Received: 22 March 2022 | Revised: 12 July 2022 | Accepted: 18 August 2022

DOI: 10.1002/mgg3.2055

Molecular Genetics & Genomic Medicine Open Access WILEY

REVIEW ARTICLE

A review of causal discovery methods for molecular network analysis

Jack Kelly[1] | **Carlo Berzuini**[1] | **Bernard Keavney**[2,3] | **Maciej Tomaszewski**[2,4] | **Hui Guo**[1]

[1]Centre for Biostatistics, School of Health Sciences, Faculty of Medicine, Biology and Health, University of Manchester, Manchester, UK

[2]Division of Cardiovascular Sciences, Faculty of Medicine, Biology and Health, University of Manchester, Manchester, UK

[3]Division of Cardiology and Manchester Academic Health Science Centre, Manchester University NHS Foundation Trust, Manchester, UK

[4]Manchester Heart Centre and Manchester Academic Health Science Centre, Manchester University NHS Foundation Trust, Manchester, UK

Correspondence
Jack Kelly, Centre for Biostatistics, School of Health Sciences, Faculty of Medicine, Biology and Health, University of Manchester, Manchester, UK.
Email: jack.kelly@manchester.ac.uk

Funding information
British Heart Foundation and The Alan Turing Institute, Grant/Award Number: SP/19/10/34813

Abstract

Background: With the increasing availability and size of multi-omics datasets, investigating the casual relationships between molecular phenotypes has become an important aspect of exploring underlying biology andgenetics. There are an increasing number of methodlogies that have been developed and applied to moleular networks to investigate these causal interactions.

Methods: We have introduced and reviewed the available methods for building large-scale causal molecular networks that have been developed and applied in the past decade.

Results: In this review we have identified and summarized the existing methods for infering causality in large-scale causal molecular networks, and discussed important factors that will need to be considered in future research in this area.

Conclusion: Existing methods to infering causal molecular networks have their own strengths and limitations so there is no one best approach, and it is instead down to the discretion of the researcher. This review also to discusses some of the current limitations to biological interpretation of these networks, and important factors to consider for future studies on molecular networks.

KEYWORDS

Bayesian networks, causal inference, causal molecular network, mendelian randomisation, omics

1 | INTRODUCTION

Molecular networks are important to understanding biological process beyond the analysis of a single gene or molecule (Han, 2008). The operation of molecular phenotypes at all levels is not isolated and interactions make up complicated networks that contain a wealth of information. In an age where data is being produced more than ever, these networks can become increasingly complex. A molecular network contains a set of nodes and edges. Nodes represent information from multi-omics, including but not limited to genes, messenger RNAs (mRNAs), proteins, DNA methylation patterns and protein phosphorylation. Edges represent the relationship between the nodes and so can symbolise direct and indirect relationships between molecular phenotypes and transcriptional regulation.

One of the primary advantages of molecular networks is in elucidating genetic and biological mechanisms underlying disease. Even in diseases with known causative genes (eg. *CFTR* mutation causing Cystic fibrosis (Elborn, 2016) and mutations in *HTT* leading to Huntington's disease (Ha & Fung, 2012)) these genes act as part of a large network and never in isolation. Dysregulated biological processes and important 'hubs' within them can be identified as disease drivers, which potentially help identify drug targets that impact sets of associated genes rather than important individual genes, though this has yet to be translated to clinically useful therapies (Chagoyen et al., 2019).

Undirected networks have been an important approach for the investigation of biological processes and identification of hub genes in disease. Traditionally, protein–protein interaction networks have been built using a combination of in vivo and in vitro methods to understand interactions, however these approaches have huge time and financial costs, and result in noisy networks with high false positive rates (Rao et al., 2014). Approaches to omics data using in silico methods have been used as an alternative to better understand these undirected associations (Kotlyar et al., 2015). Most commonly, co-expression molecular networks are built on the basis of correlation structures (Villa-Vialaneix et al., 2013). It has become popular to use specific R software to infer undirected networks from transcriptomics data. For example, weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008) is particularly user-friendly as the authors have produced extensive tutorials and guides to increase accessibility to researchers. Although providing limited mechanistic understanding, undirected networks are important as they are often precursors of the study of causal networks.

Many undirected networks (as shown in Figure 1a) rely on using correlation between nodes to infer symmetric associations. However, causal networks aim to differentiate the directed regulatory relationships from just associations. This approach identifies directed (as shown in Figure 1b) or mixed networks (as shown in Figure 1c). It is worth noting that directed relationships in a network do not necessarily have a causal interpretation, as they may merely depict temporal orders in the data generating process. Only if the confounders between the nodes have been adjusted for will these relationships have a causal meaning.

Identifying causal relations from gene expression data was proposed over 20 years ago (Friedman et al., 2000). Since then, a large number of causal inference methods have been developed using omics data. This approach is advantageous in the study of biology as it allows for inferring causality without interventions, especially when randomised controlled trials are infeasible due to high cost and ethical issues (White & Vignes, 2019).

As the technology becomes more accessible and affordable, there is an increasing range of omics data that is being collected, which allows for integrative analysis to develop a more complete picture of how different types of omics interact with one another (Eales et al., 2021). Causal inference in molecular networks is a growing area of research. However, complex high dimensional causal networks have limited use and their contribution to the literature is heavily restricted as they are often difficult to interpret. There needs to be approaches that allow for identification of biologically important sub-networks and a small number of targets for future research or therapeutic intervention.

In this review, we will discuss the current literature using causal discovery methods on molecular networks and challenges that the area is facing. We will also discuss factors that influence interpretation of causal networks, including clustering and visualisation. Previous reviews (Glymour et al., 2019; Yazdani, 2020) have focussed on introducing methodologies of building causal networks and given few biological examples, however here we will focus on published methods and their applications specifically to molecular networks and subsequent biological interpretation.

