Causal Inference and Structure Learning of Genotype–Phenotype Networks Using Genetic Variation

Adèle H. Ribeiro, Júlia M. P. Soler, Elias Chaibub Neto, and André Fujita

Abstract A major challenge in biomedical research is to identify causal relationships among genotypes, phenotypes, and clinical outcomes from high-dimensional measurements. Causal networks have been widely used in systems genetics for modeling gene regulatory systems and for identifying causes and risk factors of diseases. In this chapter, we describe fundamental concepts and algorithms for constructing causal networks from observational data. In biological context, causal inferences can be drawn from the natural experimental setting provided by Mendelian randomization, a term that refers to the random assignment of genotypes at meiosis. We show that genetic variants may serve as instrumental variables, improving estimation accuracy of the causal effects. In addition, identifiability issues that commonly arise when learning network structures may be overcome by using prior information on genotype–phenotype relations. We discuss four recent algorithms for genotype–phenotype network structure learning, namely (1) QTL-directed dependency graph, (2) QTL+Phenotype supervised orientation, (3) QTL-driven phenotype network, and (4) sparsity-aware maximum likelihood (SML).

Keywords Structural learning • Network analysis • Genotypes • Phenotypes • Genetic variations

A.H. Ribeiro • A. Fujita (✉)
Department of Computer Science, Institute of Mathematics and Statistics,
University of São Paulo, São Paulo, Brazil
e-mail: adele@ime.usp.br; fujita@ime.usp.br; andrefujita@gmail.com

J.M.P. Soler
Department of Statistics, Institute of Mathematics and Statistics,
University of São Paulo, São Paulo, Brazil
e-mail: pavan@ime.usp.br

E.C. Neto
Department of Computational Biology, Sage Bionetworks, Seattle, WA, USA
e-mail: elias.chaibub.neto@sagebase.org

K.-C. Wong (ed.), *Big Data Analytics in Genomics*,
DOI 10.1007/978-3-319-41279-5_3

1 Introduction

The development of high-throughput technologies, such as DNA microarrays and next- generation sequencing, has allowed the study of complex biological systems. However, the vast quantities of data from such large scale studies have been challenging researchers aiming to discover the complex network describing causal associations among genotypes, phenotypes, and other clinical outcomes.

In general, the associations found in observational studies cannot be interpreted as causal. However, genetic variants information has proven useful to determine causal effects from observational data. In several recent studies, genomic data and information on quantitative variation in phenotypes have been combined in order to discover causal relationships among phenotypes. Some of the most promising phenome projects, including the Consortium for Neuropsychiatric Phenomics at UCLA (www.phenomics.ucla.edu) and the National BioResource Project—Rat (http://www.anim.med.kyoto-u.ac.jp/NBR/), are listed in David Houle et al.'s paper [53].

Causal discovery methods from observational data are of great interest to researchers in many fields, such as functional genomics and proteomics, molecular biology, and epidemiology [39, 49, 76, 106, 107, 134].

In observational epidemiological studies, genetic variants that mimic the influence of modifiable environmental exposures have a key role in causal inference. If the link between a genetic variant and environmental exposure can indeed be shown, associations between genotype and disease outcome or genotype and intermediate phenotype may elucidate the importance of environmentally modifiable factors as causes of disease [112, 113]. These findings are crucial for the understanding of genetic mechanisms associated with diseases and for the development of therapeutic strategies [11, 72].

Causal networks describing the regulatory interactions between different genes are called gene regulatory networks (GRNs) and have been inferred from both observational and interventional gene expression data [77]. Approaches for reverse engineering GRNs from purely observational data (i.e., data collected without any biological or experimental interference on the level of individual genes) need a large sample size and capture only parts of biologically relevant networks [77]. However, it has been shown that it is possible to greatly improve accuracy and performance in network reconstruction by incorporating data from experimental interventions and perturbations (e.g., from gene knockout or knockdown experiments) [91].

It is widely accepted that the most trustworthy method to infer causal relationships from data are experimental studies such as randomized controlled trials. However, the number of variables in biological systems is usually very large, so that it is unfeasible to carry out randomized experiments to discover all possible causal relationships. Nonetheless, throughout this chapter, we show that it is possible to discover the structure of causal networks and to infer causal effects based on observational data alone if certain assumptions are met. In practical context, some of these assumptions may be very restrictive and may not necessarily hold in biological data. However, causal analysis may still be useful, since the conclusions might be indicative of some causal connections in the data [37].

Causal inference from observational studies is a complicated task that encompasses two major challenges: (1) ensuring accuracy of discovered results and (2) reducing computational complexity.

The absence or inadequacy of randomization, combined with the presence of (measured and unmeasured) confounding factors, often leads to spurious conclusions in observational studies. More reliable causal relations can be extracted from data by relying on two imperative pillars: proper randomization and instrumental variables. In systems genetics, Mendelian randomization plays an important role in causal inference. The random segregation of alleles from parents to offspring during meiosis closely resembles the random allocation of treatments (exposure variables or interventions) in randomized controlled trials. In other words, the genotype can be considered as an effect of a randomized intervention, allowing tests of causal hypotheses. Thus, a robust association from a genetic variant to a phenotype that is allegedly free of confounding factors (e.g., behavioral and environmental factors) can be interpreted as a causal relationship. Moreover, when certain assumptions are met, genetic variants can be used as instrumental variables, allowing causal inferences about the effect of phenotypes on outcomes. Causal inference can be greatly improved by exploiting instruments, since biases and effects of confounding factors are minimized [102]. In Sect. 2, we discuss in detail the Mendelian randomization approach, and particularly the conditions on which genetic variants can be defined as instrumental variables.

The computational complexity challenge of causal structure learning problem arises out of the need to develop efficient algorithms to handle large amounts of data. There is no available algorithm for finding an entire causal model in polynomial time. In other words, the causal structure learning problem is NP-hard. Since the possible number of causal networks is exponential in the number of variables, currently, some heuristics are used to limit the search space. The most commonly used approaches for reducing computational complexity assume certain network structure properties, such as sparsity and acyclicity. In Sect. 6, we review some current approaches being used for causal structure learning.

Structural equation models (SEMs) and probabilistic graphical models (PGMs) are widespread adopted methodologies for representing causal associations among variables. While SEM provides a functional representation of the causal mechanisms by which a variable's value is generated, PGM provides an equivalent but graphical representation of these causal mechanisms by using graph theory and probability theory. Thus, by combining elements from SEMs and PGMs, it is possible to model causal relationships using a mathematically rigorous and intuitive language. In Sect. 3, we present the direct correspondence between both representations.

We will address two classes of models which differ in their ability to accommodate feedback loops: (1) the non-recursive SEMs, which are capable of modeling cycles involving both just two variables (direct feedback loops or reciprocal associations) and three or more variables (indirect feedback loops); and (2) the recursive SEMs, which assume all causal effects as unidirectional, so that no two variables are causes of each other. The SEMs can be graphically represented by directed graphs in such a way that there is a direct equivalence between the two representations. More specifically, non-recursive SEMs can be graphically represented by directed cyclic graphs (DCGs), and recursive SEMs can be graphically represented by directed acyclic graphs (DAGs).

Biological systems have been extensively modeled using DAGs because algorithm development is facilitated by using results that are valid under the assumption that the causal structure is acyclic. However, for modeling cyclic phenomena, which are the most prevalent in biological systems, acyclic structures are very restrictive [130]. DCGs can be used as a more appropriate alternative for modeling data in the steady-state (equilibrium state of a time invariant dynamic system), since feedback loops can capture the redundancy and stability of the underlying system. Genetic regulatory networks have been modeled as DCGs, since they are capable of reproducing the stable cyclic pattern of gene expressions [22, 133].

Algorithms for discovering causal structures are often based on a functional representation (a recursive or non-recursive SEM) or on a graphical representation (a directed acyclic or cyclic graph) of causal processes. SEM-based structure learning approaches use optimization methods for estimating the model parameters and techniques for improving efficiency and interpretability such as sparsity-enforcing regularization. In this chapter, we focus mainly on the fundamental concepts used by causal structure learning algorithms that are based on a graphical model. In Sect. 4, we cover the main definitions and properties connecting graphical structure and probability distributions, including the concept of d-separation which allows to derive the conditional independencies entailed by a causal structure. This theory was developed mainly by Judea Pearl [87], Peter Spirtes, Clark Glymour, Richard Scheines [97, 118], and Thomas Richardson [96].

One of the main issues of the causal structure learning theory is the identifiability problem. There are some models which encode precisely the same set of conditional independence relations. Thus, they are considered statistically equivalent and indistinguishable from observational data. In this case, it is not possible to uniquely identify the true underlying causal model. By including information on genetic variants causally associated with phenotypes [e.g., quantitative trait loci (QTL) or quantitative trait nucleotide (QTN)], new conditional independence relations are created, and statistically equivalent phenotype networks may become identifiable. In Sect. 5, these concepts will be discussed in detail.

There are several algorithms available in the literature that were designed to solve the specific problem of discovering the structure of a genotype–phenotype–outcome network. Among these, in chronological order, are [5, 22, 24, 25, 32, 46, 71, 73, 74, 108, 126]. In Sect. 7, we describe in detail four of the most popular algorithms: QTL-directed dependency graph (QDG) [24], QTL + Phenotype supervised orientation (QPSO) [126], QTL-driven phenotype network (QTLnet) [25], and sparsity-aware maximum likelihood (SML) [22]. These are recent algorithms with source code freely available and easily accessible to the users.

2 Mendelian Randomization

2.1 Randomized Controlled Trial

A widely accepted approach for finding causal relationships is to perform intervention experiments, also known as randomized controlled trials. Such experiments are critically based on randomization and confounding factors control. Treatments (or interventions) are randomly assigned to the subjects and statistical tests are performed to verify whether differences between treatment and control groups are significant. For instance, to verify whether a new medication is superior in comparison with placebo, randomized controlled trials are usually conducted and the randomization is imperative to verify whether the treatment has a causal effect on the disease.

Experiments with randomization have three important implications, namely elimination of selection bias between groups, assurance of allocation concealment, and justification of randomization-based statistical tests. Under random assignment of treatments, selection biases are removed since confounding factors are more likely to be distributed evenly among groups, and statistical tests can be properly performed.

Note that the determination of a causal effect critically depends on whether all confounding factors are properly randomized. A proper randomization can indeed increase the chances of evenly distribute known and unknown confounding factors among groups. However, considering that there is a non-zero probability that confounding factors are not fully balanced among groups, it is recommended to perform some form of restricted randomization (e.g., randomization within homogeneous blocks with respect to a specific known confounding factor) when it is crucial that biases from a particular confounding factor are avoided [17].

2.2 Randomized Allocation of Allelic Variation in Genes

Unfortunately, in observational studies (where the data is collected without any intervention) an association between a phenotype and a disease (or other outcome of interest) may not be causal. The main reasons are [20, 36]:

- **Confounding variables:** suppose e are interested in the association between a phenotype and an outcome. Measured or unmeasured factors that affect both variables may create spurious associations when not considered in the model. They are called confounding variables. For instance, let X be a phenotype and Y be an outcome of interest. Consider a variable Z that directly affects both X and Y. The causal relationships among these variables can be represented by the scheme: $X \leftarrow Z \rightarrow Y$. If a pairwise correlation analysis is performed, there may be a significant association between the phenotype and the outcome, even when there is not a direct influence between them. This spurious association vanishes only if the variable Z is considered in the analysis. In this case, we say that Z is a confounder of the relationship between X and Y.

- **Reverse causation:** when two variables are causally related, but in the contrary direction to a common presumption (outcome affecting the phenotype), we say that there is a reverse causation. It is a type of misinterpretation in which the effect is allowed to occur before its cause. For instance, given a strong association between low circulating cholesterol levels and risk of cancer, one could suspect that low cholesterol levels increases the risk of cancer. However, it is possible that the causality goes in the opposite direction, i.e., early stages of cancer may, many years before diagnosis, lead to a lowering in cholesterol levels [62].
- **Various biases:** unobserved or imprecisely measured factors can bias estimates of the association between two variables even if the causal direction is correctly specified. Studies with small sample size are more affected by such biases.

These issues show that causal inference can be hard to achieve, or even an impossible task, if only observational associations are considered [112, 114]. However, it is possible to provide evidence for or against a causal relationship and, usually, to quantify causal effect by making specific assumptions and by using additional information.

The region of the genome affecting variation in a quantitative trait (phenotype) is known as quantitative trait locus (QTL), and QTLs have essentially been detected by using panels of microsatellites, mainly for population-based studies of plants and animals, and for family-based studies of humans. Quantitative trait nucleotides (QTNs) are often identified through genome-wide association studies by using single nucleotide polyomrphism (SNP) markers. Genetic variants (QTLs or QTNs) have been used to distinguish causation from association in biological studies.

Based on the Mendel's second law, alleles are randomly assigned from parents to offspring during gamete formation. This random allocation of alleles provides a design analogous to an intervention experiment. Thus, Mendelian randomization can be interpreted as a natural randomized controlled trial, in which different genotypes, rather than treatments, are randomly allocated to individuals. Considering that the variation in genotypes always precedes the differences in phenotype, this natural randomization allows us to use statistical tests to determine whether there is a causal relation from a genetic variant and a phenotype. Since the influence of genotype on phenotype is, in general, independent of any confounding, reverse causation or other biases, causal interpretation may be appropriate.

In the following, we summarize some concepts that provide a foundation for causal inference based on Mendelian randomization.

- **The law of segregation (Mendel's first law):** states that during the gamete formation the members of the allelic pair of each hereditary factor (some gene or genetic locus) segregate from each other independently so that each gamete carries only one allele for each factor and offspring acquire one allele randomly chosen from each parent. Since genetic variants segregation occurs randomly and independently of environmental factors, causal studies are less susceptible to confounding;

- **The law of independent assortment (Mendel's second law):** states that genetic variants segregate independently of other traits. In other words, alleles related to different traits are transmitted independently of one another from parents to offspring. Note that the independent assortment law is violated when two loci are linkage, i.e., when they are on the same chromosome and their genetic distance is small. In this case, the recombination fraction is less than 50 % in a single generation, that is, allele combinations in different loci are not inherited independently of each other [68].
- **Unambiguous causal direction:** since the randomization of marker alleles during meiosis precedes their effect on phenotypes, reverse causality is not an issue. In other words, the direction of the causal effect is always from the genotype to the phenotype.
- **Life-long effects:** genetic variants have life-long effects on exposures as opposed to interventions in randomized controlled trials which only occur over short periods of time.

The same advantages of a randomized controlled trial may be achieved from natural experimental setting provided by Mendelian randomization when there is no interaction with uncontrolled external confounders, such as maternal genotype and environmental perturbation. In this regard, the canalization or developmental compensation phenomenon need to be emphasized. When a genetic or environmental factor is expressed during fetal development or post-natal growth, the expression of other genetic variants may be influenced leading to changes that may alter development in such a way that the effect of the factor is damped (or buffered). This resistance of phenotypes to environmental or genetic perturbation can bias causal inferences and makes it difficult to relate randomized controlled trials and Mendelian randomization studies. In randomized controlled trials, the randomization of the intervention to subjects often occurs during their middle-age. On the other hand, in Mendelian randomization approaches, the randomization occurs before birth. Thus, we must be aware that some findings of studies using Mendelian randomization approach may be unrepresentative of clinical interventions on the exposure performed in a mature population.

Mendelian randomization has been particularly useful for investigating causal effects of an exposure of interest (phenotype) on an outcome (e.g., disease or other clinical outcome), when a genetic variant robustly associated with the exposure is not associated with any confounding factor and it is not associated with the outcome through any other path than through the exposure of interest [112]. In a statistical point of view, causal inference is improved whenever a genetic variant meets the requirements to be used as an instrumental variable [31, 68]. In this case, it is possible to estimate the long-term causal effects of exposures on outcomes. Genetic variants robustly associated with phenotypes have been used in several studies as instrumental variables, improving causal inference in a non-experimental setting [19, 20, 101, 114].

Instrumental variable analysis within the Mendelian randomization context is particularly powerful for experimental crosses (e.g., F2, backcrosses, intercrosses, etc.), which are conducted in controlled conditions and closely mimic randomized experiments. For studies in humans and other natural populations, a more careful analysis is needed, since the population structure and cryptic relatedness might still act as confounders [4]. When a genetic variant is in linkage disequilibrium with another genetic variant (i.e., alleles at the two loci are non-randomly associated), both affecting the outcome or the same metabolic pathway, the instrumental variable assumption that the genetic variant is associated with the outcome only through the exposure of interest may be violated. In the human genome, linkage disequilibrium can occur even between completely unlinked loci (e.g., alleles on separate chromosomes) due to population structure, natural selection, genetic drift, and mutation [111]. Thus, it can be quite a challenge to identify violations of the instrumental variable assumptions.

