is a modified RFCI [10] algorithm (sRFCI) that infers causal directions in the inferred graph, accounting for potential hidden variables and confounders. Step 3 is a modified DAVS [11] algorithm (sDAVS) that estimates the size of causal effects across the learnt graph structure. In simulation, we show that CI-GWAS gives improved power as compared to current MR methods with equivalent accuracy for the discovery of direct effects. In the UK Biobank, we find that unique biological insight can be obtained from high-dimensional graphical inference. Our approach bridges the gap between genetic epidemiology and statistical causal inference and we provide open source software to screen for causal links among hundreds of phenotypes and many millions of genetic markers in minutes with 385 NVIDIA GeForce GTX 1080 cards.

Results

Causal discovery model: CI-GWAS

Our aim was to create a framework for causal discovery based on graphical inference for large GWAS, or biobank data sets, that scales to millions of variables and hundreds of thousands of records. An overview of CI-GWAS is given here and in Figure 1, with full details in the Methods. We conducted a simulation study (see Methods) in which we explore the ability of CI-GWAS to recover a true underlying graph from simulated data, in comparison to commonly applied MR methods. We describe these findings within each subsection below and in Figures 2, 3, and S3.

Step 1: cuda-skeleton. Step 1 of our framework is a GPU-based algorithm for variable selection, called cuda-skeleton. The algorithm learns the skeleton, an undirected graph that describes the dependencies between all variables, *i.e.* all markers and traits (see Methods). Our approach is based on the skeleton inference of the recently proposed GPU accelerated implementation of the PC-stable algorithm [11, 12]. It accepts either a plink file of individual-level genotype-trait data, or summary statistic data consisting of (i) a matrix of SNP-SNP correlations (LD matrix); (ii) marker-trait correlations that can be obtained from standard GWAS association test statistics; (iii) a matrix of trait-trait correlations (see Methods). We accommodate the order 10^9 markers and order 10^2 traits performing variable selection in separate marker-blocks, after blocking the matrix of marker linkage disequilibrium (LD) with an algorithm inspired by [13] (see Algorithm 1), and later merging the block results into one graph. A sensitivity analysis shows that this approach results in consistent causal marker selection within the UK Biobank (Figure S4) across different block sizes.

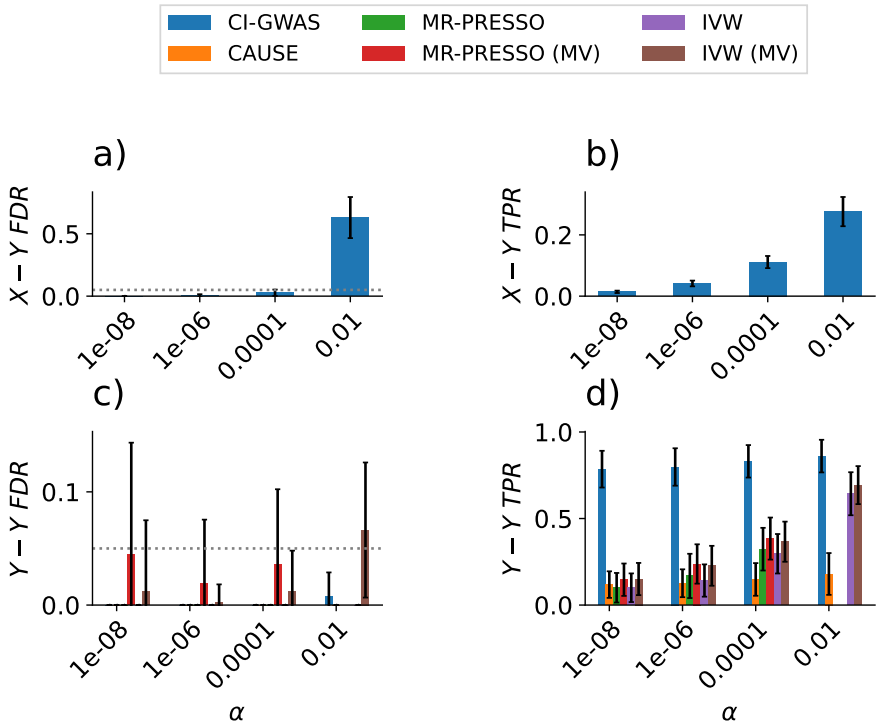

Fig. 2: **Performance of dependency inference of CI-GWAS and Mendelian randomization methods in a simulation study using UK Biobank data.** Across different α values, (a) shows the false discovery rate (FDR) of marker-trait adjacencies (X-Y) for CI-GWAS, (b) shows the true discovery rate (TPR) for the same, (c) shows the FDR for trait-trait adjacencies (Y-Y), and (d) shows the TPR for the same, except for MR-PRESSO, where a large number of selected instrument variables prevented the algorithm from returning estimates. For Mendelian randomization (MR), α is the significance threshold for the instrument variable selection. For CI-GWAS, α is a model tuning parameter (see text for details). A detailed description of all calculations and MR methods used is given in the Methods section. Error bars give the SD across simulation replicates. Dotted lines indicate FDR=0.05 Bonferroni correction was applied to all MR results except for CAUSE.

In our simulation cuda-skeleton selects markers with a low false discovery rate (FDR), reasonable true positive rate (TPR), and exhibits a low FDR and high TPR for trait-trait adjacencies, dependent on the value of the tuning parameter α, which controls the sparsity of the graph (see Methods for more details) (Figure 2). Theory for skeleton construction gives the appropriate α as $\alpha = 2(1 - \Phi(n^{1/2}c_n/2))$ [14], with n the sample size, Φ the CDF of the Normal distribution and c_n the number of true adjacencies in the data. As c_n is not known in practice, we explored realistic parameter settings within our simulation study. We find that an $\alpha = 10^{-4}$ gives low

FDR (0.0 for trait-trait and < 0.05 for marker-trait) and high TPR (~ 0.83) for trait-trait adjacencies, but a lower TPR of 0.15 for marker-trait links (Figure 2). We also investigate the ability of cuda-skeleton to estimate the degree of pleiotropy (the exact same causal DNA variants) among marker-trait relationships, by simulating two scenarios, one where causal markers are not pleiotropic and one where there is widespread pleiotropy (for full details see Methods). We find that cuda-skeleton can detect marker-trait pleiotropy, although as suggested by the TPR for marker-trait adjacencies, the degree of pleiotropy is underestimated (Table S2).

We compare CI-GWAS performance to commonly used MR methods: IVW implemented in the MendelianRandomization R package [15], MR-PRESSO [9], their multivariable extensions and CAUSE [16]. For the MR methods, α refers to the significance threshold used for instrument variable selection from standard GWAS summary statistics that we generated from the simulated data (see Methods for more details), where the commonly used threshold is 5×10^{-8} [15]. In comparison to MR methods, cuda-skeleton infers causal adjacencies among traits with same or lower FDR (Figure 2c), but with more power (TPR, Figure 2d). We find elevated FDR for multivariable MR and generally lower TPR for all MR methods across all significance thresholds without Bonferroni correction (Figure S1). Thus, for all MR methods we present results after Bonferroni-correction, with the exception of CAUSE as Bonferroni correction resulted in zero findings. Note that as described in the Methods, for single variable MR, we consider any significant effect correct when there is a direct or indirect effect in the true graph. For multivariate MR and CI-GWAS, we only consider significant effects correct where there is a direct effect in the true graph.