2 | CAUSAL MOLECULAR NETWORKS

Applications of different causal methods to omics data is covered in this review. The simplest causal network only involves the causal relationship between a pair of variables,

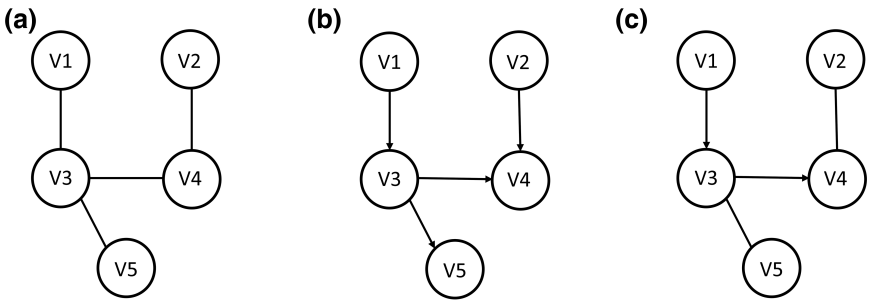

FIGURE 1 (a) an example of an undirected network, (b) a directed network and (c) a mixed network. Mixed networks have both directed and undirected edges.

investigating whether a single exposure can cause a single outcome. Causal networks can be made increasingly complex to investigate the relationships between thousands of variables. With applications to molecular phenotype data, the main approaches used to build causal networks have been Mendelian randomisation (MR) and Bayesian networks (BN), including the PC algorithm, as shown in Figure 2. Here, we consider MR, approaches to BNs and we then focus on a combination of approaches to reduce the limitations of any single method. A summary of the methodologies discussed here are shown in Table 1.

2.1 | Mendelian randomisation

MR uses single-nucleotide polymorphisms (SNPs) as 'instrumental variables' (IVs) to infer the causal effect of an exposure on an outcome. It mimics randomised controlled trials by assuming that SNP genotypes are randomly assigned to individuals within a population. MR requires three key assumptions (Figure 2a); (a) IVs are associated with the exposure of interest; (b) IVs are independent of confounders (both observed and unobserved) between exposure and outcome; (c) IVs only affects the outcome through the exposure of interest.

Horizontal pleiotropy occurs when the IV influences outcome outside of its effect on the exposure, breaking the assumption that genotype only affects the outcome through the exposure of interest. Several adaptations of MR have been developed to reduce the impact of horizontal pleiotropy. Popular approaches include MR-Egger (Bowden et al., 2015) (which models pleiotropy assuming that effects of the IV on exposure and outcome are independent), MR-PRESSO (Verbanck et al., 2018) (which corrects for pleiotropic outlier effects) and Causal Analysis Using Summary Effect estimates (CAUSE) (Morrison et al., 2020) (which accounts for correlated and uncorrelated pleiotropic effects). MR-PRESSO and MR-Egger are often both applied to data and results compared to reduce the impact of pleiotropy and outliers. These approaches have been used to provide evidence to support the casual effect of estimated glomerular filtration rate, a measure of kidney function, on chronic kidney disease, kidney stone formation, diastolic blood pressure and hypertension (Morris et al., 2019). Additionally, they have been used to show the causal effect of blood pressure on renal outcomes commonly affecting patients with hypertension (Eales et al., 2021).

In most cases discussed above, MR analysis requires the association between IV-exposure and IV-outcome are from two independent studies (Lawlor, 2016). This is known as two-sample MR. There are a limited number of one-sample MR methods that deal with IVs, exposures and outcomes coming from a single study (Bowden et al., 2015; Zhao et al., 2018). Some expansions to MR have been developed to handle data when two studies have overlapping individuals in common (LeBlanc et al., 2018), which in classic MR approaches lead to bias. Zou et al. (2020) have

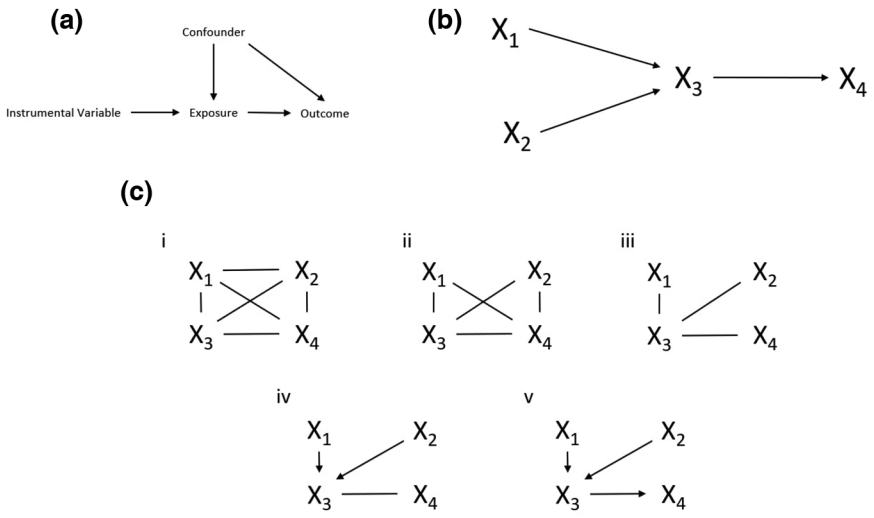

FIGURE 2 (a) Schematic representation of MR. MR infers the causal effect of an exposure (phenotype) on the outcome using instrumental variables (IVs). (b) Causal Bayesian networks connect nodes via directed edges determined by conditional independence, which is present when the relationship between two nodes is independent conditioning on all other nodes in the graph. (c) Schematic representation of the PC algorithm. The true causal graph is shown in (b). The PC algorithm initially begins with an undirected fully connected graph (i) and uses data to create a skeleton graph with undirected edges. In this case, the $X_1 - X_2$ edge is removed because X_1 is independent of X_2 (ii) and the edges between $X_1 - X_4$ are removed as the nodes are independent given X_3 . The same is true for the $X_2 - X_4$ edge (iii). Then v-structures are identified (iv) and final edges oriented (v) (Le et al., 2019).

TABLE 1 Summary of the discovery methods for analysis of causal molecular networks including the software available

