

Graphical analysis for phenome-wide causal discovery in genotyped population-scale biobanks

David Amar ^{1,2}, Nasa Sinnott-Armstrong ³, Euan A. Ashley ¹ & Manuel A. Rivas ²✉

Causal inference via Mendelian randomization requires making strong assumptions about horizontal pleiotropy, where genetic instruments are connected to the outcome not only through the exposure. Here, we present causal Graphical Analysis Using Genetics (cGAUGE), a pipeline that overcomes these limitations using instrument filters with provable properties. This is achievable by identifying conditional independencies while examining multiple traits. cGAUGE also uses ExSep (Exposure-based Separation), a novel test for the existence of causal pathways that does not require selecting instruments. In simulated data we illustrate how cGAUGE can reduce the empirical false discovery rate by up to 30%, while retaining the majority of true discoveries. On 96 complex traits from 337,198 subjects from the UK Biobank, our results cover expected causal links and many new ones that were previously suggested by correlation-based observational studies. Notably, we identify multiple risk factors for cardiovascular disease, including red blood cell distribution width.

¹Center for Inherited Cardiovascular Disease, Stanford University, Stanford, CA, USA. ²Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ³Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA. ✉email: mrivas@stanford.edu

Causal inference from observational data is a fundamental objective that has been receiving increasing attention in multiple domains including biology, epidemiology, and economics. Graphical models are a cornerstone of causal inference as they explicitly describe the generating process of the observed data. These models contain functions that describe how values are assigned to each variable, possibly depending on the values of other observed or unobserved variables. These dependencies can be summarized in a directed graph, where an edge $X \rightarrow Y$ means that the function that determines the value of Y depends on X 's value. If the graph is acyclic, the joint distribution of the data can be represented as a Bayesian network (BN) that specifies the conditional probabilities of nodes given their parents¹.

Causal discovery is a subfield of causal inference that focuses on finding evidence in data for the existence of a causal path between two or more variables^{1,2}. This is an essential preliminary step as it can be used to justify the assumptions made by statistical analyses. Algorithms for causal discovery identify patterns of conditional independencies (CI) with theoretical justification for refuting candidate models that are unlikely to have generated the observed data. This process requires two assumptions (1) graphical d-separation and the *Causal Markov Condition* (CMC), and (2) *Causal Faithfulness Condition* (CFC)^{2,3} (see Supplementary Note 1 for formal definitions). CMC states that whenever a pair of variables X and Y are separated in the graph given a set Z , then X and Y are conditionally independent given Z in every compatible distribution. CMC has been proven to hold in acyclic models and in linear models with cycles (also called feedback loops)⁴. Some results support CMC in other cyclic cases^{4,5}. CFC deals with the opposite direction: it assumes that a conditional independence (CI) in the observed distribution entails separation in the graph. CFC has theoretical justification in that the set of models that do not satisfy it are extremely unlikely (i.e., have a zero Lebesgue measure^{2,6}).

In genetics, modern population-based cohorts often aggregate large datasets with extensive phenotypic and genotypic data of the same subjects. Due to their size and depth, these datasets offer new opportunities for discovery and inference of causal relationships between traits. Consequently, a plethora of methods have been suggested for causal inference using genetic data as instruments^{7–12}. Most methods employ a graphical model called *Mendelian Randomization* (MR) in which for a given pair of phenotypes (X, Y) the effect sizes of the variants of X with both phenotypes are analyzed to estimate the causal effect of X on Y . When the effect sizes are estimated in the same dataset, we denote the analysis as a *single-sample* MR. When a different dataset is used to estimate the effect sizes of the exposures, we called it *two-sample* MR, which assumes that the two populations are compatible¹³. Standard MR methods assume linear effects and report a summary of a linear fit. IVW regression¹⁴, for example, uses inverse variance weights to average the causal estimates of the instruments⁹. MR assumes that the genetic variants are independent of confounders that affect both phenotypes, and that there is *no horizontal pleiotropy*: the instruments affect Y only through X ^{11,15}. These are strong assumptions that cannot be justified from the data when analyzing X and Y alone, especially if the genetic variant directly affects both phenotypes (which we call direct horizontal pleiotropy)¹⁵. This is exacerbated when analyzing multiple phenotypes because the assumptions are made for each X, Y pair. Implicitly, MR also assumes that the graph is acyclic. However, this problem can be mitigated in case-control situations, where the effects of the instruments on X are measured using the controls only⁹. Nevertheless, detecting cycles is particularly salient in the context of population biobanks, where medications, lifestyle changes, and variable temporal

dynamics often confound the causal directions between measurements¹⁶.

Several methods have been proposed to address some of the limitations above. MR-Egger can model horizontal pleiotropy under the assumption that the effects of the instruments on the exposure and the outcome are independent¹⁷. MR-PRESSO accounts for horizontal pleiotropy by correcting for variants with outlier effects⁸. However, the assumption that outliers are not proper instruments and should be adjusted for may not hold in practice. Latent-causal variable analysis (LCV)⁷, assumes that there is a latent variable that mediates the genetic correlation between X and Y and then compares the genome wide effect sizes of association with both traits against each other to assess if one phenotype is fully or partially *genetically causal* for the other. LCV assumes acyclicity and does not estimate the causal effects. Regardless of recent progress, extant methods focus on inference for a single trait pair from the marginal summary statistics and are limited for causal discovery by their unidentifiable assumptions.

Other causal inference methods in genomics have been proposed to either mitigate the limitations of MR or to model larger graphical structures. For example, network MR analyzes an exposure and an outcome together with a mediator¹⁸. Multi-variate MR can model multiple exposures jointly for the same outcome¹⁹. While these methods can provide accurate results, they require adding assumptions to the standard MR model. CAUSE is a recent extension of MR that uses genome-wide summary statistics to model causal effects while accounting for pleiotropy²⁰. Another type of algorithms address gene network inference by joint analysis of genetic variants and gene expression data in order to learn a large-scale graphical model with causal links among genes^{21–23}. Recently, Howey et al. explored similar methodology as an alternative for MR²⁴. They showed that learning BNs among phenotypes while including genetic variants as anchors can improve upon MR. However, this method does not provide a clear way to select edges as it simply weighs all phenotype pairs by the proportion of times they were connected in bootstrap repeats.

In this work, we leverage the theoretical framework of causal discovery to enhance and streamline MR-based analysis. Our flow, Causal Graphical Analysis Using GENetics (cGAUGE) first identifies unique CI patterns in the data and then uses them for: (1) filtering genetic instruments for downstream MR analysis, and (2) Exposure-based Separation (ExSep): an algorithm for causal discovery that does not require selecting genetic instruments in advance. The theoretical justification of these algorithms holds even in the presence of unobserved confounders and cycles. We present extensive simulations to illustrate how cGAUGE improves upon MR and BN methods by reducing their empirical false discovery rate by up to 30%. We apply cGAUGE to 67 complex traits (including 41 biomarkers) and 29 diseases using data from 337,198 subjects from the UK Biobank^{25,26}. We find many expected causal links and new ones that were previously speculated in correlation-based observational studies. These new discoveries include causes of behavioral phenotypes and multiple risk factors for cardiovascular disease, including red blood cell distribution width (RBW), which is discovered using our ExSep (non-MR) test.

