GSE Genetics Selection Evolution

REVIEW Open Access

Inferring causal phenotype networks using structural equation models

Guilherme JM Rosa[1,2*], Bruno D Valente[1,3], Gustavo de los Campos[4], Xiao-Lin Wu[1,5], Daniel Gianola[1,2,5], Martinho A Silva[3]

Abstract

Phenotypic traits may exert causal effects between them. For example, on the one hand, high yield in dairy cows may increase the liability to certain diseases and, on the other hand, the incidence of a disease may affect yield negatively. Likewise, the transcriptome may be a function of the reproductive status in mammals and the latter may depend on other physiological variables. Knowledge of phenotype networks describing such interrelationships can be used to predict the behavior of complex systems, e.g. biological pathways underlying complex traits such as diseases, growth and reproduction. Structural Equation Models (SEM) can be used to study recursive and simultaneous relationships among phenotypes in multivariate systems such as genetical genomics, system biology, and multiple trait models in quantitative genetics. Hence, SEM can produce an interpretation of relationships among traits which differs from that obtained with traditional multiple trait models, in which all relationships are represented by symmetric linear associations among random variables, such as covariances and correlations. In this review, we discuss the application of SEM and related techniques for the study of multiple phenotypes. Two basic scenarios are considered, one pertaining to genetical genomics studies, in which QTL or molecular marker information is used to facilitate causal inference, and another related to quantitative genetic analysis in livestock, in which only phenotypic and pedigree information is available. Advantages and limitations of SEM compared to traditional approaches commonly used for the analysis of multiple traits, as well as some indication of future research in this area are presented in a concluding section.

Background

In animal breeding and quantitative genetics, relationships among phenotypic traits are traditionally studied via probabilistic relationships between them, using standard Multiple Trait Models (MTM) - see, for example, [1,2]. Although such models can be used satisfactorily to infer how probable events are, they are not stable enough to predict how probabilities would change as a result of external interventions [3,4]. In biological systems, phenotypic traits may exert causal effects between them. For example, on the one hand, high yield in dairy cows may increase the liability to certain diseases and, on the other hand, the incidence of a disease may affect yield negatively. Likewise, the transcriptome may be a function of the reproductive status in mammals and the latter may depend on other physiological variables. Such phenotypic relationships can be studied using statistical models that account for recursiveness and feedback between traits.

Information regarding phenotype networks describing such interrelationships can be used to predict the behavior of complex systems, e.g. biological pathways underlying complex traits such as diseases, growth and reproduction, and ultimately it can be used to optimize management practices and multi-trait selection strategies in livestock. For instance, a correlation between traits y_1 and y_2 can be due to a direct effect of y_1 on y_2 (or y_2 on y_1) or to extraneous variables that jointly affect y_1 and y_2. Knowledge about the causal structure underlying phenotypic relationships is necessary to predict the effect of interventions (e.g., management practices) applied to trait y_1 or y_2. For example, if trait y_1 affects y_2, and y_2 has no effect on y_1, an intervention on y_1 will cause changes on y_2, but the reverse would not hold true.

* Correspondence: grosa@wisc.edu
[1]Department of Animal Sciences, University of Wisconsin - Madison, Madison, WI 53706, USA
Full list of author information is available at the end of the article

Similar situations can be considered from a genetic improvement standpoint. Conventionally, genetic correlation is defined as the proportion of variance that two traits share due to genetic causes, and it indicates how much of the genetic influence on two traits is common to both, e.g., due to pleiotropism. However, different scenarios can cause a pleiotropic effect of a specific gene (g) on two traits (y_1 and y_2), as illustrated in Figure 1: (a) the expression of the gene changes trait y_1, and the phenotypic change on trait y_1 affects trait y_2; (b) the expression of the gene acts on trait y_2, and the phenotypic changes on trait y_2 modify trait y_1; or (c) the expression of the gene changes both traits directly, which may or may not have a phenotypic causal effect between them. Knowledge about these different sources of genetic correlation between traits could be used to further improve selection decisions and increase the genetic progress of breeding programs.

As an alternative to the traditional MTM used in animal breeding and genetics, Structural Equation Models (SEM; [5,6]) can be applied to study recursive and simultaneous relationships among phenotypes in multivariate systems. Therefore, SEM can produce an interpretation of relationships among traits which differs from that obtained with standard MTM, where all relationships are represented by symmetric linear associations among random variables, i.e., as measured by covariances and correlations. Unlike MTM, in SEM one trait can be treated as a predictor of another trait, providing a functional (causal) link between them.