We also considered a combination of cuda-skeleton and MR; previous research has suggested that graphical inference can be used to discover instrument variables suitable for MR ([17]). We used cuda-skeleton to select instrument variables and then tested whether this improved the TPR for trait-trait adjacencies. Contrary to what was suggested previously, we find that TPR decreases when cuda-skeleton selected instruments are used. Even giving the true causal markers as instrument variables to MV-IVW resulted in a drop in power, likely due to weak-instrument bias (Figure S3).

Step 2: sRFCI. 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 overcome the problem of unobserved variables, we base our method on the RFCI algorithm [10] for the inference 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. Therefore, in Step 2 of CI-GWAS, we employ a novel order-independent algorithm that orientates edges between adjacent variables to

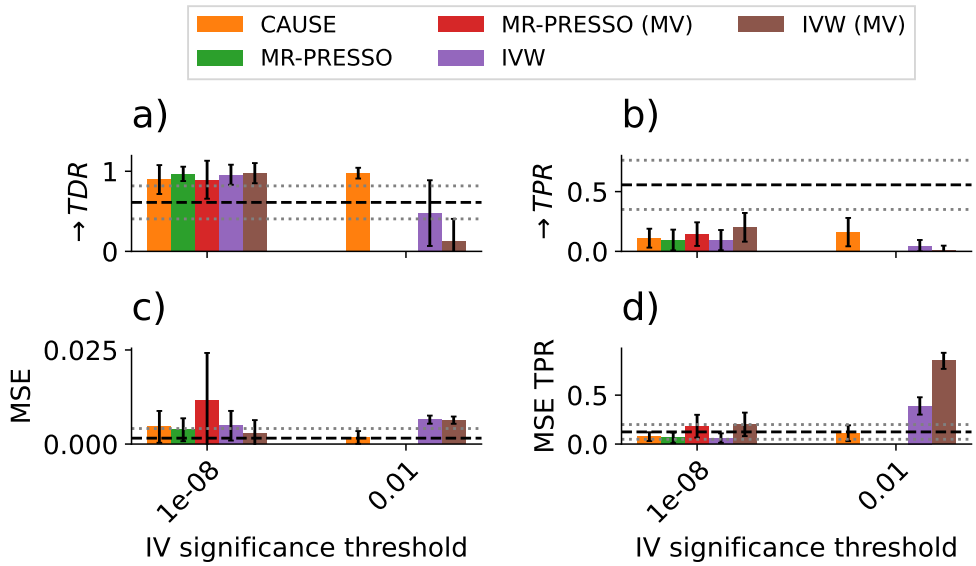

Fig. 3: **Performance of orientation and causal effect size estimation of CI-GWAS compared to Mendelian randomization methods in a simulation study using UK Biobank data.** At two instrument variable (IV) selection thresholds, (a) shows the true discovery rate (TDR) for the direction of causal relationships, (b) shows the true positive rate for the same, (c) shows the mean squared error (MSE) for relations where an effect size was inferred and (d) shows the TPR, i.e. the recovery rate of true non-zero effects. Dashed lines give the CI-GWAS value at $\alpha = 10^{-4}$ and 1SD. A detailed description of all calculations and MR methods used is given in the Methods section. MV refers to multi-variable. Note that for MR-PRESSO, a large number of selected instrument variables prevented the algorithm from returning estimates for threshold 0.01. Error bars give the SD across simulation replicates. Bonferroni correction was applied to all MR results except for CAUSE.

determine ancestral relationships, whilst maximising the inclusion of selected adjacent variables (see Methods for full details).

The outcome of our sRFCI step is a Partial Ancestral Graph (PAG), where the types of dependencies among traits are described by different types of edges. Specifically, an arrowhead ">" at a variable means that it is not an ancestor of the variable at the other end of the edge. A tail "−" at a variable means that it is an ancestor of the other. A circle mark "∘" means that the edge mark cannot be inferred unambiguously and there may or may not be an ancestral relationship. In simulation, we compared the true orientations to the orientation of edges discovered by cuda-skeleton for CI-GWAS. For MR, we compared the true orientations to the direction of the causal effect inferred by MR by testing all pair-wise relationships in both directions.

We find that the sRFCI CI-GWAS step gives a lower true discovery rate (TDR) of correctly orientated edges than MR methods with Bonferroni correction, but a higher

TPR (Figure 3a,b). Without Bonferroni correction, we find that MR methods have a TDR comparable to CI-GWAS, but still lower TPR (Figure S2a,b). We wish to stress that our simulation shows that cuda-skeleton reliably discovers adjacencies with high power. This means that CI-GWAS gives a more complete graph and it finds more correctly orientated edges than MR, but with the consequence that a proportion of the edges may be incorrectly orientated. Thus, there is a trade-off between accuracy in orientation and power.

Step 3: sDAVS. Having learnt a PAG, we estimate the average causal effect (ACE), which represents the expected effect of an intervention to alter the exposure by 1 SD. We use the DAVS algorithm [11], adjusted so that it requires only summary statistic data. In our simulation study, we find that mean squared error (MSE) of the average causal effect estimates returned by sDAVS is comparable to those of CAUSE and multivariate IVW (Figure 3c). Note that for MSE calculation, we only considered non-zero estimates of causal effects, meaning that missing causal effects do not contribute to the error. In our simulation setting, all methods recover few causal effects (Figure 2d) and CI-GWAS offers comparable performance to MR, but a fuller graph is returned allowing for better exploration of which risk factors have the most potential to be causal.

UK Biobank application

We applied CI-GWAS to 17 traits and 2.17 million single nucleotide polymorphism (SNP) markers in UK Biobank data, as described in the Methods and Table S1. We find a small number of non-pleiotropic marker-trait adjacencies (parental markers) and small number of pleiotropic links (Figure 4a, b). The markers discovered represent joint marker-trait associations that are conditional on all traits and all other markers, with the conditioning on other markers controlling for LD (essentially they can be viewed as multi-trait genome-wide "fine-mapping" results, that select trait-specific causal DNA variants). In simulation, we underestimate pleiotropy and thus we investigated this further by moving back within the graph, which revealed increasing marker sharing especially for traits of higher polygenicity like human height. This finding is consistent with trait-specific causal markers being in LD as opposed to being directly pleiotropic (Figure S5). However, we caution that these estimates are likely an empirical lower bound as there is insufficient power to fine-map all genome-wide plieotropy to the single-locus level.