Methodologies	Data source required	Advantages	Disadvantages	Software available
Mendelian Randomisation				
Mendelian randomisation (MR)	GWAS, omics	• Only requires summary statistics, fast to run • Estimates causal effect size	• Data must meet certain (possibly untestable) assumptions • Incapable of modelling complex relationships	• MendelianRandomization (R package) (Yavorska & Burgess, 2017) • TwoSampleMR (R package) (Hemani et al., 2018) • MR-Base (Hemani et al., 2018)
Bayesian MR	GWAS, omics	• Flexibility of modelling complex data structure (overlapping samples, horizontal pleiotropy, interactions, multiple exposures) • Estimates causal effect size	• Data must meet certain (possibly untestable) assumptions • Computationally intensive, applicable to small-to-medium causal networks	
Bayesian networks				
Bayesian networks (BN)	omics	• Can generate larger causal networks • Causal edges probabilities are given • Estimates causal effect size	• Computationally intensive limiting network size	• Bnlearn (R and Python package) (Scutari, 2010) • BayesNetty (Howey et al., 2020, 2021)
PC algorithm (Spirtes et al., 2000)	omics	• Relatively fast compared to other BNs	• Although faster than alternatives, computationally challenging when run on very large datasets • Causal effect size is not inferred	• Bnlearn (R and Python package) (Scutari, 2010) • Pcalg (R package) (Hauser & Bühlmann, 2012; Kalisch et al., 2012)
Combination methods				
Bayesian and MR (findr (Wang et al., 2019)/MRPC (Badsha & Fu, 2019))	GWAS, omics	• Undirected network construction followed edge directions inferred using MR	• Still computationally intensive and applications have been on subsets of omics data • Causal effect size is not inferred	• findr (R package) (Wang et al., 2019) • MRPC (R package) (Badsha & Fu, 2019)
Genome Granularity DAG (GDAG) (Yazdani et al., 2016a)	GWAS, omics	• Undirected network construction followed edge directions inferred using MR	• Still computationally intensive and applications have been on subsets of omics data • Causal effect size is not inferred	
Causal Graphical Analysis Using GEnetics (cGAUGE) (Amar et al., 2021)	GWAS, omics	• Approach has greater power and lower false discovery rate than BNs	• Computationally intensive • Causal effect size is not inferred • Greater power than BN, reduced power in presence of horizontal pleiotropy	• cGUAGE (R code: https://github.com/david-dd-amar/cGAUGE) (Amar et al., 2021)
Time series causal networks				
Granger causality (Granger, 1969)	Time series omics	• Allows causal inference using time series omics data	• Time intervals between measurements needs to be enough for a noticeable change to take place • Needs to be no confounders	• lmtest (R package) (Zeileis & Hothorn, 2002) • statsmodels.tsa.stattools.grangercausalitytests (Python package) (Seabold & Perktold, 2010)
Optimal Causation Entropy (OCE) (Sun et al., 2015)/ PCMCI (Runge et al., 2019)	Time series omics	• Outperform Granger causality using time series omics data • Can generate large scale causal networks	• Assumes stationarity which can be violated by confounders	• TIGRAMITE (Python package) (Runge et al., 2019)

developed a more flexible Bayesian MR method that can handle one, two and overlapping samples. Bayesian MR has an advantage in its flexibility of coping with complex data structures, such as overlapping samples, horizontal pleiotropy, study heterogeneity and multiple exposure and outcomes, all in a single model (Berzuini et al., 2020; Zou et al., 2020, 2021).

Advanced MR methods have been developed more recently, such as MR-ConMix (contamination mixture method for robust and efficient estimation) (Burgess et al., 2020) and GRAPPLE (Genome-wide mR Analysis under Pervasive PLEiotropy) (Wang et al., 2021), that utilises both strongly and weakly associated SNPs to identify multiple pleiotropic pathways. Both have discussed the future importance of including multiple exposures in the study of genetics and MR. The Causal Inference Test (cit) (Millstein et al., 2016) is a more conservative method that applies the principles of MR and is more robust to pleiotropic effects and reverse causation.

These advancements in MR methodologies provide researchers with more options to design models that better fit the assumptions of MR. Inferring causality using MR has been increasingly applied (Bowden & Holmes, 2019; Nordestgaard & Nordestgaard, 2016), however have been focused on smaller-to-medium scale and applications to large scale omics networks have been limited. A thorough review of MR has recently been published by Sanderson et al. (2022). Nevertheless, MR has found applications being used in combination with other approaches to building molecular networks, which will be discussed shortly.

2.2 | Bayesian networks

Bayesian networks (BNs) were one of the first approaches proposed to investigate gene expression networks (Friedman et al., 2000). BNs use Bayesian inference to calculate probabilistic graphical models of data. BNs are directed acyclic graphs (DAGs) with directed edges and no subset of nodes that can form a closed loop. The edges of the DAG are determined via conditional independence which is present when two nodes are independent conditioning on all other nodes in the graph. An example of a BN is shown in Figure 2b. The two traditional classes of Bayesian networks are constraint-based and score-based. Constraint-based methods learn an undirected network skeleton using conditional independence testing and then assign the direction of edges between nodes that are not found to be independent. Score-based methods instead aim to optimise a scoring criterion across a search space of DAGs. Additionally, there are hybrid algorithms aggregate constrained and score-based algorithms which although have been widely applied in building causal network (Li & Guo, 2018), they have had limited applications in the molecular network literature.

Due to the high computational cost, most studies have been limited to inferring causal relationships within triplets of a gene regulatory network (Bucur et al., 2019) with limited approaches to scaling networks to larger more complete molecular networks. Much of the literature using BN to infer molecular networks has introduced limitations to the size of the networks built. Mäkinen et al. (2014) used BNs to investigate coronary artery disease, introducing genetic information as priors by not allowing genes that have no associated SNPs to be parents of genes that have an associated SNP. However, this was only done on a subset of genes rather than a full network.

Azad and Alyami (2021) used BNs to investigate causal gene expression networks in Lapatinib resistance to better understand why some breast cancer patients have unsuccessful treatment. They used different Markov Chain Monte Carlo (MCMC) sampling algorithms to identify the optimum molecular network from the BN search space. MCMC samples a probability distribution where the next sample is dependent on the current sample. The study was limited to genes within the TGF-β signalling pathway in lapatinib sensitive and resistant breast cancer cells, identifying the driver genes as being associated with the GO biological terms positive regulation of pathway-restricted SMAD protein phosphorylation and regulation of lymphocyte.

Other approaches to learning BNs using MCMC schemes have been proposed. Castelletti and Consonni (2019) used MCMC to learn the Markov equivalence class of DAGs to investigate protein signalling in observational and interventional samples. This approach requires little tuning as it uses default parameter priors and so is more accessible to researchers than other Bayesian approaches. The authors have also used a Bayesian active learning procedure to identify DAGs (Castelletti & Consonni, 2020) in the same protein signalling dataset and show that DAGs can be identified even when only using a subset of the intervention samples.

Similarly, Bhattacharya and Das (2019) applied BNs to investigate causal genes in drug pathways for cancer, using a limited set of known drug target genes and genes identified by machine learning. Using a small dataset, they identified gene to gene connections that play a role in imatinib resistance in chronic myeloid leukaemia, including a *ACADVL* to *PDIA5* connection present uniquely in non-responder populations. These two proteins have been previously shown to play important roles in cancer drug-resistance (Higa et al., 2014). Additionally, BNs have been used in the past to identify any causal effects of microRNA (miRNA) on gene expression interactions (Lee &

Jiang, 2017). However, these networks are very limited, with causal edges only from miRNA to gene expression and in many cases failed to identify known gene–gene interactions from experiment-supported databases.