Results

cGAUGE: A novel pipeline for causal discovery using genetic instruments. We present a new pipeline, cGAUGE: Causal Graphical Analysis Using GENetics. cGAUGE is a set of tools that utilize CI tests for improving causal inference among traits using genetic variables (see Fig. 1 for an overview, and Supplementary

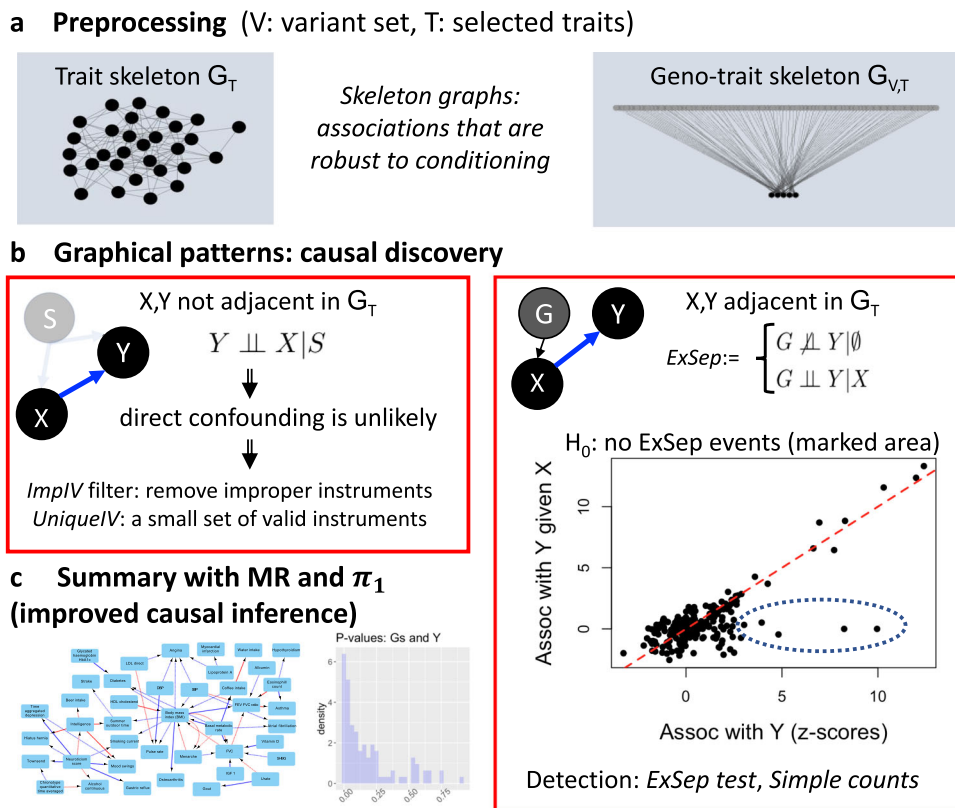


Fig. 1 The cGAUGE workflow for causal discovery. We analyze genetic and phenotypic data collected from independent subjects. **a** We first preprocess the data to infer skeleton graphs: graphs that represent associations that are robust to conditioning. Based on causal inference theory^{1,2}, surviving associations contain the subset of true causal links. We learn two skeletons: G_T among the phenotypes (ignoring the genetic data in the process), and $G_{V,T}$ between the variants and the phenotypes. **b** We then analyze the edges and the non-edges of G_T separately. We present methods that use $G_{V,T}$ to filter out improper instruments (ImpIV) or identify unique proper instruments (UniqueIV). While their theoretical justification is pertinent to G_T non-edges, we illustrate using simulations how they reduce the empirical FDR when applied to all phenotype pairs. For G_T edges we present an analysis based on ExSep events: associations between genetic variables and a trait Y that “disappear” once conditioned on a new phenotype X (i.e., $p > p_2$). Under our *local faithfulness* assumption these patterns are evidence for a causal link from X to Y . The ExSep model selection test is a method to analyze all genetic variables under the null hypothesis of no ExSep events. **c** Finally, we utilize our results for improved inference using Mendelian Randomization (MR) and also utilize the π_1 estimate for each exposure–outcome pair. This score quantifies the consistency of the associations between the exposure’s instruments and the outcome, which can be used to flag potential false positive causal links.

Note 1 for a formal explanation). cGAUGE takes as input the individual level data of a population-based biobank, a CI test (e.g., using linear or logistic regression), and two p -value thresholds: p_1 for rejecting the null of CI, and p_2 for accepting it ($p_2 > p_1$, values in between are considered unreliable). While standard statistical tests are not designed to accept null hypotheses, this is a standard assumption made by causal discovery algorithms for detecting independence^{1,2}. Alternatively, cGAUGE can take the summary statistics of all marginal and conditional tests. These are assumed to be adjusted for exogenous variables including sex, age, and genetic principal components (we used the top five by default).

cGAUGE has two types of output. First, it provides a set of filtered genetic variables that can be used as instruments for an MR analysis for a given pair of phenotypes. While the theoretical justification of these filters is pertinent to a subset of the trait pairs, we illustrate using simulations that they work well in practice for all pairs. Moreover, cGAUGE flags the trait pairs that do not have these guarantees to help users in interpreting the results. Second, it identifies phenotype pairs with evidence for causal interactions based on a new statistical test that does not require setting a significance cutoff for instruments. We now give an overview of the different steps.

We first seek associations in the data that are robust to conditioning. This is a standard preprocessing step as these

associations contain the correct (direct, non-mediated) causal links as a subset^{1,2}. The results are summarized in graphs called *skeletons*: (1) G_T for all phenotype pairs that are significantly associated with $p < p_1$ even when conditioned on pairs of other phenotypes, and (2) $G_{V,T}$ for genetic variables vs. phenotypes, created by excluding (G, X) associations for which there exists a phenotype Y such that conditioning upon it (in addition to the exogenous variables) results in $p > p_2$.

G_T separates the phenotype pairs into *edges* and *non-edges*. For non-edges we present two practical algorithms to filter the instruments of an exposure X when analyzing an outcome Y . The first is based on the observation that given the phenotypes that separate X and Y while learning G_T , their adjacent genetic variants in $G_{V,T}$ are not valid instruments when analyzing X and Y (see Theorem 2.1 in Supplementary Note 1). We call this filter *ImpIV* as it removes improper instruments. However, note that (1) there is no guarantee that the remaining instruments are valid and (2) by definition, for G_T edges, *ImpIV* does not change the set of instruments.

The second filter is based on the observation that genetic variants that are linked only to X in $G_{V,T}$ are valid instruments for analyzing X and Y (see Theorem 2.2 in Supplementary Note 1). We call this filter *UniqueIV*, as it finds unique skeleton-based instruments. As this set of instruments is identified for an

exposure regardless of the outcome, this set can be used for all MR analyses (i.e., including for G_T edges). However, note that UniqueIV may remove most if not all of X 's instruments, potentially limiting the statistical power of the downstream MR analysis.

Given a set of instruments discovered by taking all GWAS results of X , or after applying one of our filters above, the causal discovery process now relies on an MR analysis. However, we also compute π_1 : the proportion of non-null p -values (i.e., when examining the p -values of the association of X 's instruments with Y) under the assumption that the p -values follow the two-groups mixture model²⁷: p -values or their inverse normal z -score follow a mixture distribution of nulls and non-nulls (see Supplementary Note 1). π_1 can be estimated using empirical Bayes approaches²⁸ and it directly measures the association significance consistency while avoiding some parametric assumptions made by MR (e.g., that all instruments are pertinent to the same linear effect). Our simulations below illustrate how the π_1 estimates can be used to filter out false positives.

For an edge (X,Y) in G_T we show that if there are cases of genetic variants that are marginally associated with Y but are independent of Y given X , then this serves as evidence for X being a cause of Y . We denote this pattern as *ExSep*. The proof holds under the faithfulness assumption. However, we show that it is still valid even under a weaker and more realistic assumption that we call *local faithfulness*: the assumption that an emerging CI reflects having at least one pathway that is blocked in the true causal diagram (Supplementary Note 1, Theorem 2.1). We examined two approaches to test for ExSep events for a given (X,Y) pair: (1) *Naive*: using a simple threshold for the number of events with $CI(G,Y)$ with $p < p_1$ and $CI(G,Y,X)$ producing $p > p_2$, and (2) *Model selection (MS) test*: an approach that tests the null hypothesis of no class of ExSep variants.

Denote the z -scores of the associations of all genetic variables with Y as \mathbf{z}_1 and all the z -scores with Y given X as \mathbf{z}_2 . Note that these are the inverse normal scores of p -values and not effects sizes. Each z -score can represent a *null* case of no association or a *non-null* case of a true association. Thus, this fits the two-groups model discussed above, which we assume for simplicity is a mixture of two Gaussians^{29,30}. We can therefore model the joint distribution of \mathbf{z}_1 and \mathbf{z}_2 as a mixture of four bivariate normal distributions corresponding to all four combinations of null and non-null cases. However, under the null hypothesis that there are no non-null \mathbf{z}_1 cases whose \mathbf{z}_2 statistic is null, we can model the data using a mixture of three Gaussians only. Assuming that the marginal two-groups models are known and fixed, the unknown parameters of the null and non-null models are the correlations between \mathbf{z}_1 and \mathbf{z}_2 within each bivariate normal distribution, and the prior probability of each cluster. We use a grid-search heuristic to fit these models and compute their likelihoods. These are then used to test the null hypothesis using a likelihood ratio test (see Supplementary Note 1).