We annotate these significant causal marker discoveries and find that almost all are protein coding variants (see Supplementary Data Table 1 for full results). For BMI, our significant causal markers are all protein coding variants in FTO, glucose transporter SLC2A3, CACNG3, CCND2, TNNI3K, and ANAPC4. We find a variant near SOCS3 (a cytokine gene involved in immune response and inflammation) and TSLP (directly implicated in epithelium allergy and asthma), and inside exons of EFS (CAS family immunoligical self-tolerance), IL33 (involved in allergic inflamation), and TNFRSF8 (TNF receptor linked to asthma) for asthma. We find that ZSWIM5, which regulates alcohol dehydrogenaise ADHFE1, is causally linked to alcohol consumption. We find causal links between LPA (Lipoprotein(A)), PCSK9 (regulates cholesterol in the bloodstream), CMSD1 (known candidate cardiovascular disease gene), CPXM2 (a

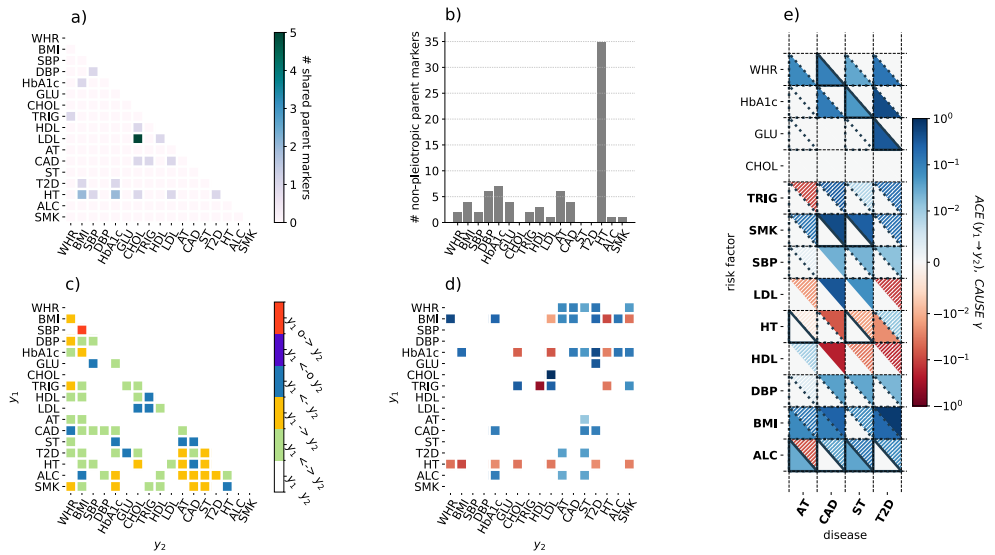

Fig. 4: **CI-GWAS results for 17 traits in the UK Biobank.** (**a**) Number of ancestral markers shared by pairs of traits in the inferred skeleton. (**b**) Number of exclusive parent markers in the inferred skeleton for each trait. (**c**) Inferred PAG edge types between traits. Bi-directed arrows indicate the presence of an unmeasured confounder of y_1 and y_2 and the absence of a causal relationship. (**d**) Inferred average causal effects (ACE) between traits. (**e**) Comparison of inferred causal effects of risk factors onto diseases, between CI-GWAS (lower triangle) and CAUSE (upper triangle). Boldfaced labels indicate traits that were part of both the CAUSE and CI-GWAS analyses. Dotted triangle outlines indicate inferred causal paths for CI-GWAS, solid outlines indicate inferred direct effects. Non-striped triangle faces indicate significance of causal relationships in CAUSE at $p < 0.05$. Colours give the strength and sign of the ACE. CAUSE results were taken from Supplementary Table 3 of [16]. CI-GWAS was run with $\alpha = 10^{-4}$. Three letter codes give the trait name as described in Table S1. Inferred ACEs are given in Table S3.

novel candidate for hypertension in mouse models [18]), BCL2 (regulates apoptosis in the myocardium) and PCSK9 (a target for therapeutic agents for lowering LDL-C) with cardiovascular disease. We find that PCSK9, LPA and RP1-81D8.3 are pleiotropic for cardiovascular disease (CAD) and cholesterol (CHOL), triglyceride levels (TRIG) and low-density lipoprotiens (LDL), respectively, which supports a shared genetic basis among CAD and lipid metabolism. For type-2 diabetes (T2D), we identify CCND2 (SNP rs76895963 whose rare allele is associated with reduced susceptibility in GWAS studies [19]), which is a provisional therapeutic target for beta cell proliferation and is pleiotropic for HbA1c levels in blood. Finally, we find that systolic blood pressure is causally linked to TBX5 (a DNA-binding protein involved in heart development), SIK1 (an antiproliferative factor in hypoxia-induced pulmonary arterial remodeling [20]), CDK6 (supporting a role of microRNAs in cardiovascular disease [21]) and RGL3 (a

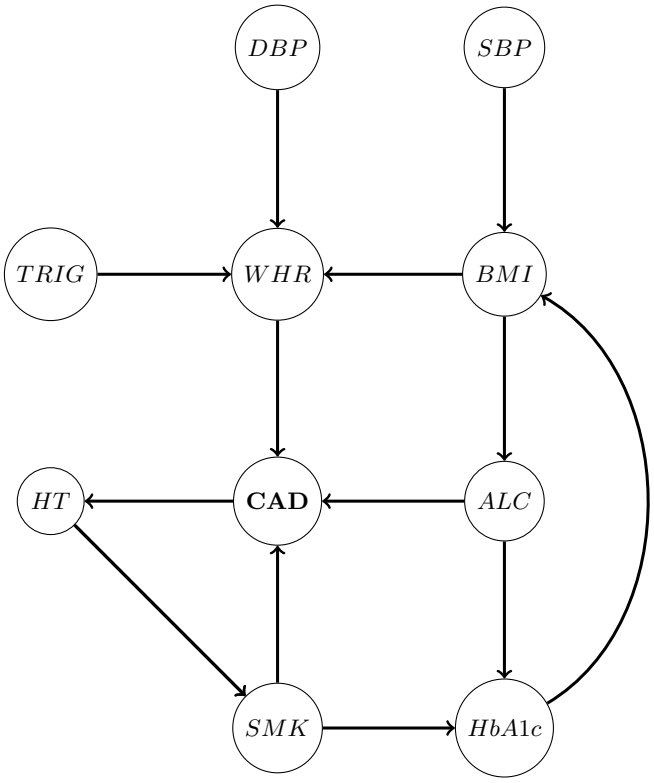

Fig. 5: **Directed graph between traits linked to cardiovascular disease (CAD) as inferred by CI-GWAS in the UK Biobank**. Arrows represent the inferred orientation of the edges among traits. We note that the inferred graph is an invalid PAG and we caution in inferring orientations between variables in large-scale observational data.

known hypertension gene). These results (and others in Supplementary Data Table 1) imply that our approach identifies trait-specific sets of "core genes". These are sometimes found as top associated loci in genome-wide association studies, but not always, and they are generally not the leading associated variants (Supplementary Data Table 1). This is expected as GWAS findings are not outcome-specific and single-trait analyses obtain results that reflect complex phenotypic relationships. In conclusion, the cuda-skeleton step of CI-GWAS discovers marker-trait causal adjacencies that are generally exonic, with clear pathways to each outcome.