Identifying the optimal BN is very difficult, and many approaches have been proposed with the aim to improve this process within transcriptional networks (Azad & Alyami, 2021). For example, Howey et al. have developed BayesNetty (Howey et al., 2020, 2021), an accessible software for building Bayesian networks using genetic and phenotypic data. This software allows users to apply algorithms accessible in the R package bnlearn (Scutari, 2010) to biological relationships. Howey et al. (2020) used BayesNetty to implement the score-based BN approach called hill climbing to investigate a small number of interactions between metabolites and phenotypic data. They use genetic anchors to ensure there can be no directional edges towards genetic variants and found it outperformed MR in highly pleiotropic scenarios. This software includes approaches that can effectively impute missing data using a version of nearest neighbour imputation and the ability to add weights to certain edges, allowing researchers to incorporate prior knowledge concerning directions between nodes. These improvements have only shown to be moderate and remain computationally intensive for generating large networks. Large amounts of information could be missed if only a subset of data is used to build causal networks which is generally the approach used with BN due to the high computational cost. It is possible to sacrifice accuracy of networks for speed using approximate solutions (Guo & Constantinou, 2020), however this is not guaranteed to make it possible to build networks using data that is as highly dimensional as omics data.

The PC algorithm (Spirtes et al., 2000) (named after its initial authors, Peter Spirtes and Clark Glymour) is a constraint-based approach to estimating Bayesian networks, starting with a fully connected undirected graph and recursively deleting edges based on conditional independence properties. This generates a completed partially DAG (CPDAG) which consists of both directed and undirected edges. The steps the PC algorithm takes to build causal networks are shown in Figure 2c. The PC algorithm is fast for high dimensional and sparse problems, which makes it more suited towards uses with molecular network data (Maathuis et al., 2010).

Zhang et al. (2012) used the PC algorithm with gene expression data to identify conditional independence between pairs of genes to build gene regulatory networks. Le et al. (2013) predicted the causal mRNA targets of miRNAs using a method named Intervention-calculus when the DAG is Absent (IDA) (Maathuis et al., 2010). IDA has been shown to have use in investigating the impact of regulators on gene expression (Ye et al., 2021) but has seen little practical use to investigate disease. Zhang et al. (2014) applied the method to epithelial-mesenchymal transition and multi-class cancer datasets and results were validated by transfections experiments.

Zhang et al. (2014) used the IDA approach to infer miRNA-mRNA pair interactions, and identified differences in causal effects between different conditions. They have used IDA to infer causality of long non-coding RNA (lncRNA) on mRNA within modules identified using WGNCA to identify lncRNAs in specific biological functions (Zhang et al., 2018), an approach that has also since been used to investigate pan-cancer (Ye et al., 2021).

Despite being faster than alternatives, the PC algorithm is still slow when applied to high dimensional datasets, and so as data is integrated runtime will increase (Le et al., 2019). The PC algorithm has seen limited use on its own in applications to molecular networks. However, it has been used more recently in combination with other approaches to infer causality in biological data.

2.3 | Combination of approaches

Research is trending towards the use of a combination of approaches to building causal molecular networks, with the aim to reduce the limitations of individual approaches and build more robust networks. MR, in particular, has been combined with other methods to help topologies and speed up construction of causal network by putting constraints on edge directions.

Yazdani et al. (2016a) proposed an approach to identifying causal networks named genome granularity DAG (GDAG). Initially, strong IVs are generated from phenotype SNP data across each chromosome independently. The structure of the undirected network for omics data is identified, and the principle of MR is used to determine the directionality of edges using the strong IVs generated previously. They have used this approach to investigate the network of metabolites (Yazdani et al., 2016b, 2019).

Augmenting Bayesian networks with the principles of MR has become popular for building molecular networks (Yazdani, 2020). Wang et al. (2019) have tried to address the computational limitations of BNs on large-scale transcriptome-wide networks using a tool they have named findr. They used the SNPs that are directly associated with gene expression, known as expression quantitative trait loci (eQTLs). For each gene, the most strongly associated eQTL is selected as the IV in inferring the pairwise causal relationships between all genes in the network. These edges are ranked and assembled into a DAG (Wang & Michoel, 2018). This method is much more efficient and outperforms traditional ways of building BNs, though has rarely been practically applied in the literature.

Badsha and Fu (2019) have developed MRPC, which incorporates the principle of MR into the PC algorithm. The principle of MR is generalised to account for a variety of causal relationships between SNPs and molecular phenotypes. MRPC begins by learning the graph skeleton using the PC algorithm with an online false discovery rate correction and any edges are oriented to point from SNPs to molecular phenotypes. MRPC then looks for v-structures in the network between any 3 nodes and uses the principle of MR to help orient edges. Although MRPC has been shown to be very effective for building molecular networks, there is still room to develop further. Within small to medium networks MRPC performs exceptionally, however for very high dimensional data as is common with multi-omics data, it is still computationally expensive and could be further optimised.

A recent paper by Zuber et al. (2020) proposed a multivariable MR and Bayesian model averaging (MR-BMA) approach that can include information from many IVs using only summary statistics from genetic association studies. It assumes the proportion of true causal risk factors is sparse when compared with all risk factors, which they demonstrate is usually the case with metabolomics data. Using MR-BMA, they identify high density lipoprotein (HDL) cholesterol as a potential causal risk factor for age-related macular degeneration, supported by previous literature (Burgess & Davey Smith, 2017). This approach has also been used to identify Apolipoprotein B as key lipid risk factor for coronary artery disease (Zuber et al., 2021). All the above methods using the principle of MR require that the three assumptions of MR are satisfied. As multi-omics data is large and complex, using MR to sidestep the problems of confounding and reverse causation is important for causal network inference.

Causal Graphical Analysis Using GEnetics (cGAUGE) has also been proposed to construct causal networks by Amar et al. (2021). cGUAGE first identifies conditional independencies in the data that are used to identify IVs for downstream MR, and for the construction of large-scale networks, which is called ExSep. Initially, the skeleton is found using the PC algorithm. Edges between nodes are then oriented. If SNPs are marginally associated with a node X_2, but are independent of X_2 given another node X_1, then this is used as evidence that X_1 causally affects X_2. cGUAGE does not infer causal effect size, so there is a lot of future potential in integrating ExSep with MR and other approaches to infer the skeleton and quantify causal effects.