Simulations. Consider the graph in Fig. 2a. There is no causal link between the two traits X and Y , but they are both affected by an unobserved confounder U . X has 10 binary instruments and U has 20. We tested the performance of MR-Egger, IVW, and MR-PRESSO on 100 datasets simulated from this graph (2000 samples each), with summary statistics computed using linear regression and instruments selected at $p < 10^{-04}$. All causal quantities were sampled independently from the same uniform distribution $U_c = U[(-0.9, -0.1) \cup (0.1, 0.9)]$, and each instrument was generated randomly with probability p_G , sampled from $U_g = U[0.05, 0.4]$ for each instrument. U, X, Y were all generated with standard normal errors. All three methods tended to erroneously

predict a causal link from Y into X (e.g., at $p < 0.01$). Moreover, the tests for heterogeneity (IVW) or pleiotropy (MR-Egger, MR-PRESSO) produced insignificant results ($p > 0.2$ in $>80\%$ of the cases), illustrating that utilizing these additional tests could not salvage the analysis from making errors. In contrast, applying UniqueIV before MR results in no causal links between X and Y at the same 0.01 significance.

We also simulated data from a simpler MR model with a single genetic variant G with either an $X \rightarrow Y$ link or a feedback loop (see Fig. 2b). Over 100 repeats, we tested the association between G and Y with and without conditioning on X . The results had multiple cases in which the association between G and Y was not detected (at $p < 0.001$), illustrating how faithfulness may not hold in finite samples. This also demonstrates why local faithfulness is more realistic: it fits the graphs even if the CI test for G and Y given X is insignificant and G and Y are associated marginally.

We next simulated single-sample synthetic data of larger graphs over 15 continuous traits (with standard normal noise). Each trait had randomly selected ingoing and outgoing neighbors such that the expected in/out degree was set to $\text{deg} \in (1, 1.25, 1.5, 1.75, 2)$. We then added k binary instruments for each trait, with k randomly selected from $U[10, 20]$. To add horizontal pleiotropy, for each instrument we decided whether it should have additional adjacent traits (in the true causal graph) at random with probability $p_{\text{pleio}} \in (0, 0.1, 0.2, 0.3, 0.4)$, and if so, we added between 1 and 10 additional adjacent traits (selected randomly and uniformly). To summarize, as deg increases the generated graph has more cycles, and as p_{pleio} increases we are more likely to have violations of the MR assumptions. When generating datasets, causal quantities and binary instruments were generated using the U_c and U_g distributions above. For more details on how to simulate the data see Supplementary Note 1.

Figure 3 shows examples of the effects of our instrument filters from graphs generated with $\text{deg} = 1.5$. To test the effects of our filters even on skeleton edges, the MR methods were run on all trait pairs using a 10% BY FDR threshold³¹. The figures show the number of discoveries and their empirical FDR. MR-Egger is not presented as it consistently had greater empirical FDR values. The results of all MR methods and all possible combinations of deg , p_{pleio} , p_1 , p_2 are available in Supplementary Data 1 and 2. All MR-methods, when run on the set of all exposure-associated instruments, had multiple cases of unreasonably high mean empirical FDRs, which was as high as 40%. Both ImpIV and UniqueIV consistently lowered the empirical FDR of IVW and MR-PRESSO. ImpIV's effect was moderate in many cases, whereas UniqueIV had reasonable empirical FDR ($<15\%$) whenever $p_1 \leq 0.001$. Supplementary Fig. 1 shows the same simulations as in Fig. 3 but with a 1% FDR cutoff. The empirical FDR of the MR methods still remains unreasonably high with high horizontal pleiotropy ($>25\%$). It also illustrates that using UniqueIV with MR-PRESSO is conservative as its empirical performance remains far below the FDR threshold used.

Moving to non-MR methods, we tested the performance of the ExSep-based approaches. Supplementary Fig. 2 shows an example (see Supplementary Data 1–4 for the full results). As in the MR analysis above, discoveries were made at 10% BY FDR adjustment. The results show that the naive analysis had poor performance when $p_{\text{pleio}} > 0.1$. The MS test had a greater number of discoveries and a lower empirical FDR. As shown in Supplementary Data 3, there were a few cases in which the MS test had an FDR between 15% and 20%.

Finally, we examined the simple π_1 estimates as a method for causal discovery. For a given threshold t we computed the number of discoveries and empirical FDRs for trait pairs with $\pi_1 > t$ but also the false positive rate among pairs with $\pi_1 < t$