The PAG estimated in Step 2 is presented in Figure 4c, where non-white squares indicate the discovery of an adjacency and colours give the edge type inferred in the sRFCI orientation. We find many bi-directed links, where the adjacency discovered by cuda-skeleton reflects the existence of latent confounders, rather than a causal link (note they do not represent bidirectional causality). Uni-directed links imply the existence of an adjacency that cannot be removed from the graph or attributed to

latent confounding as shown in Figure 4c. In Figure 4d we present the average causal effects (ACE) estimated in the sDAVS step, which are returned when a path is found between two variables and proper adjustment sets are available. We compare these estimates to those given by CAUSE [16] for the effects of risk factors on common complex disease outcomes. Given the links shown in Figure 4c, we find significant evidence for potential paths between many risk factors and disease outcomes (dotted borders in lower triangles of Figure 4e), although sDAVS has insufficient data to give estimates of the ACE for many of them. All ACE estimates can be found in Table S3.

Our results show that inferred causal relationships from observational data are highly complex and that inference is limited in high dimensional biobank data. We focus on variables with direct adjacencies as we propose these are risk factors with a link to the outcome that is inferred to not act through other traits (solid lower triangle outlines). We find that waist-hip ratio (WHR) is adjacent to cardiovascular disease (CAD), HbA1c blood levels are adjacent to risk of stroke (ST), blood glucose levels (GLU) are adjacent to whether an individual has been diagnosed with type-2 diabetes (T2D), and alcohol consumption (ALC) is adjacent to all disease outcomes directly. Focusing on the links around cardiovascular disease (CAD) in Figure 5 as an example, we find direct adjacencies of WHR, ALC, and smoking (SMK). This implies that these are the leading risk factors for CAD within the UK Biobank data. Linked to these three leading risk factors are then other previously reported CAD risk factors [22]. Contrary to a recent study [16], we find no adjacency between LDL and CAD from cuda-skeleton, rather we find evidence of pleiotropy in the marker-trait relationships from cuda-skeleton between blood lipid levels and CAD as reported above. It has been reported before that the estimated effect of LDL onto CAD becomes insignificant when controlling for HDL and apolipoprotein levels in a multivariate MR analysis [23]. Similarly, we condition HDL and other measures of blood lipid contents when assessing the dependency between LDL and CAD. Above a pleiotropic link, we then find a path from blood lipids to CAD through WHR. Likewise, we find paths from body mass index (BMI) and systolic and diastolic blood pressure (SBP, DBP) to CAD involve WHR and ALC. However, we stress that the orientation of all of these adjacencies will not be correct, nor are they unbiased of ascertainment as discussed below. However, among all the risk factors for CAD, CI-GWAS can select among them to determine which are most likely to have a direct adjacency and which have a more complex path, whilst taking into account genetic pleiotropy among variables.

Discussion

Here, we presented a method to make statistical causal inference in a genomic setting while allowing for the presence of multiple confounders and multiple correlated phenotypes. We have demonstrated that causal discovery can be extended from MR methods to multivariable graphical inference for large GWAS or biobank data sets with millions of variables and hundreds of thousands of records. In this work, we attempt to learn a causal graph by inferring the adjacencies, orientating the edges and then estimating the strength of the causal effects. The algorithms we present have strong theoretical

support and have been heavily studies in the causal inference literature [10–12]. Current MR methods generally give consistent results, however they all build upon each other and assume a similar fixed directed acyclic graph. Here, we find that previously inferred potential causal relationships actually reflect complex paths among multiple risk factors and common complex diseases. While this complex network of adjacencies makes causal learning difficult, our results suggest that within highly correlated observational genotype-trait data, we can select risk factors that are most likely to have an adjacency.

High-dimensional non-sparse graphical causal inference is an active area of research. Multivariable MR [9, 24, 25] has been proposed for this problem, but this does not build a full graph. Structural equation models have also been suggested [26], 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 [27]. 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. Thus, we focus here on applying and developing a principled, theoretically supported method that provides a crucial step toward high-dimensional causal learning for genomics.

Within this framework, there are a number of current limitations that require future research. Firstly, we show in simulations that CI-GWAS does not find all causal markers, which limits its ability to orientate causal relationships correctly. In RFCI and similar algorithms, the orientation of edges relies on separation sets and given the oracle set of these conditional independencies and separation sets, theory states that RFCI will recover the correct orientations [28]. Due to the small magnitude of marker-trait effects, noise, strongly correlated traits and the general tendency of statistical tests to make mistakes, our separation sets will differ from the oracle set and therefore result, inevitably to some degree, in wrong orientations. In the absence of an oracle set, there is a trade-off between speed and accuracy: there is a mode proposed for RFCI that uses a large number of possible separation sets to make decisions about orientations, which is infeasible for graphs of any practical size and density. Our modifications to the RFCI algorithm presented here aim to make RFCI applicable in the genomic setting, while retaining as much accuracy as possible, but will benefit from future work. Orientation of discovered causal adjacencies accounting for latent confounding is currently the major limitation of all causal inference methods, including MR, when applied to genomic data. Additionally, sDAVS leaves many inferred causal paths without estimates. Our inference framework will benefit from future work that will experiment with other existing methods for estimation of effects based on an inferred graph, or developing new methods for this.

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 publically available, but they could easily be shared as they represent sufficient population-level statistics and not individual-level data. Additionally, data

sets have structure that can inhibit unbiased causal inference. Selection biases, like the ascertainment of healthy older people into biobank studies, can also bias causal inference [29]. 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 is impossible to observe within the UK Biobank as the majority of people are being treated with some form of drug. We adjust for the effects of drug taking prior to analysis, but if the majority are on drugs this is likely ineffective. Furthermore, genotype-phenotype associations likely reflect a complex combination of direct, parental genetic and epigenetic effects, and we note that parental genotypes and their relationship to offspring traits can also be easily accommodated within our framework. A potential solution to data bias issues may be to utilise longitudinal records, where discovered adjacencies in cuda-skeleton can then be orientated longitudinally and combined with therapeutic information. With repeated measures data, statistical causal effects can 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.

In summary, our approach is applicable to screen for potential causal relationships within other forms of data in life-science, from genomics, to the microbiome, to single-cell sequencing and phenotyping data. 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 now facilitate extensive investigation of potential causal hypotheses across a wide-range of data types.