The instrumental variable approach and its assumptions are described in detail in the next section.

2.3 Genetic Variants as Instrumental Variables

Suppose a study for investigating a causal relationship between an exposure (e.g., a phenotype) X and an outcome (e.g., a clinical trait or disease) Y, when it is known a genetic variant M which is associated with X as illustrated in Fig. 1.

Considering linear relationships among the variables, the true causal equation for the outcome Y is $Y = \alpha + \beta_2 X + \beta_3 Z + \varepsilon$, where α is the regression intercept, β_2 and β_3 are direct causal effects from, respectively, X and Z, and ε is the error term.

Suppose that the simple regression model $Y = \alpha + \beta_2 X + e$, where $e = \beta_3 Z + \varepsilon$, is used in order to estimate the causal effect of X on Y, possibly because Z is an unobserved confounding factor. In this case, the ordinary least squares (OLS) estimator

$$\hat{\beta}_2 = \frac{Cov(Y, X)}{Var(X)} = \frac{Cov(\beta_2 X + \beta_3 Z + \varepsilon, X)}{Var(X)},$$

with expectation

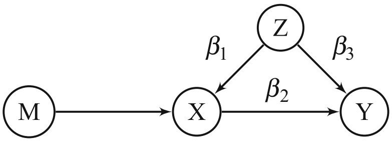


Fig. 1 Usual scenario investigated in Mendelian randomization. The causal network illustrates the assumed relationships among genetic variant M, exposure X, outcome Y and confounders Z

$$\mathbb{E}(\hat{\beta}_2) = \beta_2 + \beta_3 \frac{Cov(Z,X)}{Var(X)},$$

will have a bias of $\beta_3 \frac{Cov(Z,X)}{Var(X)}$, if $\beta_3 \neq 0$ and $Cov(Z,X) \neq 0$.

In other words, when X and Z are correlated, there is a violation of the assumption that the error e is uncorrelated with the covariate X, i.e., $cov(X,\varepsilon) = 0$, and the OLS estimator will be not asymptotically unbiased and consistent.

However, according to the instrumental variable technique introduced by the geneticist Sewal Wright [129], it is possible to determine the causal effect of X on Y, β_2, if there is a variable M (called instrumental variable) which is correlated with X, but is uncorrelated with Z.

Under the instrumental variable assumptions, i.e., $Cov(X,M) \neq 0$ and $Cov(Z,M) = 0$, the covariance of Y and Z is

$$Cov(Y,M) = Cov(\beta_2 X + \beta_3 Z + \varepsilon, M) = \beta_2 Cov(X,M).$$

Thus, we can obtain the instrumental variable (IV) estimator of β_2:

$$\hat{\beta}_2^{IV} = \frac{Cov(Y,M)}{Cov(X,M)}.$$

The instrumental variable estimator is also known as two-stage least squares (2SLS) estimator, since it can be obtained in two stages. In the first stage, X is regressed on M, and, since M is alleged to be uncorrelated with Z, the OLS estimator of the slope coefficient will be consistent and unbiased. In the second stage, Y is regressed on $\hat{X}$, which is obtained in the first stage and represents the predict value of X explained by M, but not by Z. Since the covariate $\hat{X}$ is uncorrelated with the error term, the OLS estimator is used for concluding the estimation of β_2 [12, 16].

In the Mendelian randomization scenario illustrated in Fig. 1, the genetic variant M, which is affecting the outcome Y only through its effect on the exposure X and it is not associated with confounding factors Z, is acting as an instrumental variable, allowing inferences on the causal relation of X to Y. When an intervention is made at the instrumental variable M and a significant change is detected in the outcome Y, since M is not directly associated with the outcome, the only way of explaining the indirect effect of M on Y is by a causal effect of the exposure X on the outcome Y. By using a genetic variant as an instrumental variable, it is not possible to do an intervention experiment, however, the randomization of genotypes to individuals ensures the causal inference similarly.

In order to use a genetic variant as instrumental variable, a number of assumptions must be met:

1. The genetic variant M must be associated with the exposure X;
2. The genetic variant M must be independent of measured and unmeasured factors (represented by Z) that confound the relationship between the exposure X and the outcome Y.

3. Exclusion restriction: the genetic variant M cannot affect the outcome Y by no other way than by the exposure X.

We will provide details on each of these assumptions in the following sections.

2.3.1 Statistical Association with the Exposure

The first assumption states that the genetic variant M and the exposure X must be statistically associated.

It is important that the association between M and X is strong, otherwise M is considered a weak instrumental variable and can bias estimates of the causal effects even if all other core instrumental variable assumptions are satisfied [119]. The implicit idea is that the genetic variant must be the strongest factor which genetically divides the sample into subgroups according to the exposure of interest, similar to a randomized controlled trial. It is expected that the genetic variant effect is not inhibited by the effect of any confounder.

The genetic variant is not required to be the true functional variant that produces a subsequent effect on the exposure. However, it is necessary that the chosen instrument is in linkage disequilibrium with the functional variant, i.e., they must be statistically associated.

In order to verify the association magnitude between two variables, the Pearson product moment correlation coefficient can be used if the relationship is linear. If the data present non-linear or more complicated relationships, other measures, such as the Spearman's rank correlation coefficient and the mutual information, can be better suited.

2.3.2 Independence with Exposure–Outcome Confounders

The assumption that the genetic variant M is independent of any confounder Z for the exposure–outcome relationship is specially hard to verify considering that it must hold even for unmeasured confounders.

However, it is highly recommended to examine whether any statistical association between the genetic variant and the observed confounders of the exposure exists. The absence of such statistical associations does not guarantee that the assumption is fulfilled but, at least, increases the chances of it being true.

2.3.3 Exclusion Restriction

The third and the most troubling assumption for using genetic variants as instrumental variables is sometimes referred as exclusion restriction. It states that the genetic variant only affects the outcome through the exposure. More precisely, the genetic variant must be independent of the outcome given the exposure and all confounders (measured and unmeasured) of the exposure–outcome association.

The scheme of Fig. 1 would be an exclusion restriction violation if the genetic variant M is not independent of the outcome Y conditional on exposure X and confounders Z. For example, it would be a violation if M is affecting Y through a direct edge and through the exposure X. In this case, estimates for the association between X and Y would be biased.

Since it is always possible that the genetic variant affects the outcome via a biological pathway other than the exposure of interest, it may be very difficult to guarantee that this assumption holds. In addition, pleiotropy (i.e., phenomenon of a genetic variant influencing multiple phenotypes) and linkage disequilibrium can be violations of the exclusion restriction if such associations imply the existence of another pathway by which the genetic variant is associated with the outcome [20].

A recommendation is to use only strong instruments, i.e., genetic variants whose functionality and relationship to the exposure are well biologically understood [36]. For instance, if the exposure is a protein, then the best strategy is generally to use a marker in the gene which is responsible for encoding the protein itself.

3 Causal Model

There are many definitions of causality in the philosophical and statistical literature [44, 50, 51, 55, 70, 87, 98, 104, 105]. Throughout this chapter we adopt the Pearl's definition of causality. It is a notion of causality which relies mainly on conditional probability and interventions. In order to mathematically represent interventions and distinguish them from observations, Pearl introduced the do-operator. Denoting by $do(X = x)$ the hypothetical intervention in which the variable X is manipulated to be set to the value x, and denoting by $P(y|do(x))$ the probability of the response event $Y = y$ under the hypothetical intervention $X = x$, we say that X is a cause of Y if [38, 87]:

$$P(Y = y|do(X = x)) \neq P(Y = y|do(X = x')),$$

when all background variables remain constant. Thus, Pearl's representation of causality has close resemblance to a randomized controlled experiment, in which any change in the outcome variable must be due to the intervention, if all factors influencing their association are either constant, or vary at random.

The Pearl's comprehensive theory of causation resulted from the unification of several approaches to causation, such as the graphical, potential outcome, and SEMs.

As proved by Pearl [87], SEMs provide a language for causality which is mathematically equivalent to the potential-outcome framework, developed by Jerzy Neyman [84] and Donald B. Rubin [105]. Significant contributions to the generalization of the potential-outcome framework as a general mathematical language for causal inference were also given by James Robins [98, 99].

The potential-outcome model is based on randomized experiments and counterfactual variables. By conducting randomized controlled experiments, in general, only one potential outcome is observed for each individual, which is the one corresponding to the exposure value that actually occurred for the individual. An outcome that would have occurred if, contrary to the fact, the exposure had been assigned another value is considered a counterfactual quantity according to Rubin's notation. Within the potential-outcome framework, causal inferences are made by deriving probabilistic properties of these counterfactual quantities as in a missing data problem. The equivalence between SEMs and potential-outcome models could be demonstrated by Pearl by treating counterfactual quantities as random variables. He showed that the consistency rule [98]—which states that an individual's potential outcome under a hypothetical intervention that happened to materialize is precisely the outcome experienced by that individual—is automatically satisfied in the structural model. Thus, expressions involving probabilities of counterfactuals can be converted to expressions involving conditional probabilities of measured variables [88].

The connection between SEMs and graphical models will be described in detail in the following sections. While SEMs provide a functional representation of the causal processes relating the variables, graphical models provide a visual and, thus, more intuitive representation of these relationships.

3.1 *Functional Causal Representation*

In a general context, a causal model is an SEM representing the causal relationships between random variables [87].

Consider random variables V_i, $i = 1, \ldots, n$. Let $\text{pa}(V_i)$ denote the set of the parents of V_i, that is, the set of random variables that directly determine the value of V_i. Also, let ε_i, $i = 1, \ldots, n$ be random variables representing errors due to unknown causes.

The mechanism by which the value of each variable V_i is selected can be defined according to some function f_i (usually, but not necessarily, linear) of the parent variables and of the error variable:

$$V_i = f_i(\text{pa}(V_i), \varepsilon_i), \quad i = 1, \ldots, n. \tag{1}$$

In the context of genotype–phenotype causal networks and Mendelian randomization, the random variables are quantitative phenotypes and variant genetics associated with these phenotypes (QTLs or QTNs). In addition, it is commonly assumed that the functional relationships of the SEM shown in Eq. (1) are linear, and that the effects of genetic variants are fixed.

Before introducing the model, let us first discuss the involved notation and objects. Let:

- $\mathbf{Y}$ be a $p \times n$ matrix where each element y_{ij} represents the observed value of the ith quantitative phenotype for the jth individual;
- $\mathbf{M} = \mathbf{1}' \otimes \boldsymbol{\mu}$ be a $p \times n$ matrix resulting from the Kronecker product between the transposed n-dimensional unity vector, and $\boldsymbol{\mu}$ be the vector with the expected values of each quantitative phenotype;
- $\mathbf{Q}$ be a $p \times q$ matrix of effects of genetic variants on the phenotypes. The element q_{ij} represents the effect of the jth genetic variant on the ith phenotype;
- $\mathbf{X}$ be a $q \times n$ matrix where each element x_{ij} represents the predicted genotype of the ith genetic variant in the jth individual. For SNPs the observed genotype state is used instead of the predicted values;
- $\mathbf{P}$ be a $p \times p$ matrix containing the direct causal effects of the phenotypes on each other. The element p_{ij} corresponds to the effect of the jth phenotype on the ith phenotype;
- $\mathbf{E}$ be a $p \times n$ matrix where each e_{ij} represents the measurement error of the ith phenotype for the jth individual.

The causal linear SEM representing the expected pattern of associations among p observed phenotypes and q genetic variants for n individuals is

$$\mathbf{Y} = \mathbf{M} + \mathbf{PY} + \mathbf{QX} + \mathbf{E}. \tag{2}$$

Note that the assumption that variations in QTL or QTN genotypes precede variation in the phenotypes is expressed by the fact that the $\mathbf{Q}$ matrices is always in the right side in Eq. (2), along with the parent phenotype effects, represented in $\mathbf{P}$ matrix.

In the case of biallelic genetic variants, the total genetic effect (represented in Q and X matrices) can be partitioned in order to explicit particular effects, such as additive and dominance effects of each genetic locus and the possible interaction effects among them (epistatic effects). Typically biallelic genetic variants have three genotype states, denoted usually as AA (homozygous dominant), Aa (heterozygous), and aa (homozygous recessive). These genotypes must be encoded according to the type of effect by using two degrees of freedom.

We will illustrate this extension in order to take into account additive and dominance effects. For simplicity, we will not consider interactions between genetic variants (epistasis).

Let us precisely define the components of dominance and additive effects using biallelic genetic variants. Let:

- $\mathbf{Q^a}$ be a $p \times q$ matrix of the additive effects. The element q^a_{ij} represents the additive effect of the jth genetic variant on the ith phenotype;
- $\mathbf{X^a}$ be a $q \times n$ matrix where each element x^a_{ij} represents the predicted genotype of the ith genetic variant for the jth individual, properly encoded to represent

additive effects. For instance, we can encode the genotypes aa, Aa, and AA as −1, 0, and 1, respectively.

- $\mathbf{Q^d}$ be a $p \times q$ matrix of the dominance effects. The element q^d_{ij} represents the dominance effect of jth genetic variant on the ith phenotype;
- $\mathbf{X^d}$ be a $q \times n$ matrix where each element x^a_{ij} represents the predicted genotype of the ith genetic variant for the jth individual, properly encoded to represent dominance effects. Since the dominance effects are due the interaction between the alleles, a possible encoding for the genotypes is 1 for a heterozygous individual, and 0 otherwise.

In this specific case, the causal model for the genotype–phenotype network is

$$\mathbf{Y} = \mathbf{M} + \mathbf{PY} + \mathbf{Q^aX^a} + \mathbf{Q^dX^d} + \mathbf{E}. \tag{3}$$

The interaction effects between genetic variants (describing epistasis) were not illustrated in Model 3. However, we could easily incorporate in the model matrices $\mathbf{Q^{aa}}$, $\mathbf{Q^{dd}}$, and $\mathbf{Q^{ad}}$ representing epistatic interaction effects [27].

Since dominance effects (interaction between alleles) and epistasis (interaction between loci) are higher order effects, it is possible that they have little impact on the inferences about the response variable. Supporting this idea, Burgess et al. (2011) [21] suggest to include only the most important instrumental variables (genetic variants), based on biological knowledge, for a parsimonious modeling of the genetic association (i.e., per allele additive genetic model, rather than using the total degrees of freedom in terms of effects).

However, the literature has shown that many genetic variants are not precisely identified because of the simplicity of the adopted models [75, 120, 132]. So, while lower order effects (additive effects) may be sufficient for genetic mapping, interaction effects may be decisive to analyze complex diseases (as opposed to the Mendelian diseases).

In order to draw causal inferences from observational studies, it has been suggested to select the most relevant instrumental variables (genetic determinants of the exposure) attempting to be as parsimonious as possible. However, when only a small proportion of the variability in the exposure is explained by the genetic variant, it is possible to improve the precision of estimates by using multiple genetic variants [85, 90].

3.2 Graphical Causal Representation

The causal model in Eq. (1) can be graphically represented by a directed graph.

Definition 1. Let $\mathbf{V} = \{V_1, \ldots, V_n\}$ be a finite set and $\mathbf{E} \subseteq \{(V_i, V_j) : V_i, V_j \in \mathbf{V}\}$ a set of ordered pairs of vertices. Each element of the set $\mathbf{V}$ is called **vertex** and each element of the set $\mathbf{E}$ is called **directed edge**. The edge (V_i, V_j) represents a direct connection from V_i to V_j. The ordered pair $G = (\mathbf{V}, \mathbf{E})$ is called **directed graph** or **digraph**.

We can always build a graphical representation of an SEM by using a directed graph.

In this representation, each vertex of a directed graph corresponds to a distinct random variable. In addition, each edge (V_i, V_j), if it exists, represents a direct functional relationship from the variable V_i to the variable V_j. In this case, V_i is called parent of V_j and V_j is a child of the vertex V_i. The absence of an edge indicates both variables are not directly associated. Thus, if we draw, for each variable V_i, an edge pointing to it from each of its parents, we can build the directed graph which represents their causal mechanisms.

The error terms are not represented in the graph. However, when error terms are correlated, the corresponding pairs of variables must be connected by a bidirected (double-headed) edge.

When the relationships imply causality, the graphical representation of a causal model is referred as *causal graph* or *causal diagram* of the system. The goal of the causal structure learning methods is to discover the causal graph of a system often from observational data.

In the graphical representation of the functional model shown in Eq. (2), the vertices can represent phenotypes or genetic variants and the edges represent the causal relationships among them.

The next definitions introduce concepts that distinguish two classes of graphs according to whether or not the graph structure has cyclic patterns. This distinction is important because many results and procedures for inferring causal relationships using observational data are dependent on the type of graph structure.