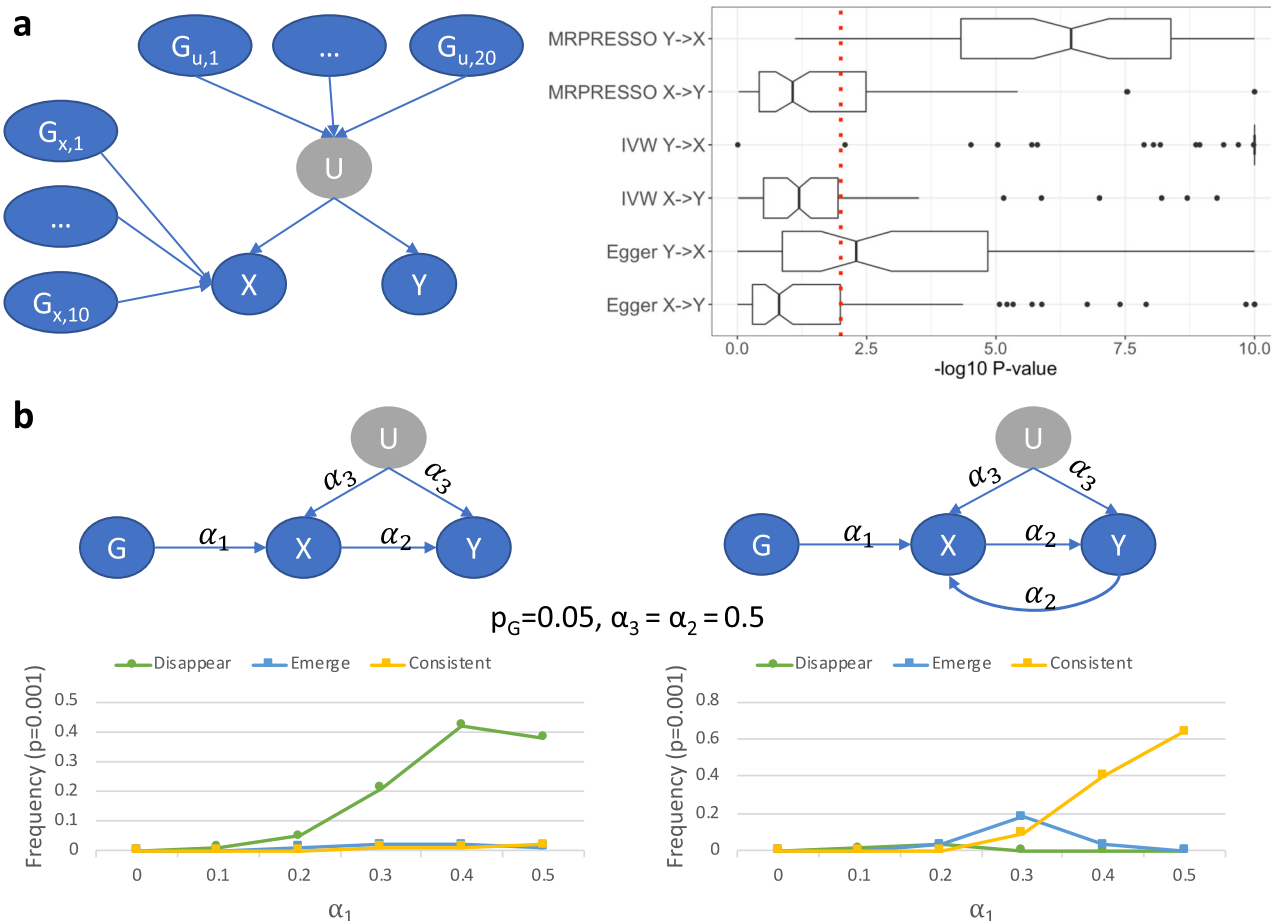


Fig. 2 Simulations of simple graphs illustrate some limitations of MR and the faithfulness assumption. **a** When simulating from linear models from the presented graph ($n = 2000$ samples), MR methods tend to erroneously predict causal links between X and Y. The left panel shows the graph, where the gray node U is an unobserved common cause (confounder) of X and Y. The right panel shows the distribution of the ($-\log_{10}$) p-values of different MR methods when analyzing X and Y. Each boxplot shows the median, and first and third quartiles. The whiskers extend from the hinge to the largest and lowest values, but no further than $1.5 \times$ (the inter-quantile range). As input for MR, summary statistics were computed using linear regression and instruments were selected using a $p < 10^{-04}$ cutoff. **b** Simulations from simple MR models with a single binary instrument show violations of faithfulness in finite samples. Each line presents the frequency of different dependency patterns between the simulated genetic instrument G and the phenotypes Y and X. In all simulated cases G and X were significantly associated at $p < 0.001$, and are thus not shown. Disappearing associations: G and Y are associated marginally, but become independent when conditioned on X ($p > 0.1$). Emerging associations: G and Y are independent but become associated when conditioned on X. Consistent: G and Y are significantly associated with and without conditioning on X. Even though under faithfulness G and Y should be associated with and without conditioning on X (in both graphs, X acts as a collider on the path from G to Y through U), we see mixed results. Nevertheless, the detected associations occur only when a non-blocked path exists between G and Y, satisfying our refined local faithfulness assumption.

(see Supplementary Data 5 and 6). The results show that trait pairs with $\pi_1 > 0.9$ had low empirical FDRs ($< 10\%$ in almost all cases). The few cases with FDR $> 15\%$ were all when p_2 was set to too low values ($p_2 < 10^{-4}$). Trait pairs with $\pi_1 < 0.3$ tended to be false positives (FPR $> 90\%$ in most cases). However, note that for π_1 to be valid for discovery in practice, we must add an assumption that pleiotropic confounding occurs in low percentages when examining the variants associated with an exposure. Nevertheless, extremely low π_1 scores can still be used for flagging potential false positives without additional assumptions.

We also compared cGAUGE to two recent methods: (1) CAUSE²⁰, and the BN methods discussed in Howey et al.²⁴. We used our simulated data with $p_{\text{pleio}} = 0$ or $p_{\text{pleio}} = 0.3$, and $\text{deg} = 1.5$ (see “Methods” section for details). The comparison with CAUSE is presented in Supplementary Data 7. Overall, both CAUSE and cGAUGE improve upon other methods in terms of FDR control. However, we observed two advantages of cGAUGE: (1) it tends to have greater power, especially for low levels of horizontal pleiotropy, and (2) out of all compared methods only

UniqueIV with MR-PRESSO keeps the empirical FDR lower than the predefined threshold in all cases. The latter is also correct in terms of worst-case performance (i.e., maximal empirical FDR observed over all simulated datasets), whereas CAUSE and UniqueIV with IVW can have 20% or greater.

When testing BNs with $p_{\text{pleio}} = 0$ the mean empirical FDR over simulated datasets was 8–8.5% among the top 10 predicted links, but was between 16% and 21.4% among the top 20 predicted links. With $p_{\text{pleio}} = 0.3$ the mean empirical FDR of either the top 10 or top 20 predicted causal links was $> 31\%$. These results again illustrate how extant methods are sensitive to high pleiotropy levels either from mediation (e.g., with $p_{\text{pleio}} = 0$) or horizontal pleiotropy (e.g., with $p_{\text{pleio}} = 0.3$).

Results on the UK Biobank data. We applied cGAUGE to the 96 traits in Supplementary Data 8 using UniqueIV and MR-PRESSO as the base MR analysis at 10% BY FDR, and with a $\pi_1 > 0.25$ cutoff. We tested $p_1 = 1 \times 10^{-6}$, 1×10^{-7} , or 1×10^{-8} , and $p_2 =$

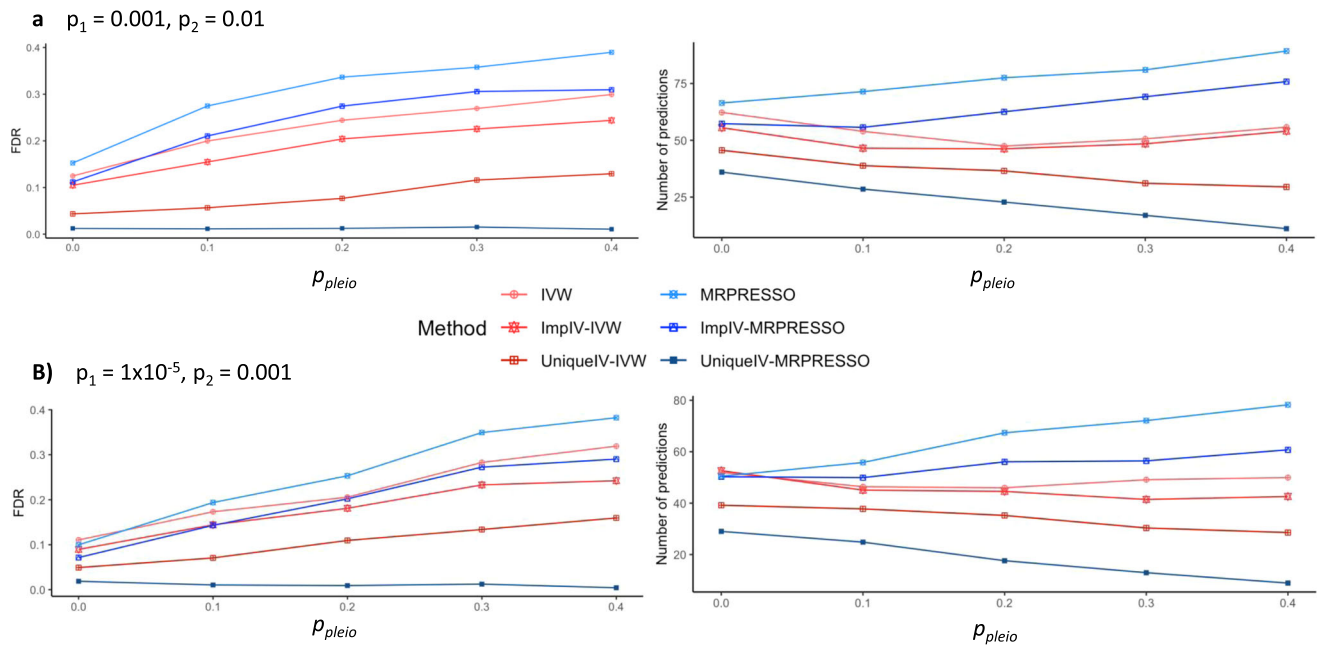


Fig. 3 Mean number of discoveries and empirical false discovery rates (FDR) of Mendelian randomization methods in simulated data from graphs with 15 continuous traits. The underlying causal diagram was generated such that the expected in- and out-going degrees of the traits were 1.5. All simulated graphs contained cycles. For each trait we added between 10 and 20 binary instruments (uniformly, i.i.d.). To add horizontal pleiotropy, for each instrument we decided whether it is horizontally pleiotropic or not with probability p_{pleio} , and if so, we added between 1 and 10 links into additional traits (uniformly, iid). When generating datasets, the traits had standard normal noise, causal quantities were randomly and uniformly sampled such that their absolute value was between 0.1 and 0.9, and binary instruments were generated randomly with a probability between 0.05 and 0.4. The plots show the mean results of the simulations for different p_{pleio} values (e.g., the mean of the empirical FDR over the simulated graphs). Discoveries from each statistical test were done at a 10% significance level after adjusting for FDR using the BY algorithm. When two methods have a similar empirical FDR, greater number of predictions correspond to greater power. **a** Results with $p_1 = 0.001$ and $p_2 = 0.01$. **b** Results with $p_1 = 1 \times 10^{-5}$ and $p_2 = 0.001$. MR-Egger is not presented as it consistently had greater empirical FDR values than the other methods.