Methods

CI-GWAS

CI-GWAS is a set of causal discovery algorithms that are based on learning a graphical model of genetic variants and phenotypes. CI-GWAS takes either individual or summary-level data, learns conditional dependencies or links between genetic variants and phenotypes, determines whether these 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 main steps: i) a GPU-based algorithm that conducts variable selection to give a sparse graph that describes likely causal relationships among markers and traits (called cuda-skeleton); ii) an order-independent separation set selection algorithm combined with an algorithm to then infer possible causal directions in the graph, under the consideration of potential hidden variables and unmeasured confounders (called sRFCI); and finally (iii) an algorithm to learn the magnitudes of the causal effects from the ancestral graph learned in the previous step (called sDAVS). We provide detailed information on each step of CI-GWAS in the following sections, alongside algorithms for their implementation. Figure 1 provides a pictorial overview of our approach.

For each step, we require the following sufficient summary statistics from a data set: (i) the matrix of SNP-SNP correlations (LD matrix); (ii) a matrix of marginal

marker-phenotype correlations; (iii) a matrix of phenotype-phenotype correlations. Our software either calculates these directly from the individual-level data, or any choice of approximations of (i) to (iii) are accepted within our algorithm. For (i), 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 sample, if the original sample matrices are unavailable [30]. All of these modelling choices can be accommodated within our software. For (ii), assuming the response variables were standardized to zero mean and unit variance, standard marginal ordinary least squares regression summary statistic estimates (GWAS summary statistics data) can be scaled by the expected marker variance (calculated from the minor allele frequency of each marker) to give the marginal correlations required. We provide open source code and scripts to run our full analysis (see Code Availability).

Step 1: Variable selection by cuda-skeleton

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: (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). Our

approach is based on the first part (skeleton inference) of the recently proposed cuPC algorithm [31], a GPU accelerated implementation of the PC-stable algorithm [12].

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 [13] 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 S4.

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.

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^{-4}
blocking	s_{max}	10^4
blocking	ϵ	10^2
blocking	corr-width	$2 * 10^3$

Step 2: Orientating edges

Having identified a set of variables that are most likely to have direct causal effects on each other, we now seek to determine the ancestral relationships among the variables. For example, if we identify that a pair V_i, V_j are adjacent, then we wish to determine which variable is ancestral, which then allows us to infer the direction of the causal relationships. We base this step on the RFCI algorithm of Refs. [10, 28] with a number of key modifications.

The separation sets S returned by cuda-skeleton represent the minimal number of variables required to remove edges among variables. In a genomics setting, markers have small causal effects on traits, and therefore traits are often missing from the separation sets of markers and other traits, which leads to wrong edge orientation. For the edge orientation within sRFCI, we now wish infer two types of separation sets: S^{max} a maximal separation set, meaning that the addition of any other variable results in $Z > \tau$, and 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 [12] provide orientation by maximally searching all potential separation sets, but this unfortunately is infeasible in a high-dimensional genomic setting as it requires checking $\sum_{k=0}^{n} \binom{n}{k}$ 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. It is possible that S^{max} is not the largest separation set among all possible sets. Likewise, S^{mpc} may not be the separation set that yields the lowest partial correlation among all possible sets. 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 [10].

Step 2. Orientate the v-structures using S^{max} and Algorithm 4.4 of [10].

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 [10]. This step involves repeated applications of orientation rules (R1) to (R10) of [28]. We implemented the suggested modifications from [12] 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 and we use this only in the following sDAVS step described below.

The output of this algorithm is a Partial Ancestral Graph (PAG) representing a the set of possible Maximal Ancestral Graphs (MAG, see Ref. [10] for definition). Adjacencies are inferred to be composed of arrows ($<$ or $>$), circles ($\circ$), and tails ($-$).

Algorithm 2 Separation set selection

Require: α, n, A^*, C

$ut \leftarrow$ unshielded triples of A^*
$op \leftarrow$ outer variables in ut
for V_i, V_j in op **do**
  $N \leftarrow \text{neighbors}(V_i, A^*)$
  $N' \leftarrow N$
  $sep \leftarrow \{\}$
  $foundSepset \leftarrow$ 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

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 [10], which has been shown to be of importance for Mendelian Randomisation methods applied to biobank data [6, 16, 32–34]. 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.

Step 3: Causal effect estimation

Having learnt the PAG, we then seek to estimate the sign and magnitude of the total causal effect. Our approach is a modified version of the Data-driven Adjustment Variable Selection (DAVS) algorithm [11] that does not rely on access to individual-level data. Thus, the input required is the PAG learnt from RFCI and the same correlation matrices already computed for cuda-skeleton.

The DAVS algorithm aims to estimate the average causal effect (ACE), which represents the expected effect of an intervention to alter the exposure by 1 SD. We modify the algorithm (for details, see [11]) to utilise the correlation matrices rather than individual-level data as follows:

Step 1. For each pair of variables V_i to V_j such that V_i is ancestral to V_j in the PAG obtained from sRFCI, V_i is selected as the mediator variable and V_j is selected as the outcome variable. A variable that has a directed edge to another variables, is its parent. Variables that are connected by a bidirected edge are spouses. DAVS creates a set Q of candidate COSO variables, containing all variables that are parents or spouses of V_i and neither parents nor spouses of V_j.

Step 2. DAVS checks if all members of Q are conditionally independent of the outcome variable V_j, given the mediator variable V_i ($Q \perp\!\!\!\perp V_j | V_i$). If this condition holds, the $ACE(V_i, V_j)$ can be estimated without any adjustments.

Step 3. If this condition does not hold, then additional steps are taken to identify sets of proper adjustment variables based on the PAG. Here we restrict the search for adjustment variables in the graph to a depth of 2, and the maximal size of candidate adjustment sets to 3.

Step 4. DAVS estimates ACE with adjustment variables. Either using no adjustments sets in Step 2, or those identified in Step 3.

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[35], 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. For computational convenience, we then removed markers in very high LD using the clumping approach of plink, where we ranked SNPs by minor allele frequency and then selected the highest-MAF SNPs from any set of markers with LD $R^2 \leq 0.8$ within a 1-Mb window. This results in the selection of a tagging set of variants, with only variants in very high LD with the tag SNPs removed. These filters resulted in a data set with 458,747 individuals and 2,174,071 markers.

Simulation study

We conducted a simulation study where we generate an underlying graph and examine the ability of CI-GWAS to: (i) recover adjacencies (to learn the skeleton); (ii) infer edge orientations; and (iii) infer the average causal effect. We aimed to assess the effectiveness of our approach in comparison to other existing methods in the field of MR, within a simulation setting that was close to our analysis conducted on biobank data.

We took the imputed SNP data of the UK Biobank, selecting chromosome 1 (169,904 genetic markers) and 458,747 individuals. Our blocking procedure (Algorithm 1) gave 32 genomic blocks for chromosome 1. We select the first 10 of these as containing causal variants. We randomly selected 5 SNP markers for each of 10 traits and 2 latent variables within each of the 10 blocks to obtain 50 underlying causal variants for each trait on this chromosome. We simulated a marker-trait and trait-trait adjacency matrix G, causal effects A, and phenotypic data as described below. We do not consider the graph among genotypes, rather we focus on our ability to recover the randomly drawn underlying causal variants in the presence real-data LD, and to estimate the causal effects among phenotypes.