2.4 | Time series data

Time series data provides the opportunity to investigate molecular networks across a biological process. Generating causal networks is made much more difficult with the problems that inherently come with this data type. Particularly, the time between measurements may be inconsistent or not reflect the rate of change that is being investigated, causal relations can greatly change over time and unmeasured confounding variables may be introduced. As multi-omics data becomes easier to generate, there has been an increased interest in using time-series data to investigate molecular networks (Barman & Kwon, 2018).

The most common approach to identifying causality in time series molecular data is Granger causality which assumes that variable X Granger-causes Y if values of X provide information that is significant about the future values of Y (Granger, 1969). Heerah et al. (2021) have proposed Granger-causal analysis of gene expression data that can handle irregularly-spaced bivariate signals. However, it has some limitations that become obvious when using multi-omics data. The time intervals between measurements needs to be enough for a noticeable change to take place and there needs to be no confounders. Both assumptions are rarely met with biological data. Stehr et al. (2019) have used Siamese neural networks for causal inference in time series data, which gives the approximate probabilities between nodes. However, this approach has only been performed on balanced synthetic data and has yet to be shown to be effective in real unbalanced data.

Multiple Bayesian approaches to inferring causality in time series gene expression data have been developed. fastBMA implements Bayesian model averaging (BMA) to efficiently identify gene regulation networks (Hung et al., 2017). Other Bayesian approaches such as Bayesian Gene Regulation Model Inference (BGRMI) (Iglesias-Martinez et al., 2016) can integrate known protein interactions and ChIP-sequencing data as prior knowledge to assist in reconstructing regulatory network of time series gene expression data.

Causal analysis of time series molecular data is still very limited. Although new methodologies are being developed in other research domains (Runge, 2018), there has been limited applications to molecular networks. Modern algorithms such as Optimal Causation Entropy (OCE) (Sun et al., 2015) and the PC algorithm with a conditional mutual information (MCI) test to reduce autocorrelation and control false positive rates (PCMCI) (Runge et al., 2019) have been shown to outperform Granger causality and be able to handle large scale networks. Applying these approaches to molecular networks would be an important step in progressing the analysis of time series causal molecular networks.

3 | BIOLOGICAL INTERPRETATION OF NETWORKS

Networks of connected genes can quickly become very complex, which severely limits biological interpretation, even in simple co-expression network (Serin et al., 2016). Nevertheless, even when interpreting simple networks it is important to distinguish between association and causality. Inappropriate use of causal language has been a particular problem in biological sciences in the past (Boutron & Ravaud, 2018).

Causal molecular networks are often high dimensional. Many studies (Azad & Alyami, 2021; Bhattacharya & Das, 2019) have identified smaller subsets of genes they are interested in through previous knowledge of pathways or clustering of undirected networks before inferring causality. However, this can miss out factors that may be relevant within the causal network but are not within the cluster or not identified by traditional univariate analysis. Alternatively, constructing a causal network and then clustering the nodes would identify any functionally close sets of variables that are likely involved in similar biological processes. Few published papers have carried out clustering within causal molecular network. As the size of these networks grow, clustering will become increasingly important to identify biological processes and important causal molecules within them.

An advantage of large causal molecular networks is drug discovery and repurposing. Previous approaches to identifying drugs have been focussed on correlating transcription signatures between disease and known drugs (Belyaeva et al., 2021) however this approach generates drugs and therapeutic targets that rarely are further researched, and have not had much success in bringing any new treatments to the clinic. Causal pathways allow for more in-depth identification of drug targets. Škrlj et al. (2021) have developed Causal Network of Diseases (CaNDis) which uses causal protein–protein interactions to identify FDA-approved drugs that can impact particular diseases. A known drug pathway signature from databases such as CMap (Lamb et al., 2006) can be matched to the causal network to impact a particular target. Causal networks can also be studied to identify upstream regulators of known disease targets that can be targeted using drugs. Unfortunately, these advancements have had little use in the literature and thus limited translation to the clinic. Further development of methodologies and additional work using these drug discovery tools when constructing molecular causal networks should be included in future research as they become more accessible.

Network visualisation is often one of the first steps once networks have been created. One of the advantages of network visualisation is the ability to better communicate the results to readers and colleagues without a full understanding of how results were generated. Appropriate visualisation therefore becomes crucial to reflect the results and get the most from the data. There are many tools that assist in generating networks, including Cytoscape (Shannon et al., 2003) and Gephi (Bastian et al., 2009). These tools generally include a large amount of customisability to visualise the network, particularly in automatically generating layouts.

However, visualising and interpreting very large and complex networks can be difficult and often overlooked in the literature. Selecting the best and most appropriate way of displaying networks is very dependent on the type of network that is being visualised, and so requires a large amount of input by someone who understands the data and how it has been analysed. In molecular networks with multi-omics data, layering the different omics types within the visualisation to show how they interact would give a much more structured view than any predesigned layout that is available. Some approaches, including Bayesian networks and MR, provide causal effect sizes which can be visualised within networks by increasing size of edges for larger effect sizes. This allows experts from other biological fields to interpret the interactions of molecular phenotypes and is more likely to lead to future research. There is potential for creating interactive networks where nodes and edges can be included or excluded by adjusting a causal effect size threshold. One of the aims of causal inference is the identification of a small number of targets for therapeutic interventions and so effective visualisation with easy interpretation can be used by other researchers to identify networks of their particular interest.

4 | CONCLUSION

Building causal molecular networks is becoming increasingly important in biology. Inferring causality from entirely observational data is much less time consuming and less expensive than traditional randomised trials or intervention experiments. Additionally, the availability of genetic and multi-omics data is massively increasing making casual molecular network inference a very powerful approach.

Here, we have reviewed the available approaches to building causal molecular networks. Traditional small-scale MR approaches infer causality between exposures and outcomes. This makes MR a powerful tool when combined with other approaches to build

large-scale networks but very limited when used on its own. Bayesian network methods, including the PC algorithm, are based on conditional independence properties and rarely scale to large multi-omics networks well. Additionally, many of the methods developed based on Bayesian networks output a Markov equivalence class that may lead to ambiguity between directed and undirected relationships.

Combinations of approaches to inferring causal networks have attracted increasing attention as they bring together the advantages of individual approaches, e.g. augmenting Bayesian networks with the principle of MR, such as MRPC (Badsha & Fu, 2019) and findr (Wang et al., 2019). This has allowed for scaling of networks to a much larger size, however computational cost is still very high. Still, these approaches have not been widely applied in the literature and there is still much to improve. Reducing the impact of unmeasured confounders and horizontal pleiotropy is important in any complex causal inference and is why MR plays an important role in these approaches. These issues are being addressed with modern MR methods such as MR-egger (Bowden et al., 2015), CAUSE (Morrison et al., 2020) and Bayesian MR, and integrating these approaches into combinations of methods should be a focus in the future.