Definition 2. A **directed path** between two vertices is a sequence of directed edges that begins at one vertex and ends at another vertex, with the restriction that all the edges are oriented in the same direction. Whenever there is a path that begins and ends at the same vertex we have a **cycle**. Cycles of length one are called **self-loops** and cycles of length two corresponds to a **bidirectional influence** or **reciprocal association**.

Definition 3. An SEM with uncorrelated error terms and which does not contain cyclic relationships is called **recursive SEM** and its graphical representation is called **directed acyclic graph (DAG)**.

Real biological systems such as GRNs often have natural cyclic behavior [28]. Thus, DAGs can be very restrictive to model such biological data. Directed graphs that can accommodate cycles and reciprocal associations have been used to model feedback processes that have reached equilibrium. For instance, equilibrium expression patterns can be modeled in reverse engineering GRNs from multiple gene expression measurements [26].

An alternative interpretation for cycles is that each feedback relation represents an infinite sequence of variables indexed by time. Thus, a cyclic graph can be viewed as a compact representation of an infinite acyclic graph [96, 117].

Definition 4. An SEM which contains at least one cycle is called **non-recursive SEM** and its graphical representation is called **directed cyclic graph (DCG)**. Note that systems with correlated error are also non-recursive, since the corresponding pair of variables are connected by a bidirected edge.

DCGs have been used to represent GRNs. In this representation, vertices are gene expression levels of genes and directed edges indicate regulation processes. The expression of a gene can be controlled by the presence of proteins called activators and repressors (or inhibitors). Thus, the gene in the tail of the edge produces a protein that regulates the gene in the head of the edge. In this case, the genome itself consists in a complex network [26, 83].

From an algebraic point of view, the edges connecting phenotypes correspond to the non-zero elements in **P** and the edges pointing from a QTL or QTN to a phenotype exist if the corresponding entries in $\mathbf{Q^a}$ or in $\mathbf{Q^d}$ are non-zero. If the **P** matrix can be rearranged as a lower triangular matrix, then we have a recursive model and, consequently, it can be represented as a DAG. Otherwise, the system contains cycles and a non-recursive SEM and a DCG are used to represent it.

Any SEM can be represented by directed graphs, even when the system involves cycles, self-loops, dependent errors, and nonlinearities. In biological context, causal models often represent linear relationships among phenotypes and genetic variants. It is commonly assumed that the system does not contain self-loops and the error terms are uncorrelated.

4 Properties Relating Functional and Graphical Models

To provide a statistical connection between the functional and the graphical representation of a causal model, some concepts are fundamental, namely conditional independence, d-separability, directed Markov property, and causal faithfulness.

The graphical model (a directed acyclic or cyclic graph) that precisely encodes the conditional independence relations among the variables of the system is called PGM. When that precise connection between graphical and functional representations can be established, some theoretical results can be used for causal inference and network structure learning. These concepts are presented in the following sections.

4.1 d-Separability

The concept called d-separation is a fundamental criterion used in network structure discovery algorithms. In fact, it can determine whether or not a directed edge exists between two variables. Under d-separation criterion, it is even possible to determine the direction of some edges.

Before giving a precise definition of d-separation, it will be introduced some related concepts: conditional and unconditional independence and undirected path. These concepts can be defined on random variables or on sets of random variables as follows:

Definition 5. Let $\mathbf{V} = \{V_1, V_2, \dots V_n\}$ be a set of random variables. Consider $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ three subsets of $\mathbf{V}$ and P the joint probability distribution function over the variables in $\mathbf{V}$. We say that the sets $\mathbf{X}$ and $\mathbf{Y}$ are **conditionally independent** given $\mathbf{Z}$ if, for any configuration x of the variables in the set $\mathbf{X}$ and for any configurations y and z of the variables in the sets $\mathbf{Y}$ and $\mathbf{Z}$ satisfying $P(\mathbf{Y} = y, \mathbf{Z} = z) > 0$, we have

$$P(\mathbf{X} = x|\mathbf{Y} = y, \mathbf{Z} = z) = P(\mathbf{X} = x|\mathbf{Z} = z).$$

This relationship is denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_P$ or simply $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$.

Definition 6. Using the same notations of the Definition 5, $\mathbf{X}$ and $\mathbf{Y}$ are **unconditionally independent** or **marginally independent** if

$$P(\mathbf{X} = x|\mathbf{Y} = y) = P(\mathbf{X} = x)$$

for any configurations x and y of the variables in the sets $\mathbf{X}$ and $\mathbf{Y}$ satisfying $P(\mathbf{Y} = y) > 0$.

This relationship is denoted by $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\emptyset$ or simply $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$.

The conditional and unconditional independencies encoded by a given directed cyclic or acyclic graph can be determined by a graphical criterion based on the definitions of undirected path and collider:

Definition 7. Let $G = (\mathbf{V}, \mathbf{E})$ be a directed graph. A sequence of distinct edges $\{E_1, \dots, E_k\}$ in G is an **undirected path** if there exists a sequence of vertices $\{V_i, \dots, V_{k+1}\}$ such that for $1 \leq i \leq k$ either $(V_i, V_{i+1}) = E_i$ or $(V_{i+1}, V_i) = E_i$, and $E_i \neq E_{i+1}$. An **acyclic undirected path** is an undirected path in which every vertex in the path occurs no more than once.

In words, an undirected path is a sequence of connected edges ignoring their directions. It is also common to define undirected path as an ordered sequence of vertices that must be transversed, ignoring the direction of the edges. However, this definition is only valid for acyclic graphs, since a pair of vertices can uniquely identify an edge. A proper definition of undirected path which is valid for structures with reciprocal associations uses a sequence of edges rather than a sequence of vertices [117].

Definition 8. Let X, Y, and Z be vertices of a graph and U be an undirected path containing X, Y, and Z in this order.

Y is a **collider** in U if there are edges pointing into it from both X and Z (i.e., Y is common effect of X and Z), preventing transmission of causal effects along such a path.

When Y is a collider and, additionally, X and Z are not connected by an edge, Y is called **unshielded collider** [118]. In addition, the formation $X \rightarrow Y \leftarrow Z$ is called **v-structure** (using Pearl's notation [87]) or **immorality** (using Koller and Friedman's notation [64]).

Having introduced such fundamental concepts, d-separation can be defined as follows [87]:

Definition 9. Let $G = (\mathbf{V}, \mathbf{E})$ be a causal graph and $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be disjoint sets of vertices of $\mathbf{V}$. $\mathbf{X}$ and $\mathbf{Y}$ are **d-separated** given $\mathbf{Z}$ in G, if for any undirected acyclic path U between a vertex in $\mathbf{X}$ and a vertex in $\mathbf{Y}$:

- U contains an unshielded collider such that neither the middle vertex (the collider) nor any of its descendants is in $\mathbf{Z}$; or
- U contains a vertex which is not a collider and it is in $\mathbf{Z}$.

When $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in the graph G we write $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$.

Using d-separation criterion, the premise of conditional independence can be observed under two assumptions: global directed Markov and causal faithfulness. We will precisely define them in the following sections, but they assure us that two vertices are conditionally independent given a set of variables $\mathbf{Z}$ if and only if $\mathbf{Z}$ d-separates both variables. Thus, the graphical property of d-separation enables us to determine what conditional independence relations are entailed by a given graphical causal model. For each pair of variables, we can test whether they are independent given all sorts of conditioning variables sets.

Sometimes it is possible to prune away edges that represent spurious associations or even to orient edges using observational data alone. Whenever both variables become independent by conditioning on other variables, we can rule out the edge between them. In addition, unshielded collider formations can be identified testing if two variables become dependent by conditioning on the collider. As the edges going into colliders are oriented, orientations of other edges can be induced.

For instance, consider the following d-separation statements present in the graph of Fig. 2:

$$B \perp\!\!\!\perp C \mid A; \quad A \perp\!\!\!\perp D \mid \{B, C\}; \quad B \perp\!\!\!\perp E \mid D; \quad C \perp\!\!\!\perp E \mid D; \quad A \perp\!\!\!\perp E \mid D.$$

These pairs of vertices are d-separated because every path between them is blocked, that is: (1) in every path containing a collider, the collider and its descendants are not in the conditioning set; and (2) in every path containing only non-colliders and non-descendant of colliders, at least one vertex of the path is in the conditioning set. Therefore, we can conclude that there is not an edge connecting them.

We can orient some of the edges if more tests are made. For instance, we can identify the unshielded collider $C \rightarrow D \leftarrow B$, knowing that $B \not\perp\!\!\!\perp C \mid \{A, D\}$. After orienting these edges, we can also identify the true direction between D and E. By conditioning to D, neither the path between B and E ($B \perp\!\!\!\perp E \mid D$) nor the path

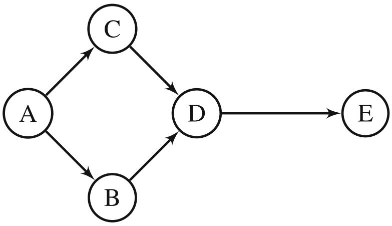


Fig. 2 A DAG representing causal relationships among five variables. By using d-separation criterion, only the direction of the edges $A \rightarrow C$ and $A \rightarrow B$ cannot be recovered

between C and E ($C \perp\!\!\!\perp E \mid D$) can be blocked, implying that D is not a collider. Thus, the edge $D \rightarrow E$ can be recovered.

Since conditional independence is a symmetric relationship, we cannot orient the two remaining edges. Even knowing that A, B, and C are not colliders, we cannot discard any of the three following orientations because for all of them we have the d-separation statement $B \perp\!\!\!\perp C \mid A$:

- A chain: $B \rightarrow A \rightarrow C$;
- Another chain: $B \leftarrow A \leftarrow C$;
- A fork: $B \leftarrow A \rightarrow C$.

Graphs with the same set of d-separation statements usually correspond to observationally equivalent models and their structures cannot be fully recovered using observational data alone. In Sect. 5, we will discuss the equivalence problem in more detail.

4.2 Global Directed Markov Property

In order to provide a probabilistic interpretation of the graphs, it is necessary to introduce a property to ensures that the graph with a set of vertices **V** can also represent a set of probability distributions over V.

If the graph accurately describes the structure entailed by a causal model, then the separation properties of the graph can be associated with conditional independencies and causality relations among variables. In other words, we can use d-separation criterion as a graphical tool to recover the underlying causal mechanisms relating the variables.

Definition 10. Let G be a directed acyclic or cyclic graph with a probability distribution P. We say that P satisfies the **global directed Markov property** for G if for all disjoints sets of variables **X**, **Y**, and **Z** the following statement is true: if **X** is d-separated from **Y** given **Z** in G, then **X** is conditionally independent from **Y** given **Z** in P.

In other words, we say that P satisfies the global directed Markov property for G when:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_P, \quad \text{foralldisjointsets}\mathbf{X}, \mathbf{Y}\text{and}\mathbf{Z}.$$

Assuming the global directed Markov property holds, the conditional independencies found by d-separation in a given graph G hold for every causal model that can be represented graphically by G.

A wide range of statistical models, including recursive linear SEMs with independent errors, regression models, and factor analytic models, satisfy the global directed Markov condition for its associated DAG.

Considering recursive and non-recursive SEMs with independent errors, if the relationships are linear, then the following result guarantees that the probability distribution P of the SEM satisfies the global directed Markov property [65, 117]:

Theorem 1. *Let L be a* linear *recursive (or non-recursive SEM) with jointly independent error variables and G be the directed acyclic (or cyclic) graph naturally associated with L. Consider the probability distribution P over the variables of L.*

Under these conditions, P satisfies the global directed Markov property for G.

Thus, the Theorem 1 allows us to use d-separation criterion in a graph to read off the conditional independence relations entailed by the associated linear SEM. In addition, all conditional independence relations which hold in a linear SEM are precisely encoded by its natural graphical representation.

The natural graphical representation of a non-recursive SEM is a DCG. It has been shown that there is no DAG that is capable of encoding the conditional independence relations entailed by a non-recursive SEM [94].

4.2.1 Local Directed Markov Property in DAGs

By construction, in DAGs, every variable is d-separated from its non-descendants given its parents. Thus, for DAGs, there is a local property equivalent to the global directed Markov property [67]:

Definition 11. Consider a directed **acyclic** graph G with a probability distribution P, both defined over a set of random variables $\mathbf{V} = \{V_1, \ldots, V_n\}$.

We say that P satisfies the **local directed Markov property** with respect to G if for every variable $V_i \in \mathbf{V}$, in the probability distribution P, V_i is independent of all other non-descendants variable (all other vertices except its parents and descendants), given its parents in G.

In Pearl's terminology [86] we say that G is an **independency map** (or **I-map**) of P when all the Markov assumptions implied by G are satisfied by P.

Thus, the local directed Markov property is sufficient to relate an acyclic graphical representation G to a probability distribution P. The equivalence of the global and local directed Markov properties in DAGs holds even when the probability distributions represented by the graph have no density function [67].

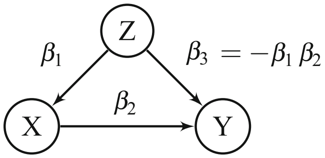


Fig. 3 Example of an unfaithful distribution to the graph. The direct effect of Z on Y is exactly the additive inverse of the indirect effect through X, leaving no total effect

According to Theorem 1, the probability distribution P of a linear recursive SEM L with jointly independent error variables satisfies the global Markov property for the DAG G that provides the natural graphical representation of L. Moreover, given the equivalence of the global and local directed Markov properties, P also satisfies the local directed Markov property for G.

We point out, however, that the local and global Markov properties are not equivalent for directed *cyclic* graphs (DCGs) [117].

4.3 Causal Faithfulness

Assuming that a probability distribution P satisfies the global directed Markov property for a graph, all d-separation statements obtained graphically hold as conditional independence statements in P. However, this assumption does not ensure that all conditional independence statements implied by P are represented in the graph.

Consider, for instance, the system in Fig. 3 represented by an DAG with three vertices, X, Y, and Z, such that Z directly affects Y, but it also affects Y indirectly, mediated by X.

Assuming that the relations are linear, the total effect of one variable on another is the sum of its direct effect and indirect effects [15]. Moreover, the indirect effect can be calculated by using Sewall Wright's multiplication rule [34, 128], i.e., by multiplying the structural coefficients on the corresponding path.

In Fig. 3, the direct effect is given by β_3, and the indirect effect is given by the product of the coefficients on path through X, that is, $\beta_1\beta_2$. By summing the direct and indirect effects, we have the total effect of Z on Y is equal to $\beta_3 + \beta_1\beta_2$. Since in this specific case β_3 is defined as $-\beta_1\beta_2$, the total effect is equal to zero.

In order to clarify a bit more the calculation of the total effect under linearity condition, consider the corresponding recursive linear SEM with independent errors that can be derived by describing each variable as a linear function of its parents and of an error variable:

$$\begin{aligned} Y &= \beta_3 Z + \beta_2 X + \varepsilon_Y \\ X &= \beta_1 Z + \varepsilon_X. \end{aligned} \tag{4}$$

By substituting the expression for X into the expression for Y, we can express Y as a function of Z and error variables:

$$\begin{aligned} Y &= \beta_3 Z + \beta_2(\beta_1 Z + \varepsilon_X) + \varepsilon_Y \\ &= (\beta_3 + \beta_2\beta_1)Z + \beta_2\varepsilon_X + \varepsilon_Y. \end{aligned} \tag{5}$$

Thus, the total effect of Z on Y is given by $\beta_3 + \beta_1\beta_2$ and it vanishes when the parameter β_3 is set to exactly $-\beta_1\beta_2$.

Although there are no conditional independencies entailed for all values of free parameters, with that specific choice of the β_3 parameter, the direct effect is cancelled out by the indirect effect and Z and Y will be apparently not associated. In such a case, we say that the population is unfaithful to the graph of the causal structure that generated it.

Under the assumption that a probability distribution P is faithful to a graph, we have the guarantee that the conditional independencies entailed by P can be read off from the graph by applying d-separation criterion.

Definition 12. Let G be a directed acyclic or cyclic graph G with a probability distribution P. We say that P satisfies the **causal faithfulness condition** for G if for all disjoints sets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, the following statement is true: if $\mathbf{X}$ is conditionally independent from $\mathbf{Y}$ given $\mathbf{Z}$ in P, then $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in G.