Specifically, 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 horizontal pleiotropy, 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 provide a function to simulate data of this kind (see Code Availability). 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) [15]. MR-PRESSO was implemented using code from the original publication (version 1.0) [9] with outlier correction. We employed the CAUSE R package (version 1.2.0.0331) [16] to perform the CAUSE analysis. Unless stated we used default options given by each software. We applied either a significance threshold of $p < 0.05$ or a Bonferroni corrected threshold of $p < \frac{0.05}{720}$ to the trait-trait causal relations (each trait is compared to nine other traits, in two directions, in 40 replicates). Additionally, we re-ran MR methods using instrument variables identified by CI-GWAS at $\alpha = 10^{-4}$ and $\alpha = 0.01$.

We present true positive rate (TPR) and false discovery rate (FDR) for marker-trait causal edges, TPR and FDR for trait-trait edges, the orientation discovery rate (TDR), and finally we the mean square error (MSE) of the causal effect estimates. We make two matrices of dimension *pxp* for each replicate. The first matrix contains the true traits adjacencies. The second matrix contains indicators of which trait are dependent through a latent variable which is ancestral to both, which is represented

by a bi-directed edge in a PAG. The logical OR over these two matrices gives the true expected adjacency matrix. 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 adjacency matrix from CI-GWAS. We compute FP, which is the number of inferred, but not true edges. We then calculate $TPR = TP/P$ and $FDR = FP/(TP + FP)$. For MR, the TPR is calculated in exactly the same way, where if a significant causal effect estimate is found in MR then we use this as an inferred adjacency. For the FDR calculation for 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 FDR in the same way as for CI-GWAS.

To calculate the orientation TDR, we divide the number of correctly orientated unidirected edges by the total number of inferred unidirected edges. We regard all edge types that contain the true ancestral relationship between the adjacent variables as correct. We exclude bidirected edges from the orientation TDR, because these do not contribute to causal effect estimation in the later DAVS step. For MR, we compare the true causal paths with the matrix of inferred significant relationship between pairs of traits, again asking what fraction of the inferred relationships are true. If MR infers a significant effect in both directions, we regard this as an bidirected edge and exclude it from the TDR calculation. For MV-MR, we compare the orientation in the same was as for CI-GWAS.

To compute the MSE, we use the matrix of inferred effects and the matrix of true causal effects (T) over paths, because our sDAVS step estimates the average causal effect is across the paths linking two variables.

$$M = (I - A_{y-y}^T)^{-1}$$
$$T = MM^T$$

with A_{y-y} the true causal effect between adjacent traits.

We exclude all missing or 0 entries in the inferred effect matrix from the calculation, thus our 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 FDR and TPR for adjacency recovery above, and we wish to assess the error of positive discoveries.

Finally, to assess the ability of CI-GWAS to capture pleiotropic markers, we compared the mean of the inferred pleiotropy ($\frac{\text{number of influenced traits}-1}{\text{total number of traits}-1}$) over all markers to the true *probPleio* in the simulation. The results are given in Table S2.

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 [36], 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 [37] version 2.30.0 with

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

We then took the leave-one-chromosome-out GWAS results presented in [38], selecting LD independent SNPs within a 5MB window, with LD< 0.01 and significance threshold $p < 5e^{-8}$, which is the same UK Biobank selected data as used within this work. Using these LD independent SNPs, we asked whether they were in LD> 0.01 with the CI-GWAS identified parental markers for each trait. If so, we recorded that rank. We did this to assess whether CI-GWAS selected parental markers are those that are identified by standard association testing on the same data.

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

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

[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] 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).

[9] 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).

[10] 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).

[11] 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).

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

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

[14] Kalisch, M. & Bühlmann, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research* **8**, 613–636 (2007).

[15] 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).

[16] 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).

[17] 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).

[18] Grabowski, K. *et al.* Cpxm2 as a novel candidate for cardiac hypertrophy and failure in hypertension. *Hypertension Research* **45**, 292–307 (2022).

[19] 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).

[20] Pires, N. M. *et al.* Acute salt loading induces sympathetic nervous system overdrive in mice lacking salt-inducible kinase 1 (sik1). *Hypertension Research* **42**, 1114–1124 (2019).

[21] Papageorgiou, N., Tslamandris, S., Giolis, A. & Tousoulis, D. MicroRNAs in cardiovascular disease: Perspectives and reality. *Cardiol. Rev.* **24**, 110–118 (2016).

[22] Collins, D. R. J. *et al.* Global cardiovascular risk assessment in the primary prevention of cardiovascular disease in adults: systematic review of systematic reviews. *BMJ Open* **7** (2017). URL https://bmjopen.bmj.com/content/7/3/e013650.

[23] Richardson TG, P. T. A.-K. M. F. B. D. S. G. H. M., Sanderson E. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis. *PLoS Med.* **17** (2020).

[24] 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).

[25] 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).

[26] 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).

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

[28] 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).

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

[30] 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).

[31] 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.

[32] 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).

[33] 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).

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

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

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

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

[38] Orliac, E. J. *et al.* Improving gwas discovery and genomic prediction accuracy in biobank data. *Proceedings of the National Academy of Sciences* **119**, e2121279119 (2022). URL https://www.pnas.org/doi/abs/10.1073/pnas.2121279119.

Supplementary Information

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

Nick Machnik, Mahdi Mahmoudi, 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).

α	pleiotropy frequency	estimated pleiotropy frequency	variance of estimated pleiotropy frequency
1.0e-02	0.2	0.048171	0.007458
1.0e-02	0.0	0.005727	0.000636
1.0e-04	0.2	0.072082	0.007363
1.0e-04	0.0	0.000335	0.000037
1.0e-06	0.2	0.048584	0.005095
1.0e-06	0.0	0.000299	0.000033
1.0e-08	0.2	0.025887	0.002853
1.0e-08	0.0	0.000000	0.000000

Table S2: Frequency of pleiotropy in simulated true graphs and the CI-GWAS inferred graphs for different values of α. Calculated in simulations with 600 causal variables, 10 traits and 20 replicates per true pleiotropy frequency. Means and variances were calculated over all markers with at least one child trait.