0.01, or 0.001. These ranges are in line with MR publications for p_1 ^{32,33}, and settings of causal discovery algorithms for p_2 ³⁴. Comparing the choices for p_1 and p_2 , the results are generally robust especially with $p_2 = 0.001$ (>0.7 Jaccard coefficient, Supplementary Fig. 3). All trait pair results and the discovered instrument sets are available in Supplementary Data 8–16.

Figure 4a shows G_T inferred with $p_1 = 1 \times 10^{-7}$, resulting in 669 edges and 95 nodes. Clustering using the MCODE algorithm^{35,36} detected groups of densely connected related phenotypes. Changing p_1 to 1×10^{-6} resulted in a similar network and clusters (Supplementary Fig. 4). Analyzing $G_{V,T}$, we observed that up to 42.5% of the original GWAS results can be filtered, depending on p_1 and p_2 (Fig. 4b). These surviving variants (per GWAS) are more likely to contain true direct causal loci as compared to those excluded. In addition, when using $p_2 = 0.001$ all p_1 values result in a similar $G_{V,T}$.

With $p_1 = 1 \times 10^{-7}$, $p_2 = 0.001$, we identified 290 causal links using our MR analysis (see Supplementary Data 9). Figure 5 shows a subset of the network that focuses on causal links from biomarkers into diseases and other phenotypes. Both LDL and lipoprotein A increase the risk for heart disease (e.g., angina $q < 10^{-17}$, 0.58 log odds ratio (LOR)/mmol/L and 3.8×10^{-3} LOR/nmol/L, respectively)^{37,38}. Other expected links include urate to gout (LOR/ $\mu\text{mol/L}$ 0.025, $q = 1.97 \times 10^{-60}$, MS test $p = 9.64 \times 10^{-54}$)³⁷, body mass index (BMI) to pulse rate ($\beta = 0.43$, $q = 1.85 \times 10^{-5}$)³⁹, BMI to diabetes (LOR/kg/m² 0.25, $q = 4.28 \times 10^{-15}$)⁴⁰, and basal metabolic rate (BMR) to diabetes (LOR/kJ, 4.8×10^{-4} , $q = 0.003$)⁴¹. The network also contains causal links into behavioral phenotypes. For example, intelligence has a negative effect on mood swings ($\beta = -0.15$, $q = 0.001$, MS test $p < 1 \times 10^{-100}$), whereas depression has a positive effect ($\beta = 0.5$, $q = 0.05$).

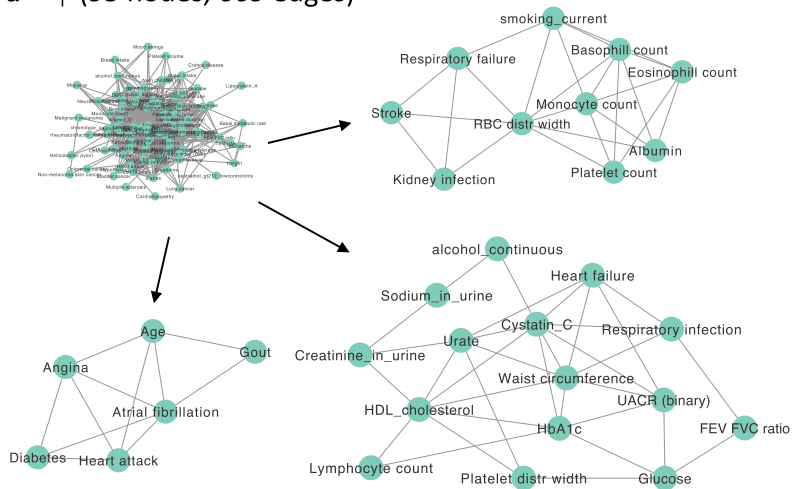
While many links in Fig. 5 are expected, we detect multiple interactions that were previously suggested in observational studies. For example, we identify risk increasing factors for atrial fibrillation including blood pressure (SBP, LOR/mmHg 0.02, $q = 0.098$)⁴², and BMR (LOR/kJ 7×10^{-4} , $q = 5.9 \times 10^{-8}$), which is in line with previous reports about correlations between the disease and metabolic syndrome⁴³. Eosinophil count has risk increasing links to asthma (LOR/10⁹ cells/L 5.4 $q = 4.4 \times 10^{-29}$, MS test $p = 3.46 \times 10^{-57}$)^{44,45}, and hypothyroidism (LOR/10⁹ cells/L 1.4, $q = 0.1$)⁴⁶. The network also suggests that albumin increases FEV/FVC ratios ($\beta = 0.002$, $q = 0.03$). Serum albumin tends to be greater in normal individuals when compared to COPD patients or smokers, which tend to have lower FEV/FVC ratios^{47,48}.

Figure 6 shows the top 40 pairs that had insignificant MR results and significant ExSep MS test results (all significant results are available in Supplementary Data 10). The network presents intelligence as a main hub that affects the Townsend deprivation score, age of menarche, smoking status, forced expiratory volume (FVC), angina, and height, which is also a feedback loop. Height-related causal links may result from temporal information, as discussed previously in O'Connor and Price⁷. For example, positive correlation between height and intelligence is a well-recognized phenomenon in children⁴⁹, and nutritional status as a child is known to affect menarche age, height, lung capacity, and BMR^{50–53}. The network also suggests RBW as a cause of myocardial infarction. The link between RBW and cardiovascular disease has observational evidence⁵⁴.

Discussion

In this work, we presented methods that utilize CI tests to enrich the causal analysis tool set of genetic biobanks. Our ImpIV and

a G_T (95 nodes, 669 edges)



b $G_{V,T}$ vs. GWAS results

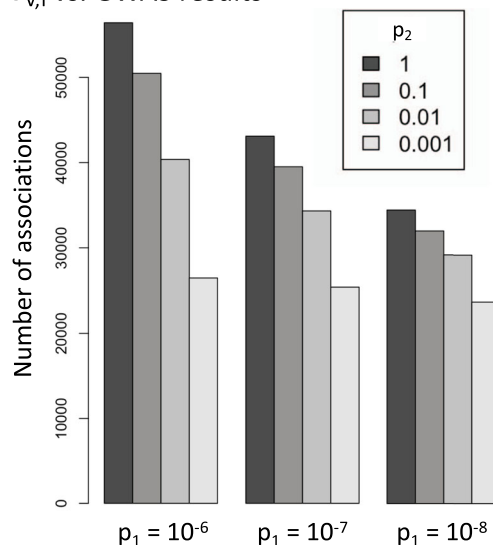


Fig. 4 Inferred skeletons that represent associations that are robust to conditioning. **a** The skeleton inferred among the phenotypes (G_T). The edges represent phenotype pairs that remain associated at $p_1 < 1 \times 10^{-7}$ when conditioned on other phenotypes. In this computation for a pair of variables X and Y , we tested the association between X and Y conditioning on all other phenotypes or all other phenotype pairs. Arrows point out to clusters detected with MCODE. **b** The effect of conditional independence filtering on genome-wide association analysis (GWAS) results. Here we show the effect of removing variant-phenotype pairs (detected in a standard GWAS) for which there exists another phenotype whom conditioning upon renders the association insignificant ($> p_2$).