Selecting IVs is also a challenge for large-scale casual networks. Linkage disequilibrium and pleiotropic effects can violate IV assumptions. Selecting strong IVs would potentially reduce data size, thus reducing computation time, and reduce bias. However, there is a trade-off as only including strong IVs that only explain a small proportion of variation in the exposures may reduce the precision of the estimates. Therefore, the future challenge is to effectively identify and select for valid IVs that satisfy assumptions and are optimal for large causal molecular networks, which may prove to be especially difficult as it is not known if strong IVs will exist for every phenotype.

Many causal molecular network methods have focussed on use of individual level data, which can be difficult to get hold of as it is usually not included on public databases for ethical reasons. Improving the available methods that can infer causality using widely available summary statistics should be a priority for researchers so more can be done with current data. Improved methods, optimal interpretation and visualisation will advance understanding of disease processes. It is scientifically important but computationally challenging to take advantage of the increasing availability of multi-omics data that are now available, and directly translate to applications in clinical treatment of disease. Given the complex biological structure of certain outcomes, the literature points to a need to develop more flexible and comprehensive approaches to building causal molecular networks.

AUTHOR CONTRIBUTIONS

JK and HG undertook the literature search and co-wrote the first draft and approved the final version. CB, BK and MT critically appraised the manuscript and approved the final version.

ACKNOWLEDGMENTS

None.

FUNDING INFORMATION

This work was jointly supported by the British Heart Foundation and The Alan Turing Institute (which receives core funding under the EPSRC grant EP/N510129/1) as part of the Cardiovascular Data Science Awards (Round 2, SP/19/10/34813).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Jack Kelly https://orcid.org/0000-0003-2265-4649

REFERENCES

Amar, D., Sinnott-Armstrong, N., Ashley, E. A., & Rivas, M. A. (2021). Graphical analysis for phenome-wide causal discovery in genotyped population-scale biobanks. *Nature Communications*, *12*, 350. https://doi.org/10.1038/s41467-020-20516-2

Azad, A. K. M., & Alyami, S. A. (2021). Discovering novel cancer bio-markers in acquired lapatinib resistance using Bayesian methods. *Briefings in Bioinformatics*, *22*, bbab137. https://doi.org/10.1093/bib/bbab137

Badsha, B., & Fu, A. Q. (2019). Learning causal biological networks with the principle of mendelian randomization. *Frontiers in Genetics*, *10*, 460. https://doi.org/10.3389/fgene.2019.00460

Barman, S., & Kwon, Y.-K. (2018). A Boolean network inference from time-series gene expression data using a genetic algorithm. *Bioinformatics*, *34*(17), i927–i933. https://doi.org/10.1093/bioinformatics/bty584

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 361–362.

Belyaeva, A., Cammarata, L., Radhakrishnan, A., Squires, C., Yang, K. D., Shivashankar, G. V., & Uhler, C. (2021). Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing. *Nature Communications*, *12*(1), 1024. https://doi.org/10.1038/s41467-021-21056-z

Berzuini, C., Guo, H., Burgess, S., & Bernardinelli, L. (2020). A Bayesian approach to mendelian randomization with multiple pleiotropic variants. *Biostatistics*, *21*(1), 86–101. https://doi.org/10.1093/biostatistics/kxy027

Bhattacharya, S., & Das, A. (2019). Fast and robust method for drug response biomarker identification and sample stratification. *BioRxiv*, 10.1101/525337. https://doi.org/10.1101/525337

Boutron, I., & Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *PNAS*, *115*(11), 2613–2619. https://doi.org/10.1073/pnas.1710755115

Bowden, J., & Holmes, M. V. (2019). Meta-analysis and Mendelian randomization: A review. *Research Synthesis Methods*, *10*(4), 486–496. https://doi.org/10.1002/jrsm.1346

Bowden, J., Smith, G. D., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. https://doi.org/10.1093/ije/dyv080

Bucur, I. G., Claassen, T., & Heskes, T. (2019). Large-scale local causal inference of gene regulatory relationships. *International Journal of Approximate Reasoning*, *115*, 50–68. https://doi.org/10.1016/j.ijar.2019.08.012

Burgess, S., & Davey Smith, G. D. (2017). Mendelian randomization implicates high-density lipoprotein cholesterol–associated mechanisms in etiology of age-related macular degeneration. *Ophthalmology*, *124*, 1165–1174. https://doi.org/10.1016/j.ophtha.2017.03.042

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., & Howson, J. M. M. (2020). A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nature Communications*, *11*, 376. https://doi.org/10.1038/s41467-019-14156-4

Castelletti, F., & Consonni, G. (2019). Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *Annals of Applied Statistics*, *13*(4), 2289–2311.

Castelletti, F., & Consonni, G. (2020). Discovering causal structures in Bayesian Gaussian directed acyclic graph models. *Journal of the Royal Statistical Society Series A*, *183*(4), 1727–1745.

Chagoyen, M., Ranea, J. A. G., & Pazos, F. (2019). Applications of molecular networks in biomedicine. *Biology Methods and Protocols*, *4*(1), bpz012. https://doi.org/10.1093/biomethods/bpz012

Eales, J. M., Jiang, X., Xu, X., Saluja, S., Akbarov, A., Cano-Gamez, E., McNulty, M. T., Finan, C., Guo, H., Wystrychowski, W., Szulinska, M., Thomas, H. B., Pramanik, S., Chopade, S., Prestes, P. R., Wise, I., Evangelou, E., Salehi, M., Shakanti, Y., … Tomaszewski, M. (2021). Uncovering genetic mechanisms of hypertension through multi-omic analysis of the kidney. *Nature Genetics*, *53*(5), 630–637. https://doi.org/10.1038/s41588-021-00835-w

Elborn, J. S. (2016). Cystic fibrosis. *The Lancet*, *388*(10059), 2519–2531. https://doi.org/10.1016/S0140-6736(20)32542-3

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*(3–4), 601–620. https://doi.org/10.1089/106652700750050961

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*, 524. https://doi.org/10.3389/fgene.2019.00524

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, *37*, 424–438. https://doi.org/10.2307/1912791

Guo, Z., & Constantinou, A. C. (2020). Approximate learning of high dimensional Bayesian network structures via pruning of candidate parent sets. *Entropy*, *22*(10), 1142. https://doi.org/10.3390/e22101142

Ha, A. D., & Fung, V. S. C. (2012). Huntington's disease. *Current Opinion in Neurology*, *25*(4), 491–498. https://doi.org/10.1097/WCO.0b013e3283550c97

Han, J. D. J. (2008). Understanding biological functions through molecular networks. *Cell Research*, *18*(2), 224–237. https://doi.org/10.1038/cr.2008.16

Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, *13*, 2409–2464.