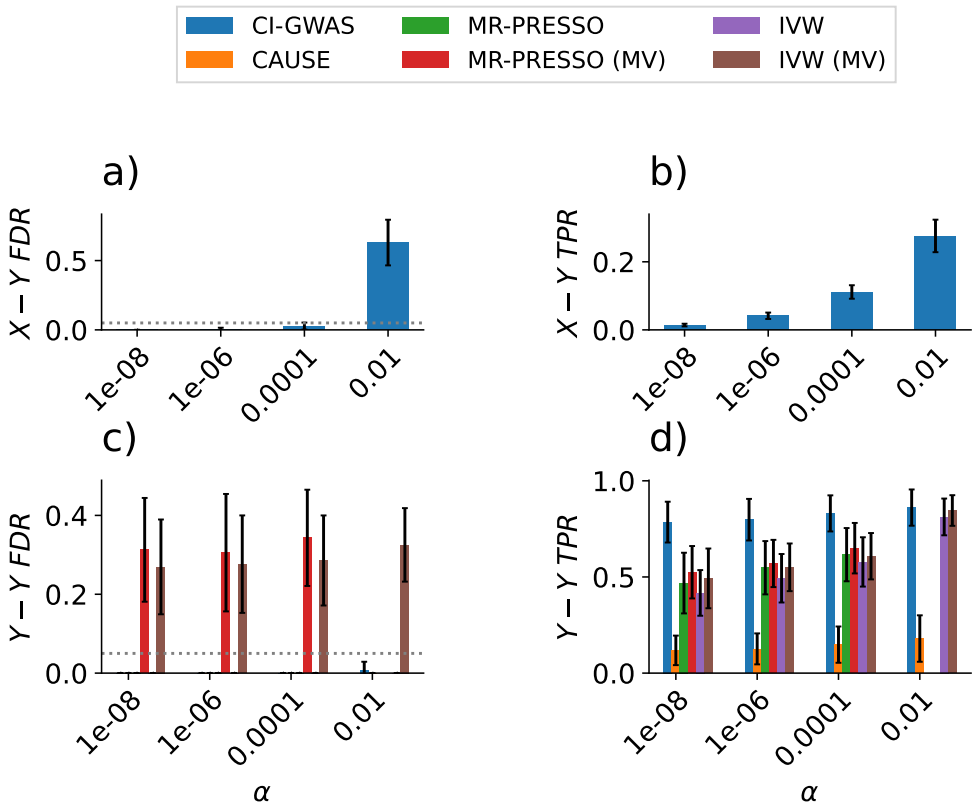

Fig. S1: **Performance of dependency inference of CI-GWAS and Mendelian randomization methods in a simulation study using UK Biobank data.** Across different α values, (a) shows the false discovery rate (FDR) of marker-trait adjacencies (X-Y) for CI-GWAS, (b) shows the true discovery rate (TPR) for the same, (c) shows the FDR for trait-trait adjacencies (Y-Y), and (d) shows the TPR for the same, except for MR-PRESSO, where a large number of selected instrument variables prevented the algorithm from returning estimates. For Mendelian randomization (MR), α is the significance threshold for the instrument variable selection. For CI-GWAS, α is a model tuning parameter (see text for details). A detailed description of all calculations and MR methods used is given in the Methods section. Error bars give the SD across simulation replicates. No multiple testing correction was performed for any methods.

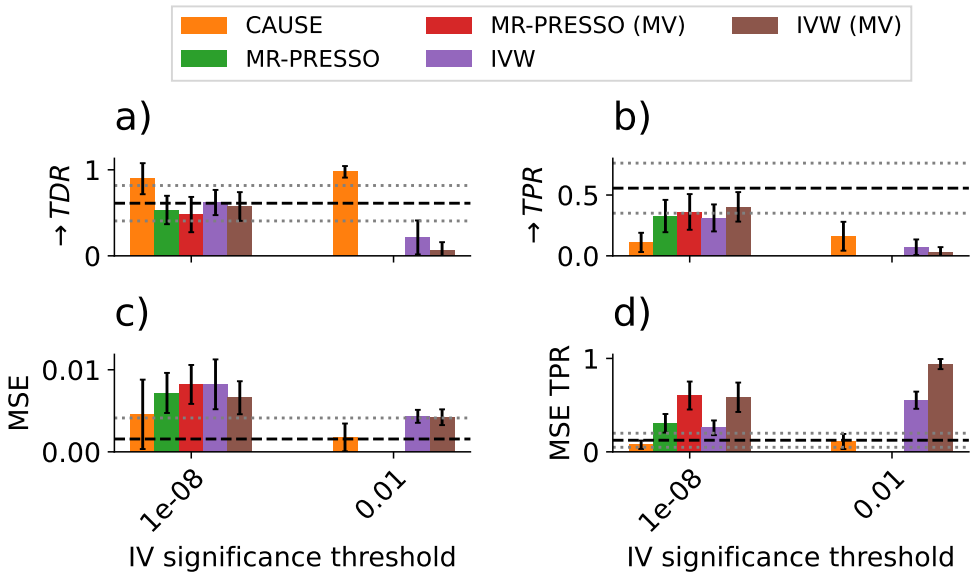

Fig. S2: **Performance of orientation and causal effect size estimation of CI-GWAS compared to Mendelian randomization methods in a simulation study using UK Biobank data.** At two instrument variable (IV) selection thresholds, (a) shows the true discovery rate (TDR) for the direction of causal relationships, (b) shows the mean squared error (MSE) for relations where a effect size was inferred. Dashed lines give the CI-GWAS value at $\alpha = 10^{-4}$ and 1SD. A detailed description of all calculations and MR methods used is given in the Methods section. MV refers to multi-variable. Note that for MR-PRESSO, a large number of selected instrument variables prevented the algorithm from returning estimates for threshold 0.01. Error bars give the SD across simulation replicates. No multiple testing correction was performed for any methods.

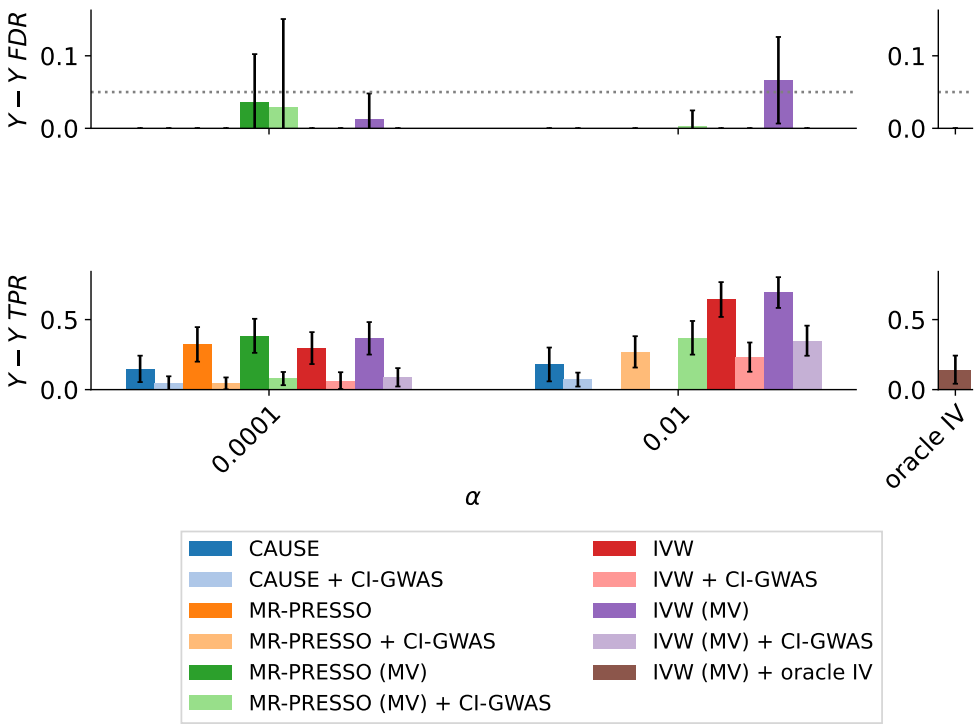

Fig. S3: **Comparison of Mendelian randomization methods when selecting instrument variables from genome wide significant results or from the cuda-skeleton step of CI-GWAS in a simulation study using UK Biobank data**. True positive rate (TPR) and false discovery rate (FDR) of trait-trait adjacencies for two values of alpha for different MR methods for simulated data based on UK Biobank data. Instrument variables were selected from genome wide significant results at an alpha threshold, or from the cuda-skeleton step of CI-GWAS (which is denoted with "+ CI-GWAS" in the legend), or from the true causal markers ("+ oracle IV" in the legend). Bonferroni correction was applied to all MR results except for CAUSE.