In other words, we say that P satisfies the causal faithfulness condition for G when:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_P \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G, \quad \text{for all disjoint sets } \mathbf{X}, \mathbf{Y} \text{ and } \mathbf{Z}.$$

The following theorem allow us to test if a probability distribution is faithful to a directed *acyclic* graph [118]:

Theorem 2. *Let G be a DAG with a probability distribution P. If P is faithful to some DAG, then P is faithful to G if and only if*

1. *for any vertices X and Y of G, X, and Y are adjacent if and only if X and Y are dependent conditional on every set of vertices of G that does not include X or Y; and*
2. *for any vertices X, Y, and Z, such that X is adjacent to Y, Y is adjacent to Z, and X is not adjacent to Z, $X \rightarrow Y \leftarrow Z$ is a subgraph of G if and only if X and Z are dependent conditional on every set containing Y but not X or Z.*

4.4 Factorization of Joint Probability Distribution Functions

Consider an arbitrary probability distribution function P, defined on n random variables $V_1, \dots V_n$. By successive application of the chain rule of probability, P can be factorized as a product of n conditional probability distribution functions:

$$P(V_1, V_2, \dots, V_n) = \prod_j P(V_j | V_1, \dots, V_{j-1}).$$

The probability distribution function P can be either specified by a probability mass function (for qualitative random variables) or by a probability density function (for quantitative random variables).

We will present a definition for the factorization of a joint probability *density* function according to a directed graph which may be either acyclic or cyclic. It was first proposed by Lauritzen et al. [67] for DAGs and its generalization was due to Thomas Richardson and Peter Spirtes [97, 117].

To precisely define that factorization property, let us first introduce the concept of ancestor of a vertex:

Definition 13. Let $G = (\mathbf{V}, \mathbf{E})$ be a directed graph. A vertex $V_i \in \mathbf{V}$ is an **ancestor** of a vertex $V_j \in \mathbf{V}$, if there is an acyclic directed path from V_i to V_j or $V_i = V_j$.

With the concept of ancestor, we can now define the factorization property of joint densities according to directed cyclic or acyclic graph.

Definition 14. Let $G = (\mathbf{V}, \mathbf{E})$ be a directed graph and $\mathbf{X}$ be a subset of $\mathbf{V}$. Consider the probability density function $f(\mathbf{V})$ for a probability distribution P with respect to a product measure μ over $\mathbf{V}$ (i.e., $P = f \cdot \mu$). Denote by $f(\mathbf{Y})$ the marginal of $f(\mathbf{V})$ for a subset $\mathbf{Y}$ of $\mathbf{V}$. Also, denote the set of ancestors of members of $\mathbf{X}$ by $An(\mathbf{X}, G)$ and the set of parents of a vertex X_i in G by $\mathrm{pa}(X_i, G)$.

For a non-negative function g, we say that P **factors according to the directed graph** G if for every subset $\mathbf{X}$ of $\mathbf{V}$,

$$f(An(\mathbf{X}, G)) = \prod_{X_i \in An(\mathbf{X}, G)} g(X_i, \mathrm{pa}(X_i, G)).$$

Directed *acyclic* graphs are built in such a way that each variable is d-separated from its non-descendants given its parents. For this reason, it is possible to use a simpler form of Definition 14. Moreover, the factorization property according to a DAG is defined for any probability distribution function (defined either over quantitative or qualitative random variables):

Definition 15. Let G be a directed **acyclic** graph and P be a probability distribution function, both defined over a set of random variables $\mathbf{V} = \{V_1, \dots, V_n\}$. Denote by $\mathrm{pa}(V_j, G)$ the set of parents of a vertex V_j in G.

We say that P **factors according to the directed acyclic graph** G if P can be written as the product of the individual distribution functions, conditional on their parent variables:

$$P(V_1, \ldots, V_n) = \prod_j P(V_j | \,\mathrm{pa}(V_j, G)).$$

4.4.1 Factorization and Global Markov Property

The following results relate factorization of probability density functions and global directed Markov property. Their proofs are due to Thomas Richardson and Peter Spirtes [97] and are based on the proofs by Lauritzen et al. [67] for DAGs.

Theorem 3. *Let P be a probability distribution that is absolutely continuous with respect to a product measure μ (and, thus, it has a non-negative probability density function).*

If P factors according to a directed graph G, then P satisfies the global directed Markov property for G.

The Theorem 3 states that if a probability distribution complies with the Definition 14, then it also satisfies the global directed Markov property for the corresponding directed (acyclic or cyclic) graph.

The reverse direction holds for acyclic graphs under the same hypotheses. However, to extend this result for cyclic graphs, a further constraint on the probability distribution P is necessary: it must has a *strictly positive* probability density function f [97, 117].

Theorem 4. *Let P be a probability distribution, defined over a set of random variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, that is absolutely continuous with respect to a product measure μ, and has a positive probability density function $f(\mathbf{V})$.*

If P satisfies the global directed Markov property for a directed (cyclic or acyclic) graph G, then $f(\mathbf{V})$ factors according to G.

The proof of Theorem 4 by Thomas Richardson and Peter Spirtes uses some ideas of Lauritzen et al. [67], which are mainly based on the moralized version of a graph (i.e., the undirected version of the graph obtained after connecting or marrying the parents of each immorality). The positivity assumption is needed to use the Hammersley–Clifford theorem. It gives necessary and sufficient conditions under which a *positive* probability distribution factorizes according to an undirected graph. A discussion on the problems involved is given by Terry Speed [115].

To conclude this section, we want to emphasize three statements that are equivalent for a probability distribution function P associated with the directed *acyclic* graph G [67]:

- P satisfies the global directed Markov property with respect to the DAG G;
- P satisfies the local directed Markov property with respect to the DAG G;

- P factors according to the DAG G, i.e., the joint probability distribution can be expressed by a product of conditional distributions for each variable given its parents. For instance, considering the system shown in Fig. 2,

$$P(A,B,C,D,E) = P(E|D)\ P(D|C,B)\ P(C|A)\ P(B|A)\ P(A).$$

4.5 Linear Entailment and Partial Correlations

In practice, algorithms for causal structure learning assume a d-separation oracle which precisely tell us whether two variables are d-separated in a directed graph given a set of other variables. Thus, we can ask the oracle whether two variables are d-separated given every possible conditioning set. If it is possible to find a conditioning set that makes two variables d-separated, then we can conclude that does not exist an edge connecting these two variables. That is a fundamental idea for reconstructing association networks (an undirected graph representing only direct associations among variables). For instance, this idea is used in the first step of the classical PC-algorithm, called PC-skeleton algorithm [118]. The direction of each edge connecting two variables that could not be d-separated by any conditioning set is determined in subsequent steps of structure learning algorithms.

Under the assumption that all involved variables have a joint multivariate normal distribution, a zero partial correlation ties with conditional independence. In addition, as shown in Definition 12, the faithfulness assumption ensures that conditional independence implies d-separability. Thus, under faithfulness and normality assumptions, it is possible to apply d-separation criterion by using a statistical test for zero partial correlations. The level of statistical significance of the test is decisive for the determination of direct associations between variables. The choice of the significance level depends on the maximum acceptable probability of making a type I error, but the most commonly used significance levels are 1 % and 5 %. In the PC-skeleton algorithm of the R package pcalg [60] (which is used in the first step of the QDG [24] and QPSO [126] genotype–phenotype discovery algorithms), the default significance level for individual partial correlation tests is 1 %. However, in the R package QTLnet [82], which implements the QDG algorithms, the suggested significance level is very small (equals to 0.05 %), probably to compensate for multiple comparisons.

The partial correlation measures the strength of the linear association between two continuous variables when the effect of a set of other random variables is controlled. We define the partial correlation coefficient in the following [59]:

Definition 16. Let X and Y be two random variables and $\mathbf{Z} = (Z_1, \ldots, Z_p)$ a set of p other random variables. Let μ_X and μ_Y be the means of X and Y, respectively, and $\boldsymbol{\mu}_{\mathbf{Z}}$ the mean vector of $\mathbf{Z}$. Also, let Σ be the covariance matrix of $(X, Y, Z_1, \ldots, Z_p)$ with the following partition notations:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ_1} & \cdots & \sigma_{XZ_p} \\ \sigma_{YX} & \sigma_{YY} & \sigma_{YZ_1} & \cdots & \sigma_{YZ_p} \\ \sigma_{XZ_1} & \sigma_{YZ_1} & \sigma_{Z_1Z_1} & \cdots & \sigma_{Z_pZ_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{XZ_p} & \sigma_{YZ_p} & \sigma_{Z_1Z_p} & \cdots & \sigma_{Z_pZ_p} \end{pmatrix} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} & \boldsymbol{\Sigma}_{X\mathbf{Z}} \\ \sigma_{YX} & \sigma_{YY} & \boldsymbol{\Sigma}_{Y\mathbf{Z}} \\ \boldsymbol{\Sigma}'_{X\mathbf{Z}} & \boldsymbol{\Sigma}'_{Y\mathbf{Z}} & \boldsymbol{\Sigma}_{\mathbf{ZZ}} \end{pmatrix}.$$

The prediction errors of X and Y given $\mathbf{Z}$ when using the best linear predictors (which minimize the mean square error) are, respectively,

$$X - \mu_X - \boldsymbol{\Sigma}_{X\mathbf{Z}}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})$$
$$Y - \mu_Y - \boldsymbol{\Sigma}_{Y\mathbf{Z}}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}),$$

with error covariance matrix that can be calculated as

$$\Sigma_{XY\cdot\mathbf{Z}} \doteq \begin{pmatrix} \sigma_{XX\cdot\mathbf{Z}} & \sigma_{XY\cdot\mathbf{Z}} \\ \sigma_{YX\cdot\mathbf{Z}} & \sigma_{YY\cdot\mathbf{Z}} \end{pmatrix} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Sigma}_{X\mathbf{Z}} \\ \boldsymbol{\Sigma}_{Y\mathbf{Z}} \end{pmatrix} \boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}} \left(\boldsymbol{\Sigma}'_{X\mathbf{Z}} \; \boldsymbol{\Sigma}'_{Y\mathbf{Z}}\right).$$

The partial correlation coefficient between X and Y eliminating the effect of $\mathbf{Z}$ is defined by the correlation between the prediction errors of X and Y given $\mathbf{Z}$, determined from $\Sigma_{XY\cdot\mathbf{Z}}$:

$$\rho_{XY\cdot\mathbf{Z}} = \frac{\sigma_{XY\cdot\mathbf{Z}}}{\sqrt{\sigma_{XX\cdot\mathbf{Z}}}\sqrt{\sigma_{YY\cdot\mathbf{Z}}}}.$$

The partial correlation coefficient shown in Definition 16 can be estimated using the sample covariance matrices. In the case of the variables having a joint multivariate normal distributed, that sample partial correlation coefficient is the maximum-likelihood estimator.

It has also been shown that zero partial correlation and conditional independence are equivalent only in Gaussian distribution [6]. However, the d-separation oracle does not necessarily need to be a statistical test for conditional independence. It can be any statistical constraint that provides the d-separability relations in a graph.

In the following, we will show some conditions linking zero partial correlation with d-separation, without any normality assumption. The main assumption is linearity of the relations among the variables.

It is noteworthy that randomization can provide the basis for making inferences without assuming a particular distribution [81]. Thus, randomization-based hypothesis tests can be used within Mendelian randomization approach when the assumption of normality is not met.

To state precisely the conditions that must be satisfied, let us first introduce some notation. Consider L a *linear* SEM with jointly independent error terms and G the directed graph corresponding to L. Note that L can be a recursive or non-recursive SEM. Thus, it can be associated with a directed acyclic or cyclic graph G. Consider also the notation $(X \perp\!\!\!\perp Y|Z)_L$ to say that X is independent of Y given Z in L, the notation $(X \perp\!\!\!\perp Y|Z)_G$ to say that X is d-separated of Y in G, and the notation $\rho_{XY \cdot Z}$ for the partial correlation of X and Y given Z.

Definition 17. Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be disjoint sets of random variables.

We say that L **linearly entails** the conditional independence relation between $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$ (i.e., $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_L$) when $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_L$ for all values of the non-zero linear coefficients and all distributions of the exogenous variables in which they are jointly independent and have positive variances.

Theorem 5. *Using the same notation of the definition 17,* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$ *if and only if L linearly entails that* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_L$.

Even if the relations between the variables are *non-linear*, d-separation is a necessary condition for a conditional independence claim to be entailed by a recursive or non-recursive SEM. Thus, whenever two variables are conditionally independent according to an SEM, they can be d-separated in the associated directed graph. However, d-separation is a sufficient condition for conditional independence only in recursive SEMs and linear non-recursive SEMs. There are non-linear non-recursive SEMs in which a d-separation relation exists in the naturally associated graph, but their conditional independence is not entailed by the model. Peter Spirtes showed that in a modified graphical representation of the SEM, called collapsed graph, d-separation statements imply conditional independence relations [117].

In the following, we show some concepts linking d-separation and partial correlation in a linear SEM.

Definition 18. Let $\mathbf{Z}$ be a set of random variables and X and Y be random variables such that $X \neq Y$ and X and Y are not in $\mathbf{Z}$.

We say that L **linearly entails** the zero partial correlation between X and Y given $\mathbf{Z}$ (i.e., $\rho_{XY \cdot \mathbf{Z}} = 0$) when $\rho_{XY \cdot \mathbf{Z}} = 0$ for all values of the non-zero linear coefficients and all distribution of the exogenous variables in which each pair of exogenous variables has zero correlation, each exogenous variable has positive variance, and in which $\rho_{XY \cdot \mathbf{Z}}$ is defined.

Theorem 6. *Using the same notation of the Definition 18,* $(X \perp\!\!\!\perp Y|\mathbf{Z})_G$ *if and only if L linearly entails that* $\rho_{XY.\mathbf{Z}} = 0$.

By the Theorem 6, when the model is linear with jointly independent errors, partial correlation marks d-separability. In other words, the partial correlations identified by applying d-separation criterion in acyclic or cyclic graphs are guaranteed to vanish [116–118]. Thus, in this case, we can use a statistical test for zero partial correlation as d-separation oracle. Actually, under these assumptions, tests for any statistic that vanishes when partial correlations vanish would suffice [118].

In DAGs, it is possible to test a small number of partial correlations that constitute a basis for the entire set. A possible basis is the one which reflects the local directed Markov property of DAGs, i.e., the set of zero partial correlations between each variable and its predecessors (non-parental non-descendant variables) given its parents [87]. The cardinality of the basis is equal to the number of missing edges in the graph. Thus, the sparser the graph, more tests are required to reconstruct the structure. The PC-algorithm [118] is a fundamental algorithm to recover DAG structures. It runs in the worst case in exponential time with respect to the number of vertices, but if the true underlying DAG is sparse this reduces to a polynomial runtime.

In linear cyclic models, it is also possible to discover features of the graph performing statistical tests of zero partial correlation in a subset of the set of all d-separation relations. The CCD algorithm [97] is a discovery algorithm for linear cyclic models that contain no latent variables. It can infer features of sparse directed graphs from a probability distribution in polynomial time.

5 Equivalent Models

It may happen that two or more causal models or structures share the same conditional independence relations. For instance, a chain ($A \rightarrow B \rightarrow C$), a reverse chain ($A \leftarrow B \leftarrow C$), and a fork ($A \leftarrow B \rightarrow C$) share the following set of conditional and unconditional independence relations:

$$I = \{A \perp\!\!\!\perp C \mid B;\ A \not\!\perp\!\!\!\perp B;\ B \not\!\perp\!\!\!\perp C;\ A \not\!\perp\!\!\!\perp C\}.$$

In this case, the same probability distribution satisfies the global directed Markov and the causal faithfulness conditions for all these graphs. We say that these three models are members of the same equivalence class. In this case, it is not possible to distinguish one from another without any other information or assumption. In other words, they are indistinguishable by observational data alone.

These concepts are formalized in Definitions 19, 20, and 21:

Definition 19. Let S_1 and S_2 be two different SEMs.

We say that S_1 and S_2 are **observationally equivalent** if every probability distribution that is generated by one of the models can also be generated by the other.

Definition 20. Let G_1 and G_2 be two directed cyclic or acyclic graphs.

We say that G_1 and G_2 are **Markov equivalent** or **faithful indistinguishable** if any probability distribution P which satisfies the global directed Markov and faithful conditions with respect to G_1 also satisfies these conditions with respect to G_2, and vice-versa.

Since the global directed Markov and causal faithfulness conditions only places conditional independence constraints on distributions, the following equivalent definition can be established:

Definition 21. Let G_1 and G_2 be two directed cyclic or acyclic graphs.

We say that G_1 and G_2 are **Markov equivalent** or **faithful indistinguishable** if the same d-separation relations hold in both graphs, or, equivalently, if they both linearly entail the same set of conditional independencies.

The Theorem 7 is an important result that holds only for DAGs:

Theorem 7. *Let G_1 and G_2 be two DAGs.*

The two DAGs are Markov equivalent if and only if they have the same skeleton (the undirected version of the graph) and the same unshielded colliders.

We can verify that the two first graphs in Fig. 4 are equivalent using the Theorem 7. Both graphs have acyclic structures, the same skeleton and the same set of unshielded colliders (empty set in this case). Note that A and B are connected in both graphs. Thus, neither the collider C of the first graph nor the collider B of the second graph is an unshielded collider. The third graph in Fig. 4 shows that the acyclicity hypothesis is important for the Theorem 7. Even though the third graph does not have the same skeleton and the same set of unshielded colliders than the first two graphs (B is an unshielded collider in the third graph), these three graphs are Markov equivalent because they hold no d-separation relations [95].