Heerah, S., Molinari, R., Guerrier, S., & Marshall-Colon, A. (2021). Granger-causal testing for irregularly sampled time series with application to nitrogen signalling in Arabidopsis. *Bioinformatics*, *37*(16), 2450–2460. https://doi.org/10.1093/bioinformatics/btab126

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Smith, G. D., Gaunt, T. R., & Haycock, P. C. (2018). The MR-base platform supports systematic causal inference across the human phenome. *eLife*, *7*, e34408.

Higa, A., Taouji, S., Lhomond, S., Jensen, D., Fernandez-Zapico, M. E., Simpson, J. C., Pasquet, J.-M., Schekman, R., & Chevet, E. (2014). Endoplasmic reticulum stress-activated transcription factor ATF6 requires the disulfide isomerase PDIA5 to modulate chemoresistance. *Molecular and Cellular Biology*, *34*(10), 1839–1849. https://doi.org/10.1128/MCB.01484-13

Howey, R., Clark, A. D., Naamane, N., Reynard, L. N., Pratt, A. G., & Cordell, H. J. (2021). A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships. *PLoS Genetics*, *17*(9), e1009811. https://doi.org/10.1371/journal.pgen.1009811

Howey, R., Shin, S. Y., Relton, C., Smith, G. D., & Cordell, H. J. (2020). Bayesian network analysis incorporating genetic anchors complements conventional mendelian randomization approaches for exploratory analysis of causal relationships in complex data. In. *PLoS Genetics*, *16*(3), e1008198. https://doi.org/10.1371/journal.pgen.1008198

Hung, L.-H., Shi, K., Wu, M., Young, W. C., Raftery, A. E., & Yeung, K. Y. (2017). fastBMA: Scalable network inference and transitive reduction. *GigaScience*, *6*, 1–10.

Iglesias-Martinez, L. F., Kolch, W., & Santra, T. (2016). BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research. *Scientific Reports*, *6*, 37140.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*(11), 1–26.

Kotlyar, M., Pastrello, C., Pivetta, F., Sardo, A. L., Cumbaa, C., Li, H., Naranian, T., Niu, Y., Ding, Z., Vafaee, F., Broackes-carter, F., Petschnigg, J., Mills, G. B., Jurisicova, A., Stagljar, I., Maestro, R., & Jurisica, I. (2015). In silico prediction of physical protein interactions and characterization of interactome orphans. *Nature Methods*, *12*, 79–84. https://doi.org/10.1038/nmeth.3178

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., & Golub, T. R. (2006). The connectivity map: Using gene-expression signatures to connect Small molecules, genes, and disease. *Small Molecules*, *313*, 1929–1935. https://doi.org/10.1038/nchembio1206-663

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. https://doi.org/10.1186/1471-2105-9-559

Lawlor, D. A. (2016). Commentary: Two-sample mendelian randomization: Opportunities and challenges. *International Journal of Epidemiology*, *45*, 908–915. https://doi.org/10.1093/ije/dyw127

Le, T. D., Hoang, T., Li, J., Liu, L., Liu, H., & Hu, S. (2019). A fast PC algorithm for high dimensional causal discovery with multi-Core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(5), 1483–1495. https://doi.org/10.1109/TCBB.2016.2591526

Le, T. D., Liu, L., Tsykin, A., Goodall, G. J., Liu, B., Sun, B. Y., & Li, J. (2013). Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics*, *29*(6), 765–771. https://doi.org/10.1093/bioinformatics/btt048

LeBlanc, M., Zuber, V., Thompson, W. K., Andreassen, O. A., Frigessi, A., Andreassen, B. K., Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., Albus, M., ... Treutlein, J. (2018). A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC Genomics*, *19*, 494. https://doi.org/10.1186/s12864-018-4859-7

Lee, S., & Jiang, X. (2017). Modeling miRNA-mRNA interactions that cause phenotypic abnormality in breast cancer patients. *PLoS One*, *12*(8), e0182666.

Li, H., & Guo, H. (2018). A hybrid structure learning algorithm for Bayesian network using Experts' knowledge. *Entropy*, *20*(8), 620.

Maathuis, M. H., Colombo, D., Kalisch, M., & Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, *7*(4), 247–248. https://doi.org/10.1038/nmeth0410-247

Mäkinen, V. P., Civelek, M., Meng, Q., Zhang, B., Zhu, J., Levian, C., Huan, T., Segrè, A. V., Ghosh, S., Vivar, J., Nikpay, M., Stewart, A. F. R., Nelson, C. P., Willenborg, C., Erdmann, J., Blakenberg, S., O'Donnell, C. J., März, W., Laaksonen, R., ... Assimes, T. L. (2014). Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genetics*, *10*(7), e1004502. https://doi.org/10.1371/journal.pgen.1004502

Millstein, J., Chen, G. K., & Breton, C. V. (2016). Cit: Hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics*, *32*(15), 2364–2365.

Morris, A. P., Le, T. H., Wu, H., Akbarov, A., van der Most, P. J., Hemani, G., Smith, G. D., Mahajan, A., Gaulton, K. J., Nadkarni, G. N., Valladares-Salgado, A., Wacher-R, N., & Franceschini, N. (2019). Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nature Communications*, *10*, 29. https://doi.org/10.1038/s41467-018-07867-7

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., & He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, *52*(7), 740–747. https://doi.org/10.1038/s41588-020-0655-9

Nordestgaard, A., & Nordestgaard, B. (2016). Coffee intake, cardiovascular disease and all cause mortality: Observational and mendelian randomization analyses in 95 000-223 000 individuals. *International Journal of Epidemiology*, *45*(6), 1938–1952. https://doi.org/10.1093/ije/dyw325

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. S. (2014). Protein-protein interaction detection: Methods and analysis. *International Journal of Proteomics*, *2014*,147648. https://doi.org/10.1155/2014/147648

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos*, *28*, 075310. https://doi.org/10.1063/1.5025050

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11),eaau4996. https://doi.org/10.1126/sciadv.aau4996

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., & Smith, G. D. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, *2*, 6.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*(3), 1–22.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M., & Ligterink, W. (2016). Learning from co-expression networks: Possibilities and challenges. *Frontiers in Plant Science*, *7*, 444. https://doi.org/10.3389/fpls.2016.00444

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303.metabolite

Škrlj, B., Eržen, N., Lavrač, N., Kunej, T., & Konc, J. (2021). CaNDis: A web server for investigation of causal relationships between diseases, drugs and drug targets. *Bioinformatics*, *37*(6), 885–887.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.