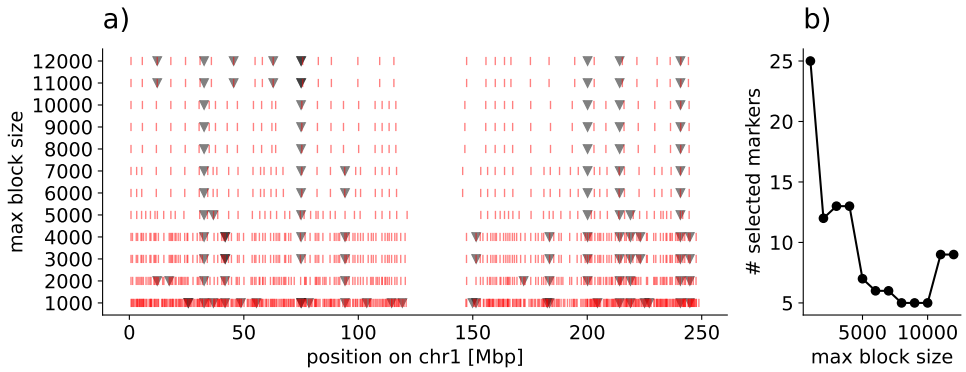

Fig. S4: **An example of the effect of maximum marker block size on the cuda-skeleton results.** The results of the cuda-skeleton are shown across various sets of block definitions generated with model parameter max block size s_{max} ranging from 1000 to 12000 markers. (**a**) gives the maximum block 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 within those blocks, 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.

exposure	outcome	*ACE*
BMI	WHR	0.483121
HT	WHR	-0.040955
HbA1c	BMI	0.191941
HT	BMI	-0.086164
BMI	HbA1c	0.233127
CAD	HbA1c	0.109513
HT	HbA1c	-0.022657
ALC	HbA1c	0.110836
HbA1c	CHOL	-0.055265
TRIG	CHOL	0.288243
HT	CHOL	-0.035631
TRIG	HDL	-0.399628
BMI	LDL	-0.013505
HbA1c	LDL	-0.064891
CHOL	LDL	1.037863
TRIG	LDL	0.268358
HT	LDL	-0.034091
WHR	AT	0.076149
BMI	AT	0.073726
T2D	AT	0.037998
ALC	AT	0.029308
WHR	CAD	0.099809
BMI	CAD	0.102236
HbA1c	CAD	0.115425
WHR	ST	0.031188
HbA1c	ST	0.047904
AT	ST	0.012000
CAD	ST	0.098009
T2D	ST	0.040488
ALC	ST	0.033585
WHR	T2D	0.148910
BMI	T2D	0.220579
HbA1c	T2D	0.461635
GLU	T2D	0.343260
CAD	T2D	0.105516
HT	T2D	-0.021061
BMI	HT	-0.086290
HbA1c	HT	-0.028711
TRIG	HT	-0.034096
BMI	ALC	0.138670
HbA1c	ALC	0.148963
WHR	SMK	0.043535
BMI	SMK	-0.036940
HbA1c	SMK	0.125995
TRIG	SMK	0.061867
HT	SMK	-0.027955

Table S3: Estimated average causal effects (ACE) for 17 traits in the UK Biobank. CI-GWAS was run with $\alpha = 10^{-4}$. Three letter codes give the trait name, with full description in Table S1. ACE values represent the expected effect of an intervention to alter the exposure by 1 SD.

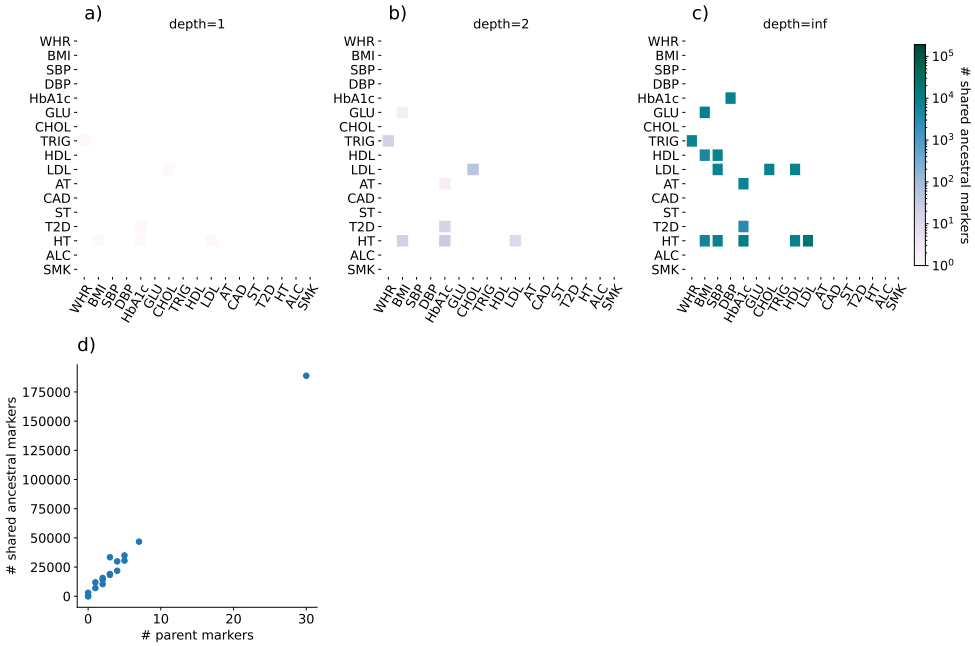

Fig. S5: **Numbers of shared ancestral markers at different graph depths in the causal skeleton in the UK Biobank data.** We inferred the skeleton using cuda-skeleton with $\alpha = 10^{-4}, l_{max} = 3, l_{max} = 14$. (**a-c**) Each matrix entry shows the number of markers that are connected to both corresponding traits by at least one path of at most length equal to the given depth in the inferred skeleton, with empty cells equal to zero. (**d**) Each point shows for one trait its number of parent markers vs. the total number of ancestral markers that it shares with other traits at infinite depth. Three letter codes give the trait name, with full description in Table S1.