A more complex theory has been developed by Thomas Richardson and Peter Spirtes [95–97, 117] to completely characterize cyclic Markov equivalence classes. We will not discuss these results in this chapter. The more interested reader may refer to [96, 97].

Logsdon et al. [74] demonstrate some results that characterize the set of perturbations (e.g., driving QTLs or QTNs) that minimizes the equivalence classes. Moreover, the authors demonstrate an important theorem, namely “Recovery Theorem,” describing how the set of equivalent DCGs can be recovered from the corresponding moralized graph. As mentioned by Logsdon et al. [74], their results can also be proven by using Thomas Richardson’s work [94].

As a result of the Recovery Theorem, it is possible to guarantee identifiability of both cyclic and acyclic models when each vertex (e.g., phenotype) has at least one unique perturbation associated (e.g., a QTL or QTN) and the genetic architecture is known. That result generalizes the assumption made by Chaibub Neto et al. in the QDG algorithm [24]. By providing a genetic mapping where every phenotype is associated with an unique genetic variant, a directed acyclic or cyclic network can be uniquely recovered [22, 74].

In studies of genetic associations, Mendelian randomization can reduce the size of equivalence classes of phenotypes by using driving genetic variants. This is

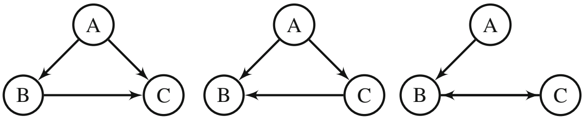


Fig. 4 Two DAGs and one DCG which are Markov equivalent. Every pair of vertices in these graphs cannot be d-separated

because the additional information of the genetic variants causing phenotypes can create new conditional independence relationships among the vertices, getting rid of some Markov-equivalence. Background knowledge can also be used to rule out some equivalent graphs. However, these approaches may not always narrow the set of possibilities to a single graph.

6 Causal Structure Learning

Inferring the causal effects may be a problem of estimation of a SEM if the causal structure is known. For a genotype–phenotype network, the structure is specified indicating the non-zero entries of the matrices **P** and **Q** of the model shown in Eq. (2). The values of these entries can be estimated using an appropriate method of SEM estimation.

When the causal structure is not known a priori, we can use an algorithm to recover or discover it. This problem is known as *causal structure learning* or *causal structure discovery*.

In Sect. 5, we showed that causal structures can be statistically indistinguishable from each other. This means we cannot distinguish equivalent models based on observational data. The use of QTL or QTN genotype information can break statistical equivalence. However, it is still possible that the true causal structure cannot be uniquely recovered. When the exact identifiability is not possible, the algorithms often report suitable summary statistics of all graphs within the equivalence class.

There are two main approaches to infer the network structure that is more compatible with the joint distribution of the data: (1) an SEM-based approach, in which the structure is determined by fitting an SEM with regularization and variable selection techniques for identifying relevant associations; and (2) an approach based on the graphical representation of the causal processes, in which d-separation criterion and greedy strategies are often used to determine an adequate structure network.

6.1 Learning Structural Equation Models

In SEM-based approaches, the network structure learning problem is reduced to the estimation of the model parameters, often using a regularizer that controls the model complexity. The regularizers have a key role in the learning of the network structure because of the variable selection effect. Additional constraints, such as sparsity and smoothness, are incorporated in the likelihood function of an SEM model. Structure learning algorithms based on SEM infer both the causal structure and the parameters of the model.

Most methods for learning genotype–phenotype networks adopt the model shown in Eq. (2) (or a slightly modified version) in which genetic variants are incorporated. The network structure is specified by the non-zero entries of the matrices **P** (causal relations among phenotypes) and **Q** (causal relations from genotypes to phenotypes). Thus, the causal structure learning problem is equivalent to estimate which entries of these matrices are non-zero. Some methodologies require more assumptions than others for estimating these parameters. The most common assumptions are listed in the following:

- The response variables (phenotypes) are assumed to follow a multivariate normal distribution;
- There are no self-loops, so that all diagonal entries of **P** are zero; Self-loops are often used in modeling of time series, implying that a variable depends on its own path. In this case, the model is static, and the state of a phenotype at any instant of time t is assumed to be not dependent on its state at the past time step $t-1$;
- The error term corresponding to the jth individual (column j of the matrix **E**) is modeled as a zero-mean Gaussian vector with covariance $\sigma^2\mathbf{I}$, where **I** denotes the $p \times p$ identity matrix. Thus, the error terms are assumed to be uncorrelated;
- No dominance or epistatic effects are considered in the model. It is assumed that all effects of genetic variants are due to the additive effects of alleles;
- The q QTLs have been predetermined by an existing method, but the magnitude of their effects is unknown.

The most popular regularization methods are those that lead to sparse variable selection. A main reason is that the reduction in the effective number of parameters to be estimated (in sparse models, many parameters are expected to be zero) reduces computational cost, and contributes to the selection of relevant features. Sparsity is often achieved by imposing the $L1$-norm (i.e., the sum of absolute values) on the parameters as a regularization. Examples of sparse regularizers include least absolute shrinkage and selection operator (LASSO) [122] and its extensions, such as adaptive LASSO [136] and Dantzig selector [23]. A limitation of the LASSO is that it tends to select only one variable from a group of highly correlated variables. To overcome this limitation, some regularizers, such as elastic net [137] and its adaptive version [138], combine $L1$-norm with the $L2$-norm (i.e., the sum of squared absolute values) regularization.

In the field of genotype–phenotype network learning, the use of the sparsity constraint is particularly attractive for inferring the structure of GRN. Several studies indicate that biological gene networks, including protein–protein interaction, metabolic, signalling, and transcription-regulatory networks, contain few highly connected vertices (also know as hubs) [1, 10, 58, 69, 93, 124]. Thus, by exploiting the sparsity prior information, it may be possible to both improve computational efficiency and achieve a biologically realistic representation. The Adaptive LASSO (AL)-based algorithm [74], for instance, was designed for determining regulatory relationships underlying observed gene expressions by using the adaptive LASSO procedure for feature selection. Another algorithm intended to infer the structure of GRNs is the SML algorithm, proposed by Cai et al. [22]. The SML algorithm infers sparse SEMs in which an $L1$-norm penalty is incorporated on the entries of the matrix **P** of Eq. (2), inducing a sparsity constraint. Later, Anhui Huang

[54] extended the SML algorithm by incorporating the adaptive elastic net penalty into the SEM likelihood. Other SEM-based structure learning algorithms deserving attention are [32, 71, 73].

6.2 Learning Causal Graphical Models

Causal structure learning algorithms which are based on the graphical representation of the causal processes often use the d-separability concept. Thus, they often assume the global Markov and causal faithfulness conditions. These algorithms search for the structure more compatible with the joint distribution of the data. One of the approaches used to reduce complexity of the algorithm is to constrain the search space by imposing specific properties to the structure of the graphs, such as acyclicity and sparsity.

The most common assumptions often made by structure learning algorithms based on graphical representation are

- **Causal sufficiency**: there are no hidden confounders, that is, all common causes of the underlying causal system have been observed and the error variables are jointly independent. That is a problematic assumption, since it is difficult to be confirmed and, in general, depends on factual knowledge. Within Mendelian randomization framework, the improvement of causal inference by using genetic variants as instrumental variables may compensate for biases introduced by small departures from causal sufficiency. Thus, identification of the true network structure may still be achieved even when causal sufficiency condition is not perfectly fulfilled [40].
- **Causal Markov condition**: the distribution generated by a causal structure (represented by a directed graph) satisfies the global directed Markov condition. It permits inference from probabilistic dependence to causal connection. Note that, for linear SEMs, this assumption holds if the error terms are independent.
- **Causal faithfulness**: all conditional independence relations present in a directed graph G are consequences of the global directed Markov condition applied to the true causal structure G. This is an assumption that any conditional independence relation holding in G is due to the causal structure rather than a particular parameterization of the model. Thus, it permits inference from probabilistic independence to causal separation.

In order to learn causal graphical models, three approaches are often used: constraint-based approaches, score-based approaches, and hybrid approaches, where techniques from constraint-based and score-based approaches are combined. In the following, we will discuss the main ideas used in constraint-based and score-based approaches.

6.2.1 Constraint-Based Approaches

The algorithms in this category are based on significance tests for the null hypothesis that a certain conditional independence statement is true. These individual constraints are used both to decide if a given pair of variables is adjacent or not as well as to orient some edges. In order to read off the implied conditional independencies, d-separation criterion is often used. Thus, the causal sufficiency, causal Markov, and causal faithfulness assumptions must be made in order to safely apply these conditional independence tests.

The most basic causal discovery method is the SGS (Spirtes-Glymour-Scheines) algorithm. The correctness of the SGS algorithm follows from the Theorem 2, which is stated only for DAGs. Thus, acyclicity is assumed.

The SGS algorithm works similarly to the exercise shown in Sect. 4.1, in which the graph of Fig. 2 is reconstructed by using d-separation criterion. The algorithm starts with a complete undirected graph. Then, the skeleton of the graph is inferred: for each pair of variables, it tests whether they are conditionally independent on any set of variables. If so, then the edge connecting the pair can be removed. The reason is that, if the dependence between two vertices can be explained away, then there cannot be a direct causal connection between them. In the next step, the algorithm finds and then orients the edges of the unshielded colliders. The unshielded collider is the only configuration for three vertices and one missing edge that can be uniquely oriented. The orientation of other edges are determined by consistency. That is recursively made until no more edges can be oriented. In this last step, the algorithm checks if any loop is created.

The SGS algorithm is statistically consistent, but it is computationally inefficient. In the edge-removal step, each pair of variables should be conditioned on all possible subsets of the remaining variables. Thus, the number of tests it does grows exponentially in the number of variables.

The PC (Peter and Clark) algorithm is very similar to the SGS algorithm, but it is more efficient, specially for sparse graphs. In the edge-removal step, it tries to condition on as few variables as possible. It only conditions on adjacent variables and the sets are sorted in order of increasing size. The PC algorithm has the same assumptions as the SGS algorithm, and the same consistency properties.

The first step of the PC-algorithm, where an association (undirected) graph is inferred, is called PC-skeleton.

The PC-skeleton is used in the first step of two popular genotype–phenotype structure learning algorithms: the QDG, proposed by Chaibub Neto et al. [24], and the QPSO algorithms, by Huange Wang and Fred van Eeuwijk [126]. Technically, these two algorithms are not constraint-based approaches because they use a score-based approach to orient the edges. Thus, they are considered *hybrid approaches*. Score-based approaches are described in the following section.

6.2.2 Score-Based Approaches

Given a score that indicates how well the network fits the data, score-based algorithms search the space of all possible structures for the network with the highest score.

The search for the global optimal network is an NP-hard problem. Thus, the time required to solve the problem increases very quickly as the number of vertices grows. By the solution of the enumeration problem of labeled directed graphs (i.e., graphs in which each vertex has been assigned a different label, so that all vertices are considered distinct), it is known that there are $2^{n(n-1)}$ different causal structures (directed graphs) with n vertices [103]. In asymptotic notation, there are $2^{O(n^2)}$ structures. Just to give an idea of how this number increases, the number of directed graphs with n labeled vertices, for n varying from 1 to 6, is 1, 4, 64, 4096, 1 048 576 and 1 073 741 824.

Thus, heuristics such as greedy search are used to find a sub-optimal structure. However, local optimal solutions can be far away from the global optimal solutions. This becomes even more critical when the number of sampled configurations is small compared to the number of vertices.

The simplest search algorithm over the structure space is the greedy hill-climbing search. A series of modifications of the local structures are made by adding, removing, or reversing an edge, and the score of the new structure is computed after each modification. The search ends when there are no more modifications that increase the score.

The scores offer model selection criteria for the network structure. There is no consensus on what is the best criterion, since that depends on the objective that one wants to achieve [56]. One of the most used measure is the Bayesian information criterion (BIC) [109], which penalizes complex models and the penalty increases with the sample size. It is an approximation for the posterior predictive distribution with respect to the model parameters. The posterior probability of a structural feature (e.g., the presence of an edge) is the total posterior probability of all models that contain it. By estimating the posterior probability of a feature, we are estimating the strength with which the data indicates the presence of it.

Score-based approaches often assume acyclicity, because every DAG has a topological ordering, that is, an ordering of the vertices as $V_1, \ldots, V_n$ so that for every edge (V_i, V_j) we have $i < j$. In this case, each vertex V_i can have parents only from the set $\{V_1, \ldots, V_{i-1}\}$. That significantly reduces the search space and has implications that can reduce the computational cost of the whole process [121].

Robert W. Robinson derived a recurrence relation to count how many labeled DAGs have n vertices [100]. By applying the recurrence relation for n varying from 1 to 6, we notice that corresponding number of labeled DAGs is 1, 3, 25, 543, 29 281, and 3 781 503. Using asymptotic notation there are $2^{O(n \log n)}$ orderings, as opposed to the $2^{O(n^2)}$ structures. Thus, the space of orders is smaller and more regular than the space of structures.

In Bayesian approaches, it is also assumed acyclicity for estimating the probability of a structural feature over the set of all orderings. That is often performed by using a Markov chain Monte Carlo (MCMC) algorithm. It was noted by empirical studies [110] that different runs of MCMC over the structure space typically lead to very different estimates in the posterior probabilities. That poor convergence to the stationary distribution has not been found running MCMC over ordering space.

Two recent methodologies using MCMC to jointly infer the causal structure of a genotype–phenotype network are the QTLnet algorithm, proposed by Chaibub Neto et al. [25], and the Bayesian framework for inference of the genotype–phenotype map for segregating populations, proposed by Hageman et al. [46].

7 Algorithms for Causal Discovery in Genetic Systems

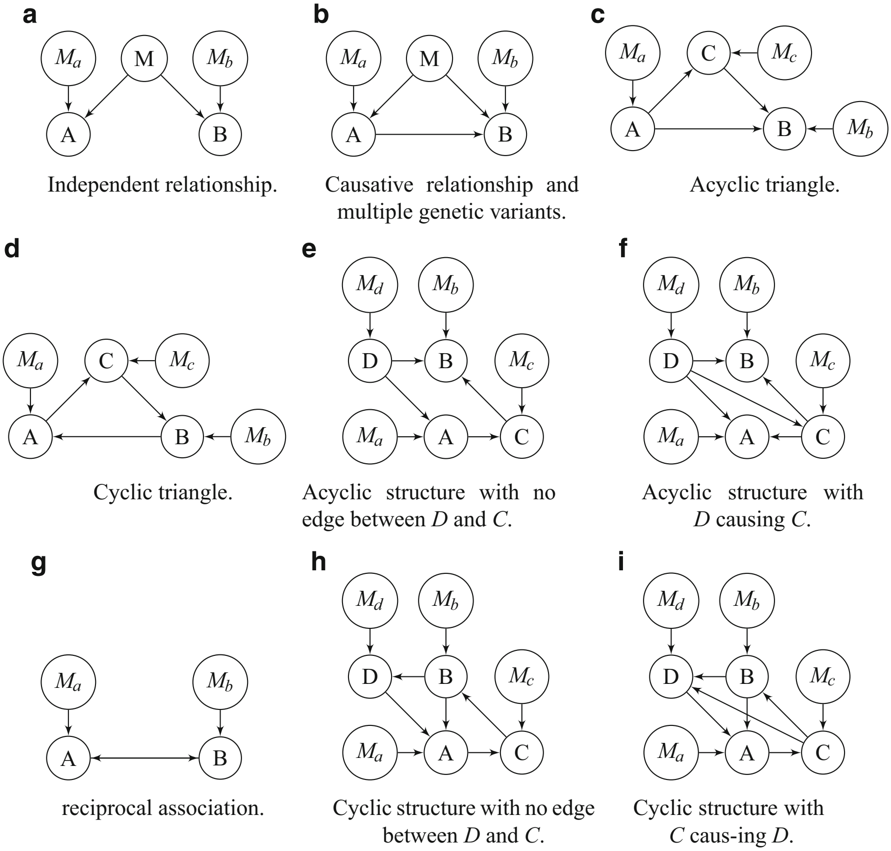
In this section, we will describe some of the most popular algorithms to infer the structure of a genotype–phenotype network, namely QDG [24], QPSO [126], QTLnet [25], and SML [22].

We have tested all proposed features by these algorithms in a simulation study. Some simulated networks are shown in Fig. 5. The letters A, B, C, and D were used to identify phenotypes. In all other networks, there is a distinct genetic variant associated with each phenotype. These genetic variants are identified by the letter M followed by the associated phenotype letter in subscript. We simulated genotype and phenotype data for 500 individuals, independently. QTL genotypes were generated from an F2 intercross using the R/QTL package [18], so that QTLs of each simulated network are unlinked and in linkage equilibrium. The phenotypes were generated according to Eq. (2), with error terms following a normal distribution with zero mean and variance 0.01.