Stehr, M.-O., Avar, P., Korte, A. R., Parvin, L., Sahab, Z. J., Bunin, D. I., Knapp, M., Nishita, D., Poggio, A., Talcott, C. L., Davis, B. M., Morton, C. A., Sevinsky, C. J., Zavodszky, M. I., & Vertes, A. (2019). Learning causality: Synthesis of large-scale causal networks from high-dimensional time series data. *ArXiv*, 1905.02291. http://arxiv.org/abs/1905.02291

Sun, J., Taylor, D., & Bollt, E. M. (2015). Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, *14*(1), 73–106. https://doi.org/10.48550/arXiv.1401.7574

Verbanck, M., Chen, C. Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature Genetics*, *50*(5), 693–698. https://doi.org/10.1038/s41588-018-0099-7

Villa-Vialaneix, N., Liaubet, L., Laurent, T., Cherel, P., Gamot, A., & SanCristobal, M. (2013). The structure of a gene co-expression network reveals biological functions underlying

eQTLs. *PLoS One*, *8*(4), e60045. https://doi.org/10.1371/journal.pone.0060045

Wang, J., Zhao, Q., Bowden, J., Hemani, G., Smith, G. D., Small, D. S., & Zhang, N. R. (2021). Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS Genetics*, *17*(6), e1009575. https://doi.org/10.1371/journal.pgen.1009575

Wang, L., Audenaert, P., & Michoel, T. (2019). High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. *Frontiers in Genetics*, *10*, 1196. https://doi.org/10.3389/fgene.2019.01196

Wang, L., & Michoel, T. (2018). Controlling false discoveries in Bayesian gene networks with lasso regression p-values. *ArXiv*, 1701.07011. https://doi.org/10.48550/arXiv.1701.07011

White, A., & Vignes, M. (2019). Causal queries from observational data in biological systems via Bayesian networks: An empirical study in Small networks. *Methods in Molecular Biology*, *1883*, 111–142. https://doi.org/10.1007/978-1-4939-8882-2_5

Yavorska, O. O., & Burgess, S. (2017). MendelianRandomization: An R package for performing mendelian randomization analyses using summarized data. *International Journal of Epidemiology*, *46*(6), 1734–1739.

Yazdani, A. (2020). Mendelian randomization and causal networks for systematic analysis of omics. *ArXiv*, 2004.06958. https://doi.org/10.48550/arXiv.2004.06958

Yazdani, A., Yazdani, A., Elsea, S. H., Schaid, D. J., Kosorok, M. R., Dangol, G., & Samiei, A. (2019). Genome analysis and pleiotropy assessment using causal networks with loss of function mutation and metabolomics. *BMC Genomics*, *20*, 395. https://doi.org/10.1186/s12864-019-5772-4

Yazdani, A., Yazdani, A., Samiei, A., & Boerwinkle, E. (2016a). Generating a robust statistical causal structure over 13 cardiovascular disease risk factors using genomics data. *Journal of Biomedical Informatics*, *60*, 114–119. https://doi.org/10.1016/j.jbi.2016.01.012

Yazdani, A., Yazdani, A., Samiei, A., & Boerwinkle, E. (2016b). Identification, analysis, and interpretation of a human serum metabolomics causal network in an observational study. *Journal of Biomedical Informatics*, *63*, 337–343. https://doi.org/10.1016/j.jbi.2016.08.017

Ye, Z., Ke, H., Chen, S., Cruz-Cano, R., He, X., Zhang, J., Dorgan, J., Milton, D. K., & Ma, T. (2021). Biomarker categorization in transcriptomic meta-analysis by concordant patterns with application to pan-cancer studies. *Frontiers in Genetics*, *12*, 1079. https://doi.org/10.3389/fgene.2021.651546

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.

Zhang, J., Le, T. D., Liu, L., & Li, J. (2018). Inferring and analyzing module-specific lncRNA-mRNA causal regulatory networks in human cancer. *Briefings in Bioinformatics*, *20*(4), 1403–1419. https://doi.org/10.1093/bib/bby008

Zhang, J., Le, T. D., Liu, L., Liu, B., He, J., Goodall, G. J., & Li, J. (2014). Inferring condition-specific miRNA activity from matched miRNA and mRNA expression data. *Bioinformatics*, *30*(21), 3070–3077. https://doi.org/10.1093/bioinformatics/btu489

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J. K., Liu, Z. P., & Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, *28*(1), 98–104. https://doi.org/10.1093/bioinformatics/btr626

Zhao, Q., Wang, J., Hemani, G., Bowden, J., & Small, D. S. (2018). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *ArXiv*, 1801.09652. https://doi.org/10.48550/arXiv.1801.09652

Zou, L., Guo, H., & Berzuini, C. (2020). Overlapping-sample mendelian randomisation with multiple exposures: A Bayesian approach. *BMC Medical Research Methodology*, *20*, 295. https://doi.org/10.1186/s12874-020-01170-0

Zou, L., Guo, H., & Berzuini, C. (2021). Bayesian mendelian randomization with study heterogeneity and data partitioning for large studies. *ArXiv*, 2112.08147. https://doi.org/10.48550/arXiv.2112.08147

Zuber, V., Colijn, J. M., Klaver, C., & Burgess, S. (2020). Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nature Communications*, *11*, 29. https://doi.org/10.1038/s41467-019-13870-3

Zuber, V., Gill, D., Ala-Korpela, M., Langenberg, C., Butterworth, A., Bottolo, L., & Burgess, S. (2021). High-throughput multivariable mendelian randomization analysis prioritizes apolipoprotein B as key lipid risk factor for coronary artery disease. *International Journal of Epidemiology*, *50*(3), 893–901. https://doi.org/10.1093/ije/dyaa216

How to cite this article: Kelly, J., Berzuini, C., Keavney, B., Tomaszewski, M., & Guo, H. (2022). A review of causal discovery methods for molecular network analysis. *Molecular Genetics & Genomic Medicine*, *10*, e2055. https://doi.org/10.1002/mgg3.2055