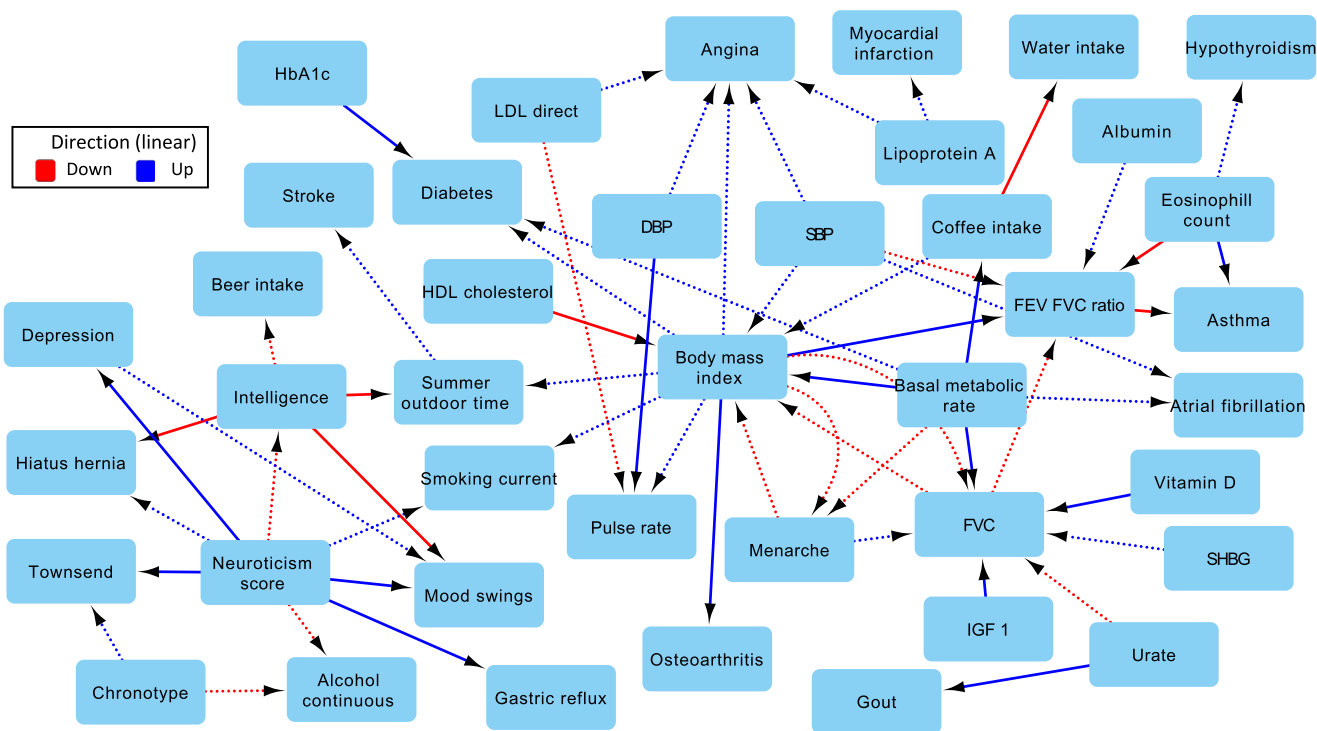


Fig. 5 cGAUGE MR analysis with $p_1 = 1 \times 10^{-07}$ and $p_2 = 0.001$. All presented links were detected at 10% BY FDR correction and had $\pi_1 > 0.25$. Solid arrows represent G_T edges (phenotypes whose correlation is robust to conditioning) and dashed arrows represent G_T non-edges. For simplicity, we excluded waist circumference and height, arrows out of diseases, and arrows into biomarkers. DBP diastolic blood pressure, SBP systolic blood pressure, HbA1c hemoglobin A1c levels, FEV forced expiratory volume, FVC forced vital capacity, IGF1 insulin like growth factor 1 levels, LDL direct low-density lipoprotein levels, HDL cholesterol high-density lipoprotein levels, SHBG sex hormone-binding globulin levels.

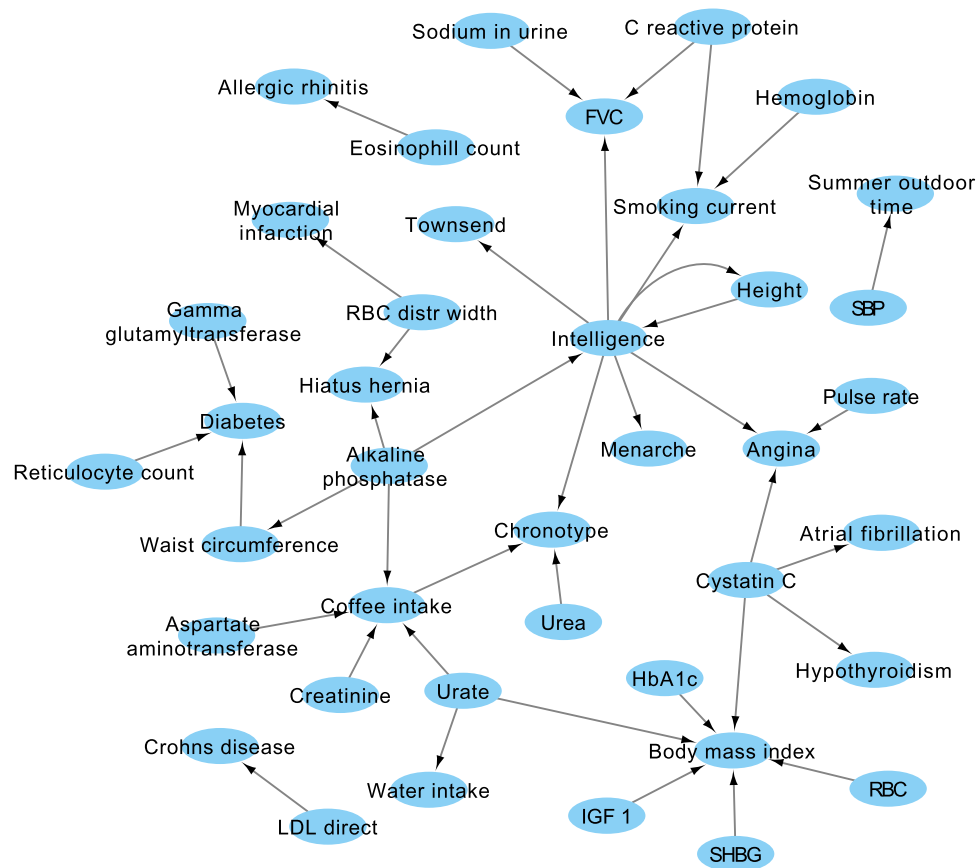


Fig. 6 The top 40 links discovered uniquely by the ExSep MS test. The null hypothesis is that there are no *ExSep* events. That is, that there are no genetic variants that become independent of the outcome when adjusted for the exposure. An arrow represents an ordered pair with evidence for causal interaction with a significant MS test p -value even at 1% FDR adjustment. The MS-test does not report the direction of the causal effect, but can be significant even if the causal interaction is not linear. All links in the figure were not discovered by the MR analysis at 10% FDR adjustment, and are thus not in Fig. 5. SBP systolic blood pressure, HbA1c hemoglobin A1c levels, FVC forced vital capacity, IGF1 insulin like growth factor 1 levels, LDL direct low-density lipoprotein levels, RBC red blood count, SHBG sex hormone-binding globulin levels.

UniqueIV filters highlight which genetic instruments to use for MR analysis. They both start with the genetic variables detected in the GWAS. ImpIV filters out improper instruments but has no guarantees for the validity of the remaining set. UniqueIV removes more instruments and has guarantees about the remaining set, but it may be small and therefore less powerful in subsequent statistical analyses. Our theorems on which these filters rely mainly pertain to trait pairs that can be rendered independent by conditioning on other traits. However, given that they can still highlight valid instruments in all cases, they provide an alternative to the current practice in the community of using the GWAS results without filtering instruments while relying on assumptions alone.

We also provide a non-MR analysis for causal discovery based on the *ExSep* pattern: cases in which the association between genetic variables and the outcome are nullified when conditioned on the exposure. We provide a statistical test for the null hypothesis that no *ExSep* events exist when analyzing an exposure–outcome pair, which utilizes the set of all independent genetic variables and does not require selecting the instruments in advance. Our test is based on a grid search and future work can test if alternative optimization techniques improve the power of the test. In addition, unlike MR, this test does not infer the causal effect size. Future work can address integrating the results from our different analyses including the traits skeleton, the MR analysis, and the *ExSep* test. A major challenge here is to maintain the validity of our theorems while considering the uncertainty of skeleton edges under the same optimization framework.

Our instrument filters and the skeleton inference processes require p -value thresholds for either accepting or rejecting the null hypothesis of (conditional) independence. This is a general limitation of causal discovery algorithms as hypothesis testing was not developed for accepting null hypotheses^{1,2,55,56}. Moreover, most causal discovery algorithms utilize a single p -value threshold and thus assume that for any statistical test they can decide between null and non-null. cGAUGE alleviates this issue in two ways. First, it leverages the large-scale nature of hypothesis testing with genetic variables using empirical Bayes techniques as these can sometimes justify inference about the proportion of null cases²⁷. Second, we use two p -value thresholds such that CI events are used only if an association was first discovered with $p < p_1$ and later had $p > p_2$ when adjusting for additional variables.