It is worth mentioning that the source codes of the four algorithms are freely available. Details about how to obtain the source code of each algorithms are provided in the following sections.

We want to emphasize that the purpose of our simulation studies is only to check the capabilities of the algorithms. Although it would be very interesting to do a comparison study of the algorithms, it is out of the scope of this chapter. By running simulations under different configurations of phenotype networks, we investigated the advantages and limitations of the algorithms, and we noted that they differ mainly in their ability to discover networks with the following properties: genetic variants with pleiotropic effects, phenotypes associated with multiple genetic variants, acyclic structure, feedback loops, and reciprocal associations. In Sects. 7.1–7.4, we show how these issues are addressed. We conclude this section providing a summary of the main features implemented by each of the algorithms in Table 1.

The QDG and QPSO algorithms are closely related. Both are designed to orient edges into a phenotype association network. In other words, the QDG and QPSO algorithms focus on discovering the causal direction among variables which are known to be statistically associated. However, they can only achieve this goal if genetic variants robustly associated with the phenotypes are previously selected.



(a) Independent relationship.

(b) Causative relationship and multiple genetic variants.

(c) Acyclic triangle.

(d) Cyclic triangle.

(e) Acyclic structure with no edge between *D* and *C*.

(f) Acyclic structure with *D* causing *C*.

(g) reciprocal association.

(h) Cyclic structure with no edge between *D* and *C*.

(i) Cyclic structure with *C* caus-ing *D*.

Fig. 5 Causal networks, in which *A*, *B*, *C*, and *D* are phenotypic vertices. Each is associated with a distinct genetic variant, M_a, M_b, M_c, and M_d, respectively. (**a**) Independent relationship. (**b**) Causative relationship and multiple genetic variants. (**c**) Acyclic triangle. (**d**) Cyclic triangle. (**e**) Acyclic structure with no edge between *D* and *C*. (**f**) Acyclic structure with *D* causing *C*. (**g**) Reciprocal association. (**h**) Cyclic structure with no edge between *D* and *C*. (**i**) Cyclic structure with *C* causing *D*

Thus, besides the phenotype association network, it is also necessary to provide a genetic map as input to these algorithms.

The association network required by both QDG and QPSO algorithms is an undirected graph among phenotypes constructed in such a way that an edge does not exist between two phenotypes if they are d-separated. The main method to infer an association network is the PC-skeleton algorithm [118], discussed in Sect. 6.2.1. Another recommended method to infer association networks is the undirected dependency graph (UDG) [24, 29].

There are several genetic mapping approaches in the literature, most of them using microsatellite or SNP markers.

Linkage mapping is the conventional genetic mapping technique, mostly used for QTLs mapping using microsatellite markers in inbred populations of plants and animals (e.g., F2, backcross, and recombinant inbred lines), and for family based QTL mapping in humans and natural populations [2]. In the simple interval mapping approach [66], the QTL position is assumed to be in a region flanked by two linked observed markers. The Haley–Knott [47] approach gives a remarkably good approximation of the interval mapping, and it is computationally more efficient. These methods model a single QTL at a time and assume that the phenotypes follow a normal distribution or a mixture of normal distributions [135]. When it is necessary to model multiple QTLs at once, composite interval mapping (CIM) [57, 131] and multiple interval mapping (MIM) [61] can be used. In order to estimate the specific position of a QTN (sometimes within a QTL), fine mapping techniques are performed. Linkage disequilibrium and/or SNP information, as well as imputation in sparse marker panels, are currently used for improving fine mapping [43, 52, 125].

The identification of QTNs that are robustly associated with a particular phenotype is mainly done by genome-wide association (GWA) in the case of complex traits and diseases [3] and increasingly by next-generation sequencing in the case of Mendelian traits [30]. In particular, exome sequencing has been a powerful approach for identifying rare variants that underlie Mendelian disorders in circumstances in which conventional approaches have failed [7]. These studies provide much higher resolution than linkage mapping and often involve studying human population.

The QDG algorithm can discover acyclic and cyclic structures, while the QPSO algorithm can only discover acyclic structures. Reciprocal associations cannot be detected by any of these two algorithms.

Unlike the QDG and QPSO, the QTLnet algorithm [25] jointly infers the phenotype network and the associated genetic architecture. Thus, it is not necessary to provide as input to this algorithm a phenotype association network and a genetic map. It assumes that the structure is acyclic and that the phenotypes are normally distributed.

These three aforementioned methods are intended to infer network structures based on concepts of graphical models. The size of the causal effects is not estimated. If it is also necessary to estimate the magnitude of the causal effects, once the model structure is inferred, it may be done by estimating the parameters of the corresponding SEM.

Some SEM-based structure learning algorithms can infer both the phenotype network structure and the magnitude of the causal effects. An example is the SML [22] algorithm which will also be discussed in this section.

As the QDG and QPSO algorithms, the SML algorithm assumes that the genetic variants (e.g., QTLs or QTNs) directly associated with the phenotypes (and preferably not violating the instrumental variable assumptions), must be known a priori and provided as input to the algorithm. It can discover acyclic and cyclic structures, even when reciprocal associations are present.

The output of these algorithms is the most likely structure network according to the data. Except for the SML method, these network reconstruction algorithms also report a goodness-of-fit statistic indicating how well the model fits the data. In particular, this statistic used is the BIC score. The causal model with the lowest BIC score is determined as the best solution.

The QDG and QTLnet algorithms report the BIC score of each fitted model and the score comparing the possible directions of each edge. Thus, the output contains information not only on the best solution, but also on all fitted models. By comparing information on different solutions, one can decide whether there are Markov equivalent networks (possibly models with similar BIC scores) and spurious edges.

In the next sections, each of these four algorithms will be described in more detail.

7.1 QTL-Directed Dependency Graph Algorithm

The QDG algorithm was proposed by Chaibub Neto et al. [24]. The goal of this algorithm is to infer causal directions into a phenotype association network. That is achieved by using information of genetic variants (e.g., QTLs) associated with each phenotype. The algorithm is designed to discover causal structures that can contain cycles but not reciprocal associations.

Both a phenotype association network and a genetic map must be provided as input to the algorithm.

The phenotype association network can be an UDG [29] or the graph skeleton built in the first step of the PC (Peter-Clark) [118] algorithm. Both options are implemented in the QDG algorithm.

In the genetic mapping (e.g., QTL mapping), each phenotype must be associated with at least one unique genetic variant. However, the algorithm can deal with both multiple and pleiotropic genetic variants.

Since the genetic mapping and the determination of the phenotype association network are independently performed, spurious edges and indirectly associated genetic variants can be removed when both are put together. In a pre-processing step of the QDG algorithm, tests are conducted to remove associations identified due to only a pleiotropic effect. In this case, the algorithm verifies whether an association between two phenotypes vanishes by conditioning to the common genetic variant. If so, the edge connecting them is removed.

The orientation edge step is a score-based procedure, similar to those described in Sect. 6.2.2. We will define the logarithm of odds (LOD) score that is used to compare the two possible orientations of each edge.

Consider A and B a pair of connected phenotypes, each with a non-empty set of associated genetic variants, M_a and M_b, respectively. Let $pa(A)$ and $pa(B)$ represent the set of phenotypes that are parents of A and B, respectively. Also, let f be the predictive density, that is, the sampling model with parameters replaced by the corresponding maximum-likelihood estimates. We use a subscript i to represent the phenotype or genotype values of each individual.

The LOD score corresponding to the direction $A \to B$ is given by:

$$
\begin{aligned}
LOD_{A\to B} &= LOD_A + LOD_{B|A} \qquad (6)\\
&= \log_{10}\left\{\prod_{i=1}^{n}\frac{f(A_i|M_{a_i}, pa_i(A))}{f(A_i)}\right\} + \log_{10}\left\{\prod_{i=1}^{n}\frac{f(B_i|A_i, M_{b_i}, pa_i(B))}{f(B_i)}\right\}.
\end{aligned}
$$

The LOD score corresponding to the direction $B \to A$ is given by the formula (6) changing the roles of A and B.

The LOD score is defined as $LOD = LOD_{A\to B} - LOD_{B\to A}$. If it is positive, then the edge is oriented in favor of the direction $A \to B$. Otherwise, it is oriented in the opposite direction. The genetic variants associated with the phenotypes play an important role in this step. As discussed in Sect. 5, they can break likelihood equivalences by creating new conditional independence relationships.

The edges are oriented in a greedy strategy. For each edge, it is computed the LOD score, and the chosen direction is the one with the higher likelihood. This process follows a randomly chosen ordering of the edges. Whenever the orientation of an edge changes, the graph is updated before moving to the next edge.

Different solutions can be obtained by running the algorithm from a different edge ordering. Thus, it is recommended to rerun the algorithm using different edge orderings to get all possible solutions. If more than one solution is obtained in this process, the graph with the lowest BIC score is selected as the best solution.

The output of the algorithm is a detailed report of all possible solutions. It contains the BIC score of each model and the LOD score of each edge. That information can be useful when deciding between equivalent models and whether a certain edge is spurious. For instance, an LOD score very close to zero means that there is not a strong evidence for any direction. In other words, that edge may be spurious.

In Fig. 5a, phenotypes A and B are correlated only due to pleiotropic effect of the common causal genetic variant M. The QDG algorithm could not remove the edge connecting A and B in our simulations for this network, but its LOD score was very close to zero (−0.04).

In the simulation studies reported by the Chaibub Neto et al. [24], the QDG algorithm could infer correctly feedback loops, but not reciprocal associations. When there is a reciprocal association, Chaibub Neto et al. noted that the direction detected corresponds to the one with highest regression coefficient. These observations are consistent with the results obtained in our simulations, since the QDG algorithm failed in recovering the network shown in Fig. 5g.

In our simulations, the QDG algorithm was able to precisely infer simple acyclic and cyclic networks, including those shown in Fig. 5b, c, d, e, h.

However, the QDG algorithm only performs well if the association network is inferred correctly. For instance, the PC-skeleton algorithm was successful in inferring the skeleton of the graphs in Fig. 5e, h, but it failed in recovering a bit more complex version of these networks, such as those shown in Fig. 5f, i. Consequently, the QDG also failed to recover these structures.

The QDG algorithm is implemented in the R/QTLnet package [82], available at https://cran.r-project.org/web/packages/qtlnet/. For reference, the QTLnet package version we used was 1.3.6, published in March, 2014. We run the tests using R environment, version 3.2.3 [92].

7.2 QTL+Phenotype Supervised Orientation Algorithm

The QPSO algorithm was proposed by Wang and van Eeuwijk [126]. It was designed to infer causal connections between pairs of phenotypes from an association network.

Likewise the QDG algorithm described in Sect. 7.1, a phenotype association network and a genetic map must be determined a priori. The genetic map can contain both multiple and pleiotropic genetic variants causally associated with phenotypes. However, the QDG and the QPSO algorithms differ in the sense that the QPSO algorithm does not assume that every phenotype has at least a unique causal genetic variant.

In order to orient the edges between pairs of phenotypes, two important assumptions are made: phenotypes follow a Gaussian distribution and the structure is locally acyclic. Because of the acyclicity assumption for local networks, the algorithm may not perform well in detecting cyclic structures. Moreover, it cannot detect reciprocal associations.

The edge orientation step is a score-based procedure, similar to those shown in Sect. 6.2.2. In each step of a heuristic search, the algorithm chooses a pair of connected phenotypes and extracts its local network. This local network is denoted by local generalized phenotype network (LGPN) and consists of the two connected phenotypes, their parents (both genetic variants and phenotypes) and other phenotypes connected by undirected edges. It is assumed that the LGPN is a conditional linear Gaussian model, in which discrete variables (QTLs or QTNs) are not allowed to have continuous parents (phenotypes).

This local network is thoroughly investigated by using the log-likelihood score which will be defined in the following.

Let A and B be two connected phenotypes. Denote by $pa(A)$ and $pa(B)$ the set of parent vertices of A and B, respectively, including genetic variants and other phenotypes. Let f be the probability density function with parameters replaced by the corresponding maximum-likelihood estimates. A subscript i is used to indicate values for the ith individual.

The log-likelihood score of the local structure is given by:

$$\sum_{i=1}^{n} \log_{10}(f(A_i|pa_i(A))\, f(B_i|pa_i(B))).$$

Using this score, both genetic variants and phenotypes identified as parent vertices can break Markov equivalence among phenotype networks.

Under the assumption that the local structure is acyclic, Wang and van Eeuwijk [126] showed a result that allows all undirected edges of the LGPN be oriented simultaneously.

According to Theorem 7, if two DAGs have different sets of unshielded colliders, then they are not Markov equivalent. Wang and van Eeuwijk [126] showed that all candidate DAGs derived from an LGPN have a distinct set of unshielded colliders if two conditions are satisfied: (1) the pair of connected phenotypes must have at least one parent vertex, and (2) each phenotype connected to the pair by an undirected edge must be nonadjacent to at least one of the parents of the pair's phenotype to which it is connected. Thus, the problem is identifiable under acyclicity assumption and the two aforementioned conditions.

The phenotype network is inferred in a greedy strategy. For each pair of phenotypes satisfying the two identifiability conditions, it is obtained the log-likelihood score of each possible configuration of the respective local network. If there are k undirected edges in the LGPN, then there are 2^k directed graphs to be tested. The configuration with the highest log-likelihood score is considered as the locally optimal directed graph (LODG). This process can be computationally very expensive, since the number of candidate directed graphs increases exponentially.

To prevent the algorithm from converging to a network that is a locally optimal solution, it is recommended that the edge orientation procedure is repeated several times from different starting points. The BIC score is used as a global evaluation metric to determine the most likely solution among those obtained in multiple runs.

Despite the acyclicity assumption in determining the LODG, it is possible that the algorithm builds a cyclic structure when combining the LODGs. However, since the algorithm was not designed with the purpose of discovering cyclic networks, the correct structure can be recovered only by chance.

The accuracy of the QPSO algorithm depends on the association network provided as input to it in a similar way to the QDG. If the association network is not the true skeleton of the network, then the algorithm will fail in recovering the true causal structure.

In our simulation studies, the QPSO algorithm could not recover the true structure of most networks we simulated. Out of the networks shown in Fig. 5, it correctly recover only the networks shown in Fig. 5b, c.

In addition, since the QPSO algorithm only provides information on the network that is solution of the problem, it is difficult to do further analysis in order to identify equivalent networks and spurious edges.

The QPSO algorithm is implemented in Matlab, and it is available upon request to the authors. All simulations were performed using Matlab 8.1.0.604 (R2013a).

7.3 QTL-Driven Phenotype Network Algorithm

The QTLnet algorithm was proposed by [25] to jointly infer causal relationships between genotypes and phenotypes. Thus, unlike the QDG and QPSO algorithms, it is not necessary to provide a phenotype association network and a genetic map as input to the QTLnet algorithm.

The algorithm infers a genotype–phenotype network by a Bayesian procedure under the assumptions that the phenotype network structure is acyclic and the phenotypes follow a normal distribution.

The phenotypes are modeled by a set of structural equations similar to the model shown in Eq. (2).

Let $\mathbf{Y}_i = (Y_{ti})_{t=1}^{T}$, for each $i = 1, \ldots, n$, be a vector with the measurements of T phenotypes for the ith individual, and ε_{ti} represent the corresponding independent normal error terms. Denote by μ_t the overall mean for the phenotype t. Consider a row vector $\mathbf{X}_{ti}$ with the QTL genotypes and observed values of other covariates associated with the phenotype t for the individual i, and a column vector $\boldsymbol{\theta}_t$ with their linear (additive) effects. Thus, the genetic architecture is defined by the elements of $\boldsymbol{\theta}_t$. The effect of the phenotype k on the phenotype t is represented by the coefficient β_{tk}. The notation $pa(Y_t)$ represents the set of phenotypic parent vertices of Y_t.

Each phenotype t of the ith individual is modeled by the following SEM denoted as homogeneous conditional Gaussian regression (HCGR) model:

$$Y_{ti} = \mu_t + \mathbf{X}_{ti}\boldsymbol{\theta}_t + \sum_{Y_k \in pa(Y_t)} \beta_{tk} Y_{ki} + \varepsilon_{ti}, \quad \varepsilon_{ti} \sim \mathcal{N}(0, \sigma_t^2). \tag{7}$$

Though the parametric family HCGR can accommodate cyclic and acyclic networks, only DAGs can be recovered by using the QTLnet algorithm.

Let $\mathcal{M}$ represent a specified network structure and Γ represent all parameters of the model. Also, let $\mathbf{q}_i = \{\mathbf{q}_{1i}, \ldots, \mathbf{q}_{Ti}\}$ be the QTL map of the ith individual, in which $\mathbf{q}_{ti}$, for $t = 1, \ldots, T$, represents the set of QTLs associated with the phenotype t. The likelihood of the HCGR model is equal to the probability of the observed phenotypes conditional to the QTL genotypes, with respect to the parameters $\mathcal{M}$ and Γ. Under the acyclicity assumption, the factorization property shown in Definition 15 holds. Thus, considering the individual likelihood functions

$$p(Y_{ti}|\mathbf{q}_{ti}, pa(Y_t)) \sim \mathcal{N}(\mu_t + \mathbf{X}_{ti}\boldsymbol{\theta}_t + \sum_{Y_k \in pa(Y_t)} \beta_{tk} Y_{ki}, \sigma_t^2),$$

we can write the likelihood of the HCGR model as a product of the all individual likelihood functions:

$$p(\mathbf{Y}_i|\mathbf{q}_i; \Gamma, \mathcal{M}) = \prod_t p(Y_{ti}|\mathbf{q}_{ti}, pa(Y_t)). \tag{8}$$

Thus, by using the DAG factorization property, the likelihood function can be written as a product of normal distributions, one for each value of the data. It means we can use the maximum-likelihood estimation (MLE) technique to estimate the parameters of the HCGR model, in the same way as the classical linear regression model.

By taking the product of the likelihood function and a prior density, up to a normalizing constant, we have the posterior probability distribution. The QTLnet algorithm estimates this posterior probability by using an MCMC algorithm. Thus, the QTLnet algorithm is similar to the Bayesian score-based approaches presented in Sect. 6.2.2.

Specifically, a modified Metropolis–Hastings algorithm was proposed to integrate the sampling of network structures and QTL mapping. It searches across the model space sampling from the derived posterior distribution. In each step of the search, it is proposed a single modification, such as an edge deletion, addition, or reversion, so that the resulting network does not contain cycles. In addition to this simple approach (described in the paper and initially implemented in the software), the R/QTLnet package now supports a more effective M-H sampler [45], which improves a lot the mixing of the Markov chain.

The algorithm does not consider the network with the highest posterior probability as solution of the problem. Instead, the solution is an average network constructed by putting together all causal relationships such that the posterior probability is maximum or above a predetermined threshold.

An advantage of the Bayesian approach is its ability to incorporate prior information in the analysis. In an extension of the QTLnet algorithm [78], it is possible to specify a prior density using, for instance, biological knowledge or sparsity to produce a more predictive network.

Since genotype and phenotype information are jointly analyzed by the QTLnet algorithm, common genetic variants are no longer hidden confounders, reducing the possibility of inferring networks with spurious edges. For instance, the QTLnet could test for the independence between the vertices A and B, conditioned to the common QTL M, correctly inferring the structure of the network shown in Fig. 5a. Additionally, it precisely inferred all acyclic networks we simulated, including those shown in Fig. 5b, c, e, f. On the other hand, since QTLnet assumes acyclicity, it could not discover any cyclic structure.

The output of the algorithm contains the set of the posterior probabilities for each possible network structure and the averaged probabilities for each edge direction. Thoroughly analyzing this information, it is possible to identify equivalent networks. In addition, averaged probabilities close to 0.5 indicate suspicious directions and should be further investigated.

The QTLnet algorithm uses the R/QTL package [18] for performing a single-QTL genome scan, by using methods such as the Haley–Knott regression. This process may involve prediction or imputation of genotypes, requiring information on the experimental cross design. For this reason, the current implementation of the QTLnet is intended for studies in segregating populations. Thus, the source code must be adapted for studies in natural populations.

As the QDG algorithm, it is possible to use the QTLnet algorithm from the R/QTLnet package [82]. We run our simulations using the last version available of QTLnet package (version 1.3.6, published in March, 2014), and the R environment, version 3.2.3 [92].

7.4 Sparsity-Aware Maximum Likelihood Algorithm

The SML algorithm was proposed by Cai et al. [22]. It is an SEM-based approach to infer sparse SEMs integrating genotypic and phenotypic information.

The network is postulated to obey the SEM shown in Eq. (2). The only assumption placed on the phenotype structure network is that there is no self-loops. Thus, the algorithm can infer both acyclic and cyclic networks.

It is assumed that every phenotype is associated with at least one genetic variant. According to the Recovery Theorem [74] discussed in Sect. 5, that restriction guarantees that the network structure can be uniquely identified for both cyclic and acyclic models.

In the paper describing the SML algorithm [22], Cai et al. comment that genetic maps with both multiple and pleiotropic genetic variants affecting phenotypes are supported. However, we could not verify that feature by our simulation studies. That is not yet implemented in the current version of the algorithm (obtained in January 2016 from the supporting information in the online version of the article—doi: 10.1371/journal.pcbi.1003068.s008). In this version, it is only possible to provide genetic variants affecting one phenotype and every phenotype must be associated with only one distinct genetic variant. Considering the Eq. (2), the one-to-one correspondence is forced by placing constraints on the elements of the matrix **Q**. It is required that all elements on the main diagonal are non-zero and all other elements are equal to zero.

Because of that limitation, the algorithm failed to discover the true structure of the network shown in Fig. 5a. It inferred a reciprocal association between the phenotypes *A* and *B*, with a causal effect, in both direction, of 0.168. For the graph in Fig. 5b, the algorithm estimated a causal effect of 0.23 from *B* to *A* and a causal effect of 1.12 from *A* to *B*. The effects of these spurious connections are relatively low. However, we cannot decide whether they are significantly non-zero because we were not provided a significance test for these coefficients.

All the common assumptions listed in Sect. 6.1 are required to carry out the SML algorithm. That is, the model assumes that the phenotypic variance is due to the additive effects, but not due to dominance or epistatic effects of genetic variants. In addition, it is assumed that phenotypes are normally distributed with independent and normally distributed error terms.

The network structure inference is achieved by estimating all the off-diagonal entries of the **P** matrix (the diagonal has only null entries, since it is assumed that no self-loop are present). The matrix **P** may or may not be a triangular matrix, implying that the phenotype network structure is acyclic or cyclic, respectively.

The genetic architecture is specified in the matrix **Q**, but the additive effects need to be estimated. Since the **Q** matrix has only one genetic variant associated with each phenotype, only the diagonal entries of the **Q** matrix need to be estimated by the SML algorithm.

In order to efficiently estimate the parameters, it is assumed that the **P** matrix is sparse, that is, most of its entries are equal to zero. The l_1-norm is used to control the number of zeros in the network structure matrix **P**.

Under the normality assumption, the l_1-regularized log-likelihood maximization problem is solved by using a non-convex optimization algorithm called *block-coordinate ascent* [13].

Although the Recovery Theorem guarantees the identifiability of the network, the algorithm can converge to a local maximum. Thus, it is recommended to run the algorithm several times from different initial values.

In our simulation studies, it was necessary to use an empirically determined threshold of 0.15 in an attempt to eliminate weak causal effects representing spurious associations. Apart from that, the SML algorithm performed very well under the assumption that there is one, and only one, genetic variant associated with each phenotype. The algorithm precisely recovered all acyclic and cyclic causal structures (with both reciprocal associations and feedback loops), including those illustrated in Fig. 5c–i. We noted, however, that the more reciprocal association, the greater the error of the estimated causal effects.

The output of the algorithm contains the estimated **P** and **Q** matrices and the algorithm does not provide a goodness-of-fit measure.

The Matlab package implementing the SML algorithm is available as supporting information in the online version of the article [22]. We run the simulations using Matlab 8.1.0.604 (R2013a).

7.5 Summary

In Sects. 7.1–7.4, we described the algorithms QDG [24], QPSO [126], QTLnet [25], and SML [22].

These algorithms were designed under different assumptions. Thus, one can be more suitable for particular tasks than others. By running simulations, we investigated whether the algorithms are capable to discover the structure of a genotype–phenotype network which was generated according to the expected assumptions. The characterization of the algorithms was based on their ability to discover networks in the following situations: genetic variants with pleiotropic effects, phenotypes associated with multiple genetic variants, acyclic structure, feedback loops, and reciprocal associations. In addition, we investigated their behaviors taking into account the input parameters and the output information.

In Table 1, we summarize the input, output, assumptions, and features of these four algorithms.

Table 1 Input, output, assumptions, and features of the algorithms QDG, QPSO, QTLnet, and SML

	QDG	QPSO	QTLnet	SML
Input				
A phenotype association network	✓	✓		
A genetic map in which each phenotype				
must be associated with a distinct genetic variant	✓			✓
can be associated with multiple genetic variants	✓	✓		✓*
can share a genetic variant with other phenotypes	✓	✓		✓*
Assumptions				
Phenotypes are normally distributed		✓	✓	✓
The phenotype network is acyclic		✓	✓	
Features				
Discovers acyclic structures	✓	✓	✓	✓
Discovers cyclic structures	✓			✓
Discovers reciprocal associations				✓
Performs multiple genetic (QTL) mapping			✓	
Estimates causal effects				✓
Output				
The most likely structure network	✓	✓	✓	✓
A list of the most likely solutions	✓		✓	
A goodness-of-fit measure of the solution	✓	✓	✓	

The check mark indicates that the corresponding algorithm has the property. (*) indicates that the feature could not be verified by our simulation studies

8 Conclusions

The best approach for inferring causality is to conduct randomized controlled trials. However, very often not all interventions can be tested because of the large number of variables and resource constraints.

Throughout this chapter, we showed some concepts and algorithms that allow us to discover causal associations among variables based on observational data. However, we need to be aware that many assumptions are made in this process. It is important to check the validity of these assumptions since they may be unrealistic from a biological point of view.

In this chapter, we explained in detail the reasons for those assumptions. The causal sufficiency is probably the most difficult assumption to retain, since it is difficult to design an experiment in which all causes involved are properly measured. However, in systems genetics, causal inference is aided by Mendelian randomization. The random assignment of genotypes to individuals (from parents to offspring during meiosis) mimics a randomized controlled trial on genetic level, assuring the causal association from genotypes to phenotypes. Moreover, it is possible to improve inferences on the causal effects by using genetic variants as instrumental variables when they are not associated with confounders of the exposure–outcome association of interest or with the outcome through a path other than through the exposure. Considering this scenario, it is even possible to draw appropriate causal inferences on the network structure when the causal sufficiency assumption is being slightly violated. What we mean is that the causal assumption is deterministic, but the causal modeling contains error terms and involves estimation of the residuals of the model as well as robust goodness-of-fit statistics. Thus, causal associations may still be identified if departures from causal sufficiency do not overly inflate the estimate of the error term variance.

We also discussed that Mendelian randomization is a powerful tool for causal structure learning from genomic data. By adding information on causal relationships from genotypes to phenotypes, some conditional independence relations among the variables are created. Thereby, QTL or QTN information has been used for reducing Markov equivalence classes.

Some concepts and definitions in Pearl's causality theory [87] were first developed for acyclic graphical models. Therefore, there are many results in causal inference under the acyclicity hypothesis. Throughout this chapter, we prioritized the exposition of the generalized theory which is applicable to both cyclic and acyclic cases and has been developed mainly by Thomas Richardson and Peter Spirtes [95–97, 117]. However, some results that facilitate the development of algorithms, such as the factorization property of the joint probability distribution as shown in Definition 15 and the Markov equivalence theorem for DAGs as shown in Theorem 7, are only stated for the acyclic case. Thus, more efforts are still needed to make the generalized theory more accessible for practical applications.

Algorithms for discovering causal phenotype networks are of great interest in genomic studies. The output of these algorithms is a directed graph in which the direction of the edges indicates the flow of information in the causal processes. The investigation of the inferred network structure allows a better understanding of the mechanisms of the underlying biological system (e.g., a gene regulation network) and identification of causes of interest (e.g., genetic determinants and risk factors in diseases).

We described in detail four algorithms for genotype–phenotype network learning, namely (1) QDG, (2) QPSO, (3) QTLnet, and (4) SML. These algorithms are similar in the sense that they leverage genetic variant information to help in determining causal directions among phenotypes. However, they were designed under different assumptions, and therefore some may be more suitable than others for a particular biological application.

The most common assumptions include acyclicity of the network structure and normality of the phenotype distribution. The acyclic structure assumption may be quite restrictive for some applications. In GRNs, for instance, modeling cyclic phenomena is particularly important. Therefore, algorithms that are capable to recover networks containing feedback loops and reciprocal associations are possibly more attractive for modeling biological networks.

Some algorithms require that every phenotype is associated with a distinct genetic variant. Under these conditions, the Recovery Theorem assures that the network structure is uniquely identifiable. One must be particularly careful when selecting genetic variants associated with phenotypes of interest. Pleiotropy and linkage disequilibrium may violate some instrumental variables assumptions within Mendelian randomization framework, and thus misleading conclusions may be more likely to be drawn.

In order to achieve efficiency in high-dimensional data analysis, sparsity has been exploited in some causal structure discovery algorithms. As shown in Sect. 4.5, the PC-algorithm, for instance, may run in polynomial runtime if the true network is acyclic and sparse. There are results in the literature supporting the claim that biological networks, such as transcriptional regulatory networks, are sparse and with few highly connected vertices [48]. However, that is not always the case and the sparsity assumption must be verified in the research field of interest.

We investigated some properties of the QDG, QPSO, QTLnet, and SML algorithms by running some simulations. This study had no intentions of comparing the algorithms, but a comparative study would be very valuable for the field. Both the accuracy of the inferred causal structure and the computational performance of the algorithms (in terms of computational time and memory requirements) should be evaluated. The accuracy and complexity of the algorithms are mainly dependent on the number of data samples used, on the number of vertices (dimension of the data), on the number of edges (degree of sparsity of the network), and also on the degree distribution of the vertices (that is, number of edges per vertex). In regard to the latter, it would be interesting a comparison between Erdös–Renyi [35] and Barabási–Albert [9] random graphs, since it must be harder to orient edges incident on (rather than emanating from) highly connected vertices. Another issues that should be taken into account in a comparative study of algorithms are programming language and high performance computing techniques to save computational time, such as parallel and distributed processing.

In this chapter, we discussed linear SEMs representing genotype–phenotype networks in which the genetic effects are considered fixed effects. Gianola et al. [42] presented an alternative modeling of the causal network among phenotypes. The authors proposed a linear recursive SEM in which the additive genetic effects are random effects following a multivariate normal distribution. Valente et al. [123] proposed an algorithm based on d-separation tests for recovering a DAG (or a class of observationally equivalent acyclic causal structures) which represents a recursive SEM in the context of mixed models applied to quantitative genetics.

We also want to emphasize that the methodologies presented in this chapter are intended to model time invariant systems. However, a better understanding of dynamic biological processes, such as signaling, metabolic, and regulatory activities, can be provided by using a model that takes into account the temporal patterns in the data. Using dynamic networks it is possible to infer causal networks exploiting the temporal aspect of time series data. In this case, cyclic associations are inferred using time delay information [41, 80]. Dynamic Bayesian networks (DBNs) have been widely used for inferring GRNs from time series gene expression data [8, 63, 79, 89].

We conclude with a recommendation for future research in genotype–phenotype network structure inference. Most of the reconstruction network structure algorithms discussed in this chapter assume that phenotypes are normally distributed. There are special correlations, such as tetrachoric, polychoric, biserial, and polyserial correlations, which measure the strength of association between continuous and categorical variables and are particularly robust to deviations from symmetry and kurtosis [33, 127]. Additionally, it was observed that SEMs with ordinal categorical indicators are best estimated using special correlation matrix instead of using Pearson's correlation matrix [14, 127]. Thus, an interesting direction for future work is to investigate SEM-based approaches for reconstructing phenotypes networks using these more robust association measures.

Acknowledgements This work was supported by FAPESP (2013/01715-3 and 2014/09576-5, 2015/01587-0), CNPq (306319/2010-1 and 473063/2013-1), CAPES, and NAP eScience–PRP–USP. We are greatly thankful to Guilherme J. M. Rosa for valuable comments.

References

1. Albert, R. Scale-free networks in cell biology. *Journal of cell science 118*, 21 (2005), 4947–4957.
2. Almasy, L., and Blangero, J. Human QTL linkage mapping. *Genetica 136*, 2 (2009), 333–340.
3. Altshuler, D., Daly, M. J., and Lander, E. S. Genetic mapping in human disease. *Science 322*, 5903 (2008), 881–888.
4. Astle, W., and Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science* (2009), 451–471.
5. Aten, J. E., Fuller, T. F., Lusis, A. J., and Horvath, S. Using genetic markers to orient the edges in quantitative trait networks: the neo software. *BMC Systems Biology 2*, 1 (2008), 34.
6. Baba, K., Shibata, R., and Sibuya, M. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics 46*, 4 (2004), 657–664.
7. Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics 12*, 11 (2011), 745–755.
8. Bar-Joseph, Z., Gitter, A., and Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics 13*, 8 (2012), 552–564.
9. Barabási, A.-L., and Albert, R. Emergence of scaling in random networks. *Science 286*, 5439 (1999), 509–512.
10. Barabási, A.-L., and Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics 5*, 2 (2004), 101–113.
11. Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., and Boone, C. Genetic interaction networks: toward an understanding of heritability. *Annual review of genomics and human genetics 14* (2013), 111–133.
12. Baum, C. F., Schaffer, M. E., Stillman, S., et al. Instrumental variables and GMM: Estimation and testing. *Stata journal 3*, 1 (2003), 1–31.