In simulations we show that our methods have substantially lower empirical FDR. Specifically, our UniqueIV filter can reduce the empirical FDR by up to 30%, whereas our *ExSep* MS test has reasonable empirical FDR in almost all tested cases. In contrast, we observe that MR-Egger and BN learning have unreasonable empirical FDRs even with zero levels of added horizontal pleiotropy. IVW and MR-PRESSO can also have >20% empirical FDR (at 1% or 10% adjustment) as added horizontal pleiotropy levels increase (e.g., when 10% of the instruments are horizontally pleiotropic). We also illustrate that using the MR internal tests for pleiotropy cannot salvage the analysis when an unobserved confounder is well explained genetically (Fig. 2a). Our simulations were done using a single-sample, and when applied to the two-sample

case, methods like MR-Egger and MR-PRESSO can have lower FDR. However, the observed empirical FDRs are still unacceptably high, and a two-sample MR can have additional challenges because of biases that lead to selection of improper instruments¹³. Finally, we observe that both cGAUGE and CAUSE reduce the empirical FDR as compared to all other methods. However, cGAUGE tends to have greater power, and it consistently keeps the empirical FDR in the desired level even in cases where CAUSE does not. These two methods, while being very different, represent a substantial progress in avoiding spurious results. Future studies can explore ways to integrate their ideas to further increase power.

We applied cGAUGE to a set of 96 phenotypes from the UKBB data. These were selected such that they cover many individuals and were not perfectly correlated as causal discovery may be invalid otherwise². cGAUGE reports hundreds of causal links, most of which are expected. However, many of the identified links are novel and confirm previous suggestions from epidemiological observational studies that reported correlations with no causal inference. Notable examples include links from blood pressure to atrial fibrillation, serum albumin to lung function, and RBW to cardiovascular disease. These are only a few examples and we provide the results for all pairs in the supplementary material. There are two important considerations when interpreting large-scale causal networks. First, while we adjusted for population structure using the top genetic principal components, and included the Townsend deprivation score to account for socio-economic status, there may still be some errors in the output networks due to statistical errors. For example, even in our simulations of data with limited pleiotropy levels, the empirical FDR was not zero. Second, some significant MR results were filtered out using our $\pi_1 > 0.25$ cutoff, including known false positives (e.g., HDL \rightarrow angina⁵⁷). π_1 quantifies the proportion of non-null exposure instruments that are associated with the outcome. Both in theory and in simulations, low π_1 suggests that the detected links may be false positives. Thus, it allows flagging problematic results.

Utilizing CI patterns is a unique property of our flow that is not covered by extant approaches that use genetic data. These tests require using the individual-level data and are thus not as easily available as the GWAS summary statistics that standard MR uses. However, if the summary statistics of these tests are provided, cGAUGE can be run without the individual level data. In our case, this amounts to all CI tests for roughly 50,000 genetic variants, which is a reasonable size dataset that can be shared by the community. Moreover, while our MR analyses in this paper are all based on single-sample MR, our UniqueIV and ImpIV filters provide static instrument sets that can be used and explored in future studies. Specifically, these can be used for two-sample MR, which requires learning the instrument set on one sample and estimating the causal effects on another. This analysis can help in reducing bias of estimated causal effects⁵⁸. The same methodology can be used to improve multivariate MR as it also requires precise instrument sets (e.g., by requiring that the instruments are not directly linked to confounders)¹⁹. In addition, our instrument sets can be used as weights when interpreting the GWAS results of an exposure. UniqueIV variants are more likely to be causal than ImpIV-only variants. Variants that are removed by ImpIV can be down weighted as cGAUGE identifies evidence for a path to some outcome that is not through the exposure.

Methods

UK Biobank data. We used 805,462 directly genotyped variants from 337,198 white British subjects from the UK Biobank^{25,26}. The MHC region was excluded (chromosome 6, positions 23–35M). Data were preprocessed as explained in ref. ⁵⁹ with a small change: we excluded variants with a MAF < 1%. 96 phenotypes (traits and diseases) were selected for the analysis (see Supplementary Table 8). These were selected to cover the phenotypes analyzed by O'Connor and Price⁷, but additional traits that had large sample sizes were added.

Biomarker data. Biomarker measurements from UK Biobank participants were adjusted for 83 covariates, including age, sex, their interaction, assessment centers, and technical factors⁶⁰, except that Townsend Deprivation Index and principal components of the genotyping matrix were not included in the regression. Residuals from these regressions were used for downstream analysis.

Single GWAS. Genome wide association analysis per phenotype was performed using PLINK (version 2.0a2)⁶¹. The baseline results for each each GWAS were adjusted for sex, age, and the top five genetic PCs. We also clumped the results using PLINK's greedy approach with the following parameters: $-\text{clump-p1 } 0.0001$, $-\text{clump-r2 } 0.1$, and $-\text{clump-kb } 500$.

Graph visualization. All networks were plotted using Cytoscape version 3.7.2^{62,63}.

Computing π_1 . When analyzing an exposure–outcome pair with a given set of exposure instruments, we examined the distribution of the p -values of the associations of the variants with the outcome. We computed the proportion of non-null p -values as a measure of association significance consistency. This measure is commonly used by FDR methods and it is estimated by comparing the observed distribution of p -values to a random uniform distribution. Specifically, we use the local FDR method implemented in limma (version 3.42.2)^{64,65}.

BN inference. We used the bnlearn R package (version 4.5) for inferring BN⁶⁶. We implemented the analysis presented in Howey et al. ²⁴: BN were inferred using the hill climbing algorithm using the Bayesian Information Criterion (BIC) as the objective function with 50 random restarts to address convergence into local optimum. Directed edges from the phenotypes into the genetic variants were not allowed (i.e., using the bl option of the algorithm). We used the BIC-CG discrete–continuous hybrid objective function that models continuous variables using a mixture of Gaussians. Discretization of the data instead of this option resulted in worse empirical FDR scores and was therefore excluded. Networks were averaged over 100 bootstrap repeats, where in each repeat the top inferred network was kept. For each phenotype pair, we computed the proportion of times they were linked in the networks and took either the top 10 or top 20 as the predictions of the algorithm.

Other analyses. We tested MR methods by taking the GWAS summary statistics of each phenotype (without the MHC region) and ignoring the CI tests. For MR, given a p -value threshold ($p = 1 \times 10^{-6}$, 1×10^{-7} , or 1×10^{-8}) we select the top GWAS regions with MAF > 1%, clump them using PLINK, and use the filtered results as the instruments. We used the MendelianRandomization R package (version 0.4.2)⁶⁷ to run MR-Egger and IVW. We used the MR-PRESSO implementation from the original publication⁸ (version 1.0) with outlier correction, and also report the results of its global test for pleiotropy.

Robustness analysis. To measure the robustness of our pipeline we tested different values for p_1 (1×10^{-6} , 1×10^{-7} , 1×10^{-8}) and p_2 (0.1, 0.01, 0.001) and computed the Jaccard score for each network type between the different combinations of p_1 and p_2 (see Supplementary Fig. 3).

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

Data availability

Code to simulate data and the UK-Biobank summary statistics used to generate the results are available at <https://github.com/david-dd-amar/cGAUGE/>⁶⁸. UK-Biobank data was retrieved using application 24983. The results from all analyses, including the instrument sets and pairwise Mendelian Randomization results are available in the Supplementary Data.

Code availability

R implementation of cGAUGE is available at <https://github.com/david-dd-amar/cGAUGE/>. Code for preprocessing UK-Biobank biomarker data is available at https://github.com/rivas-lab/public-resources/tree/master/uk_biobank/laboratory-tests.