13. BAZERQUE, J. A. *Leveraging sparsity for genetic and wireless cognitive networks*. PhD thesis, University of Minnesota, 2013.
14. BISTAFFA, B. C. *Incorporação de indicadores categóricos ordinais em modelos de equações estruturais*. PhD thesis, Universidade de São Paulo, 2010.
15. BOLLEN, K. A. *Structural equations with latent variables*. John Wiley & Sons, 1989.
16. BOWDEN, R. J., AND TURKINGTON, D. A. *Instrumental variables*, vol. 8. Cambridge University Press, 1990.
17. BOX, G. E., HUNTER, W. G., HUNTER, J. S., ET AL. *Statistics for experimenters*. John Wiley & Sons, New York, 1978.
18. BROMAN, K. W., WU, H., SEN, Ś., AND CHURCHILL, G. A. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics 19*, 7 (2003), 889–890.
19. BURGESS, S., DANIEL, R. M., BUTTERWORTH, A. S., THOMPSON, S. G., CONSORTIUM, E.-I., ET AL. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International journal of epidemiology* (2014), dyu176.
20. BURGESS, S., AND THOMPSON, S. G. *Mendelian Randomization: Methods for using Genetic Variants in Causal Estimation*. CRC Press, 2015.
21. BURGESS, S., THOMPSON, S. G., ET AL. Avoiding bias from weak instruments in Mendelian randomization studies. *International journal of epidemiology 40*, 3 (2011), 755–764.
22. CAI, X., BAZERQUE, J. A., AND GIANNAKIS, G. B. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput Biol 9*, 5 (2013), e1003068.
23. CANDES, E., AND TAO, T. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* (2007), 2313–2351.
24. CHAIBUB NETO, E., FERRARA, C. T., ATTIE, A. D., AND YANDELL, B. S. Inferring causal phenotype networks from segregating populations. *Genetics 179*, 2 (2008), 1089–1100.
25. CHAIBUB NETO, E., KELLER, M. P., ATTIE, A. D., AND YANDELL, B. S. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The annals of applied statistics 4*, 1 (2010), 320.
26. CHU, T., GLYMOUR, C., SCHEINES, R., AND SPIRTES, P. A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics 19*, 9 (2003), 1147–1152.
27. CORDELL, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics 11*, 20 (2002), 2463–2468.
28. DANKS, D., GLYMOUR, C., AND SPIRTES, P. The computational and experimental complexity of gene perturbations for regulatory network search. In *Proceedings of IJCAI-2003 Workshop on Learning Graphical Models for Computational Genomic* (2003), pp. 22–31.
29. DE LA FUENTE, A., BING, N., HOESCHELE, I., AND MENDES, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics 20*, 18 (2004), 3565–3574.
30. DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., ET AL. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics 43*, 5 (2011), 491–498.
31. DIDELEZ, V., AND SHEEHAN, N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research 16*, 4 (2007), 309–330.
32. DONG, Z., SONG, T., AND YUAN, C. Inference of gene regulatory networks from genetic perturbations with linear regression model. *PloS one 8*, 12 (2013), e83263.
33. DRASGOW, F. Polychoric and polyserial correlations. *Encyclopedia of Statistical Sciences* (1988).
34. DUNCAN, O. D. *Introduction to structural equation models*. Elsevier, 2014.
35. ERDŐS, P., AND RÉNYI, A. On random graphs. *Publicationes Mathematicae Debrecen 6* (1959), 290–297.

36. Evans, D. M., and Davey Smith, G. Mendelian randomization: New applications in the coming age of hypothesis-free causality. *Annual review of genomics and human genetics*, (2015).
37. Friedman, N., Linial, M., Nachman, I., and Pe'er, D. Using Bayesian networks to analyze expression data. *Journal of computational biology 7*, 3–4 (2000), 601–620.
38. Galles, D., and Pearl, J. Axioms of causal relevance. *Artificial Intelligence 97*, 1 (1997), 9–43.
39. Gardner, T. S., Di Bernardo, D., Lorenz, D., and Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science 301*, 5629 (2003), 102–105.
40. Geris, L., and Gomez-Cabrero, D. *Uncertainty in biology: a computational modeling approach*, vol. 17. Springer, 2015.
41. Ghahramani, Z. Learning dynamic Bayesian networks. In *Adaptive processing of sequences and data structures*. Springer, 1998, pp. 168–197.
42. Gianola, D., and Sorensen, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics 167*, 3 (2004), 1407–1424.
43. Goddard, M., et al. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics 155*, 1 (2000), 421–430.
44. Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.
45. Grzegorczyk, M., and Husmeier, D. Improving the structure MCMC sampler for bayesian networks by introducing a new edge reversal move. *Machine Learning 71*, 2–3 (2008), 265–305.
46. Hageman, R. S., Leduc, M. S., Korstanje, R., Paigen, B., and Churchill, G. A. A bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics 187*, 4 (2011), 1163–1170.
47. Haley, C. S., and Knott, S. A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*, 4 (1992), 315–324.
48. Hao, D., Ren, C., and Li, C. Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC systems biology 6*, 1 (2012), 34.
49. Hernán, M. A., Hernández-Díaz, S., Werler, M. M., and Mitchell, A. A. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology 155*, 2 (2002), 176–184.
50. Hocutt, M. Aristotle's four becauses. *Philosophy 49*, 190 (1974), 385–399.
51. Holland, P. W. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series 1988*, 1 (1988), i–50.
52. Horikoshi, M., Pasquali, L., Wiltshire, S., Huyghe, J. R., Mahajan, A., Asimit, J. L., Ferreira, T., Locke, A. E., Robertson, N. R., Wang, X., et al. Transancestral fine-mapping of four type 2 diabetes susceptibility loci highlights potential causal regulatory mechanisms. *Human molecular genetics* (2016), ddw048.
53. Houle, D., Govindaraju, D. R., and Omholt, S. Phenomics: the next challenge. *Nature Reviews Genetics 11*, 12 (2010), 855–866.
54. Huang, A. *Sparse model learning for inferring genotype and phenotype associations*. PhD thesis, University of Miami, 2014.
55. Hume, D., and Beauchamp, T. L. *An enquiry concerning human understanding: A critical edition*, vol. 3. Oxford University Press, 2000.
56. Iacobucci, D. Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology 20*, 1 (2010), 90–98.
57. Jansen, R. C. Interval mapping of multiple quantitative trait loci. *Genetics 135*, 1 (1993), 205–211.
58. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. The large-scale organization of metabolic networks. *Nature 407*, 6804 (2000), 651–654.
59. Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis*, vol. 4. Prentice hall Englewood Cliffs, NJ, 1992.

60. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software 47*, 11 (2012), 1–26.
61. Kao, C.-H., Zeng, Z.-B., and Teasdale, R. D. Multiple interval mapping for quantitative trait loci. *Genetics 152*, 3 (1999), 1203–1216.
62. Katan, M. B. Apolipoprotein e isoforms, serum cholesterol, and cancer. *International journal of epidemiology 33*, 1 (2004), 9–9.
63. Kim, S. Y., Imoto, S., Miyano, S., et al. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Briefings in bioinformatics 4*, 3 (2003), 228.
64. Koller, D., and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
65. Koster, J. T., et al. Markov properties of nonrecursive causal models. *The Annals of Statistics 24*, 5 (1996), 2148–2177.
66. Lander, E. S., and Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics 121*, 1 (1989), 185–199.
67. Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. Independence properties of directed Markov fields. *Networks 20*, 5 (1990), 491–505.
68. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine 27*, 8 (2008), 1133–1163.
69. Leclerc, R. D. Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology 4*, 1 (2008).
70. Lewis, D. *Counterfactuals.* Harvard University Press, 1973.
71. Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., and Churchill, G. A. Structural model analysis of multiple quantitative traits. *PLoS Genet 2*, 7 (2006), e114.
72. Liang, Y., and Mikler, A. R. Big data problems on discovering and analyzing causal relationships in epidemiological data. In *Big Data (Big Data), 2014 IEEE International Conference on* (2014), Washington, DC, pp. 11–18.
73. Liu, B., de La Fuente, A., and Hoeschele, I. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics 178*, 3 (2008), 1763–1776.
74. Logsdon, B. A., and Mezey, J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Comput Biol 6*, 12 (2010), 429–435.
75. Lopes, M. S., Bastiaansen, J. W., Harlizius, B., Knol, E. F., and Bovenhuis, H. A genome-wide association study reveals dominance effects on number of teats in pigs. *PloS one 9*, 8 (2014), e105867.
76. Lusis, A. J., Attie, A. D., and Reue, K. Metabolic syndrome: from epidemiology to systems biology. *Nature Reviews Genetics 9*, 11 (2008), 819–830.
77. Markowetz, F., and Spang, R. Inferring cellular networks–a review. *BMC bioinformatics 8*, Suppl 6 (2007), S5.
78. Moon, J. Y., Chaibub Neto, E., Yandell, B., and Deng, X. Bayesian causal phenotype network incorporating genetic variation and biological knowledge. *Probabilistic Graphical Models in Genetics, Genomics, and Postgenomics* (2014), 165–195.
79. Murphy, K., Mian, S., et al. Modelling gene expression data using dynamic bayesian networks. Tech. rep., Computer Science Division, University of California, Berkeley, CA, 1999.
80. Murphy, K. P. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.
81. Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. *Applied linear statistical models*, vol. 4. Irwin Chicago, 1996.
82. Neto, E. C., and Yandell, M. B. S. Package ‘qtlnet’. *Genetics 179* (2014), 1089–1100.
83. Newman, M. E. The structure and function of complex networks. *SIAM review 45*, 2 (2003), 167–256.

84. NEYMAN, J. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in English. *Statistical Science 5* (1923), 463–472.

85. PALMER, T. M., LAWLOR, D. A., HARBORD, R. M., SHEEHAN, N. A., TOBIAS, J. H., TIMPSON, N. J., SMITH, G. D., AND STERNE, J. A. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical methods in medical research 21*, 3 (2012), 223–242.

86. PEARL, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc, San Francisco, CA, 1988.

87. PEARL, J. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, UK, 2000.

88. PEARL, J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology 21*, 6 (2010), 872–875.

89. PERRIN, B.-E., RALAIVOLA, L., MAZURIE, A., BOTTANI, S., MALLET, J., AND D'ALCHE BUC, F. Gene networks inference using dynamic bayesian networks. *Bioinformatics 19*, suppl 2 (2003), ii138–ii148.

90. PIERCE, B. L., AHSAN, H., AND VANDERWEELE, T. J. Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International journal of epidemiology* (2010), dyq151.

91. POURNARA, I., AND WERNISCH, L. Reconstruction of gene networks using bayesian learning and manipulation experiments. *Bioinformatics 20*, 17 (2004), 2934–2942.

92. R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

93. RAVASZ, E., SOMERA, A. L., MONGRU, D. A., OLTVAI, Z. N., AND BARABÁSI, A.-L. Hierarchical organization of modularity in metabolic networks. *Science 297*, 5586 (2002), 1551–1555.

94. RICHARDSON, T. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence* (1996), Morgan Kaufmann Publishers Inc., pp. 454–461.

95. RICHARDSON, T. A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence* (1996), Morgan Kaufmann Publishers Inc., pp. 462–469.

96. RICHARDSON, T. A characterization of markov equivalence for directed cyclic graphs. *International Journal of Approximate Reasoning 17*, 2 (1997), 107–162.

97. RICHARDSON, T., AND SPIRTES, P. Automated discovery of linear feedback models. Tech. rep., CMU-PHIL-75, Dept. of Philosophy, Carnegie Mellon University, 1996.

98. ROBINS, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling 7*, 9 (1986), 1393–1512.

99. ROBINS, J. M. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications 14*, 9 (1987), 923–945.

100. ROBINSON, R. W. Counting labeled acyclic digraphs. In *New Directions in the Theory of Graphs*. Academic Press New York, 1973.

101. ROCKMAN, M. V. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature 456*, 7223 (2008), 738–744.

102. ROSA, G. J., VALENTE, B. D., DE LOS CAMPOS, G., WU, X.-L., GIANOLA, D., AND SILVA, M. A. Inferring causal phenotype networks using structural equation models. *Genet Sel Evol 43*, 6 (2011).

103. ROSEN, K. H. *Handbook of discrete and combinatorial mathematics*. CRC press, 1999.

104. ROSENBAUM, P. R., AND RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41–55.

105. RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*, 5 (1974), 688.

106. Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*, 5721 (2005), 523–529.
107. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature 461*, 7261 (2009), 218–223.
108. Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics 37*, 7 (2005), 710–717.
109. Schwarz, G., et al. Estimating the dimension of a model. *The annals of statistics 6*, 2 (1978), 461–464.
110. Segal, E., Shapira, M., Regev, A., Pe’er, D., Botstein, D., Koller, D., and Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics 34*, 2 (2003), 166–176.
111. Slatkin, M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics 9*, 6 (2008), 477–485.
112. Smith, G. D., and Ebrahim, S. ‘mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology 32*, 1 (2003), 1–22.
113. Smith, G. D., and Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology 33*, 1 (2004), 30–42.
114. Smith, G. D., and Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics 23*, R1 (2014), R89–R98.
115. Speed, T. P. A note on nearest-neighbour gibbs and markov probabilities. *Sankhyā: The Indian Journal of Statistics, Series A* (1979), 184–197.
116. Spirtes, P. Directed cyclic graphs, conditional independence, and non-recursive linear structural equation models. Tech. rep., CMU-PHIL-35, Dept. of Philosophy, Carnegie Mellon University, 1993.
117. Spirtes, P. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (1995), Morgan Kaufmann Publishers Inc., pp. 491–498.
118. Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*, vol. 81. MIT press, 2000.
119. Staiger, D. O., and Stock, J. H. Instrumental variables regression with weak instruments. *Econometrica: Journal of the Econometric Society 65*, 3 (1997), 557–586.
120. Sun, C., VanRaden, P. M., Cole, J. B., and O’Connell, J. R. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PloS one 9*, 8 (2014), e103934.
121. Teyssier, M., and Koller, D. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429* (2012).
122. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.
123. Valente, B. D., Rosa, G. J., de los Campos, G., Gianola, D., and Silva, M. A. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics 185*, 2 (2010), 633–644.
124. Wagner, A., and Fell, D. A. The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences 268*, 1478 (2001), 1803–1810.
125. Wang, H., Haiman, C. A., Burnett, T., Fortini, B. K., Kolonel, L. N., Henderson, B. E., Signorello, L. B., Blot, W. J., Keku, T. O., Berndt, S. I., et al. Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans. *Human molecular genetics 22*, 24 (2013), 5048–5055.
126. Wang, H., and van Eeuwijk, F. A. A new method to infer causal phenotype networks using qtl and phenotypic information. *PloS one 9*, 8 (2014), e103997.

127. West, S. G., Finch, J. F., and Curran, P. J. Structural equation models with nonnormal variables. *Structural equation modeling: Concepts, issues, and applications* (1995), 56–75.
128. Wright, S. Correlation and causation. *Journal of agricultural research 20*, 7 (1921), 557–585.
129. Wright, S. *Appendix to The Tariff on Animal and Vegetable Oils, by P. G. Wright*. New York: MacMillan, 1928.
130. Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics 20*, 18 (2004), 3594–3603.
131. Zeng, Z.-B. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences 90*, 23 (1993), 10972–10976.
132. Zhang, L., Li, H., Li, Z., and Wang, J. Interactions between markers can be caused by the dominance effect of quantitative trait loci. *Genetics 180*, 2 (2008), 1177–1190.
133. Zhao, W., Serpedin, E., and Dougherty, E. R. Recovering genetic regulatory networks from chromatin immunoprecipitation and steady-state microarray data. *EURASIP Journal on Bioinformatics and Systems Biology 2008*, 1 (2008), 248747.
134. Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics 40*, 7 (2008), 854–861.
135. Zhu, X., Zhang, S., Zhao, H., and Cooper, R. S. Association mapping, using a mixture model for complex traits. *Genetic epidemiology 23*, 2 (2002), 181–196.
136. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*, 476 (2006), 1418–1429.
137. Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 2 (2005), 301–320.
138. Zou, H., and Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics 37*, 4 (2009), 1733.