Received: 16 July 2019; Accepted: 2 December 2020;

Published online: 13 January 2021

References

1. Pearl, J. *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2009).

2. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* 2nd edn, Vol. 1 (The MIT Press, 2001).
3. Weinberger, N. Faithfulness, coordination and causal coincidences. *Erkenntnis* **83**, 113–133 (2018).
4. Pearl, J. & Dechter, R. Identifying independencies in causal graphs with feedback. In *UAI'96 Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence* 420–426 (Morgan Kaufmann Publishers Inc., 1996).
5. Neal, R. M. On deducing conditional independence from d-separation in causal graphs with feedback. *J. Artif. Intell. Res.* **12**, 87 (2000).
6. Meek, C. Strong Completeness and Faithfulness in Bayesian Networks. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence 411–418 (1995).
7. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* <https://doi.org/10.1038/s41588-018-0255-0> (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. *Nat. Genet.* **50**, 693–698 (2018).
9. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* **26**, 2333–2355 (2017).
10. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
11. Pingault, J. B. et al. Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).
12. Jiang, L. et al. Constrained instruments and their application to Mendelian randomization with pleiotropy. *Genet. Epidemiol.* <https://doi.org/10.1002/gepi.22184> (2019).
13. Lawlor, D. A. Commentary: two-sample Mendelian randomization: opportunities and challenges. *Int. J. Epidemiol.* **45**, 908–915 (2016).
14. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
15. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
16. Spence, J. D. Statins and cataracts: reverse causality? *Can. J. Cardiol.* **31**, 691.e11 (2015).
17. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
18. Burgess, S., Daniel, R. M., Butterworth, A. S. & Thompson, S. G., the EPIC-InterAct Consortium. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int. J. Epidemiol.* **44**, 484–495 (2015).
19. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
20. Morrison, J. et al. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **52**, 740–747 (2020).
21. Badsha, M. B. & Fu, A. Q. Learning causal biological networks with the principle of mendelian randomization. *Front. Genet.* **10**, 460 (2019).
22. Zhang, L. & Kim, S. Learning gene networks under SNP perturbations using eQTL datasets. *PLoS Comput. Biol.* **10**, e1003420 (2014).
23. Wang, L., Audenaert, P. & Michoel, T. High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. *Front. Genet.* **10**, 1196 (2019).
24. Howey, R., Shin, S. Y., Relton, C., Smith, G. D. & Cordell, H. J. Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. *PLoS Genetics* **16**, e1008198 (2020).
25. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.* **1**, 123–126 (2012).
26. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
27. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, 2012).
28. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
29. Amar, D., Shamir, R. & Yekutieli, D. Extracting replicable associations across multiple studies: empirical Bayes algorithms for controlling the false discovery rate. *PLoS Comput. Biol.* **13**, e1005700 (2017).
30. McLachlan, G. J., Bean, R. W. & Jones, L. B.-T. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615 (2006).
31. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
32. Eppinga, R. N. et al. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. *Nat. Genet.* **48**, 1557 (2016).
33. International Schizophrenia Consortium et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
34. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. & Bühlmann, P. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw., Artic.* **47**, 1–26 (2012).
35. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
36. Morris, J. H. et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinforma.* **12**, 436 (2011).
37. Ragab, G., Elshahaly, M. & Bardin, T. Gout: an old disease in new perspective — a review. *J. Adv. Res.* **8**, 495 (2017).
38. Nabel, E. G. & Braunwald, E. A tale of coronary artery disease and myocardial infarction. *N. Engl. J. Med.* **366**, 54–63 (2012).
39. Martins, D., Tareen, N., Pan, D. & Norris, K. The relationship between body mass index, blood pressure and pulse rate among normotensive and hypertensive participants in the third National Health and Nutrition Examination Survey (NHANES). *Cell. Mol. Biol.* **49**, 1305–1309 (2003).
40. Gray, N., Picone, G., Sloan, F. & Yashkin, A. Relation between BMI and diabetes mellitus and its complications among US older adults. *South. Med. J.* **108**, 29–36 (2015).
41. Alawad, A. O., Merghani, T. H. & Ballal, M. A. Resting metabolic rate in obese diabetic and obese non-diabetic subjects and its relation to glycaemic control. *BMC Res. Notes* **6**, 382 (2013).
42. Thomas, M. C. et al. Blood pressure control and risk of incident atrial fibrillation. *Am. J. Hypertens.* **21**, 1111–1116 (2008).
43. Watanabe, H. et al. Metabolic syndrome and risk of development of atrial fibrillation: the Niigata preventive medicine study. *Circulation* **117**, 1255–1260 (2008).
44. Green, R. H. et al. Asthma exacerbations and sputum eosinophil counts: a randomised controlled trial. *Lancet* **360**, 1715–1721 (2002).
45. Price, D. B. et al. Blood eosinophil count and prospective annual asthma disease burden: a UK cohort study. *Lancet Respir. Med.* **3**, 849–858 (2015).
46. Jafarzadeh, A., Poorgholami, M., Izadi, N., Nemati, M. & Rezayati, M. Immunological and hematological changes in patients with hyperthyroidism or hypothyroidism. *Clin. Investig. Med.* **33**, E271–E279 (2010).
47. Hackett, T. L. et al. Oxidative modification of albumin in the parenchymal lung tissue of current smokers with chronic obstructive pulmonary disease. *Respir. Res.* **11**, 180 (2010).
48. Vaz Fragoso, C. A. et al. The ratio of FEV1 to FVC as a basis for establishing chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **181**, 446–451 (2010).
49. Taki, Y. et al. Correlation among body height, intelligence, and brain gray matter volume in healthy children. *Neuroimage* **59**, 1023–1027 (2012).
50. Baker, E. R. Body weight and the initiation of puberty. *Clin Obstet Gynecol.* **28**, 573–579 (1985).
51. Vollmer, S. et al. Association between economic growth and early childhood undernutrition: evidence from 121 Demographic and Health Surveys from 36 low-income and middle-income countries. *Lancet Glob. Health* **2**, e225–e234 (2014).
52. Ong, T. J., Mehta, A., Ogston, S. & Mukhopadhyay, S. Prediction of lung function in the inadequately nourished. *Arch. Dis. Child.* **79**, 18–21 (1998).
53. Harris, J. A. & Benedict, F. G. A biometric study of human basal metabolism. *Proc. Natl Acad. Sci. USA* **4**, 370–373 (1918).
54. Danese, E., Lippi, G. & Montagnana, M. Red blood cell distribution width and cardiovascular diseases. *J. Thorac. Dis.* **7**, E402–E411 (2015).
55. Richardson, T.S. A discovery algorithm for directed cyclic graphs. In *UAI'96 Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence* 454–461 (1996).
56. Strobl, E. V. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *Int. J. Data Sci. Anal.* **8**, 33–56 (2019).
57. Voight, B. F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
58. Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* **40**, 597–608 (2016).
59. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).

60. Sinnott-Armstrong, N. et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* (2020). In press.
61. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
62. Smoot, M. E., Ono, K., Ruschinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
63. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
64. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
65. Langaas, M., Lindqvist, B. H. & Ferkingstad, E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B* **67**, 555–572 (2005).
66. Scutari, M. & Denis, J. -B. *Bayesian Networks: With Examples in R* (CRC Press, 2014).
67. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
68. Amar, D. *cGAUGE: Causal Graphical Analysis Using Genetics* <https://doi.org/10.5281/zenodo.4281327> (2020).

Acknowledgements

M.A.R. is supported by Stanford University and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). N.S.-A. is funded by the Department of Defense through a National Defense Science and Engineering Grant and by a Stanford Graduate Fellowship. The primary and processed data used to generate the analyses presented here are available in the UK Biobank access management system (<https://amsportal.ukbiobank.ac.uk/>) for application 24983, “Generating effective therapeutic hypotheses from genomic and hospital linkage data” (<http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf>), and the results are displayed in the Global Biobank Engine (<https://biobankengine.stanford.edu>). We thank Jonathan Pritchard and members of the Rivas and Ashley labs for helpful comments. This work was supported by National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under awards R01HG010140. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

D.A. developed the methods and implemented the algorithms. D.A., N.S.-A., and M.A.R., designed the experiments and prepared the data. D.A., N.S.-A., E.A., and M.A.R., analyzed the results and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-20516-2>.

Correspondence and requests for materials should be addressed to M.A.R.

Peer review information *Nature Communications* thanks Jean Morrison, Itsik Pe'er and Marie Verbanck for their contribution to the peer review of this work.

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

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



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

© The Author(s) 2021