In the last few years, genetics has been used as a means to infer phenotype networks, including causal relationships among them [7], and SEM or related methodologies have been employed for such tasks (e.g., [8-12]). These applications of SEM to reconstruct phenotype networks considered genetical genomics studies with model species, using quantitative trait loci (QTL), molecular marker, and or DNA sequence information to facilitate causal inference. However, even with livestock, in which genetical genomics studies are not common due to its cost, and reliable information regarding QTL or even sequence information may not be available, SEM have also been satisfactorily used to study phenotypic networks. SEM within a quantitative genetics mixed models context have been described by [13]. Many authors have used such an approach (e.g., [14,15]), but typically the causal structures are pre-selected using some sort of prior knowledge. More recently, Valente et al. [16] have proposed a methodology that allows searching for recursive causal structures in the context of mixed models for the genetic analysis of multiple traits, showing that under certain conditions it may be possible to infer phenotype networks and causal effects even without QTL or marker information. In this paper, we briefly review SEM and present some of their applications for phenotype network reconstruction in genetical genomics studies, in which both phenotypic and molecular information is available, as well as in the context of classical quantitative genetic analysis of multiple phenotypic traits, using pedigree information.

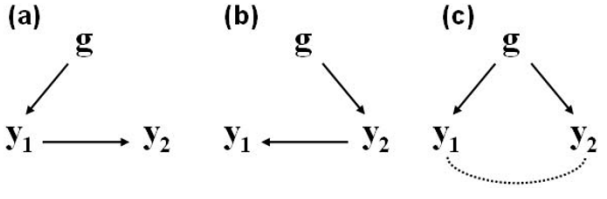

Figure 1 Some possible gene-phenotype networks involving a single gene (g) and two phenotypic traits (y_1 and y_2). Standard multi-trait statistical models could potentially detect a correlation between the two phenotypic traits and a pleiotropic effect of gene g; however, only gene-phenotype network and causal models would be able to distinguish the paths connecting them.

1. Structural equation models

Structural Equation Models [3,4] provide a general statistical modeling technique to estimate and test functional relationships among traits, which are often not revealed by standard linear models. When fitting a SEM to a set of variables, it is necessary to define a priori, for each variable, the subset of the remaining variables that have a (direct) causal effect on it. This information is called 'causal structure', and can be represented as a directed graph in which variables (measured or unmeasured) constitute nodes and causal relationships are represented as directed edges between nodes. For example, consider the graph depicted in Figure 2, in which explanatory variables x and some additional (residual) variables e directly affect variables y, which have also some causal relationships among them.

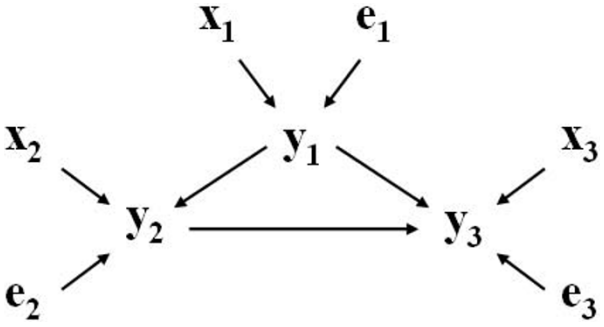

Figure 2 Example of a causal structure, in which y's represent measurements on three phenotypic traits, x's and e's represent known explanatory variables and residual factors affecting y's, respectively.

The graph in Figure 2 can be represented by a set of structural equations, given by:

$$\begin{cases} y_1 = \beta_1 x_1 + e_1 \\ y_2 = \lambda_{21} y_1 + \beta_2 x_2 + e_2 \\ y_3 = \lambda_{31} y_1 + \lambda_{32} y_2 + \beta_3 x_3 + e_3 \end{cases}$$

where β's are model parameters representing the "fixed effects" of the x covariates on y's, and λ's are structural coefficients representing the magnitude of the casual effects among y's. Hence, in matrix notation, a SEM can be represented as $\mathbf{y} = \mathbf{\Lambda y} + \mathbf{X\beta} + \mathbf{e}$, where $\mathbf{\Lambda}$ is a quadratic matrix with zeroes in the diagonal and with structural coefficients λ or zeroes in the off-diagonal, and $\mathbf{y}$, $\mathbf{X}$, $\mathbf{\beta}$ and $\mathbf{e}$ are appropriate vectors or matrices with the observations y's, exogenous variables x's, model parameters β's and residuals e's, respectively. Competing networks representing different causal structures among y's may be compared using some model selection criteria, such as likelihood ratio tests (LRT), Akaike information criterion (AIC; [17]), Bayesian information criterion (BIC; [18]), or Bayesian model selection approaches (see, for example, [19]).

Structural equation models have been intensively used in many fields, such as economics, psychometrics, social statistics, and biological sciences. In genetics, they have been used, for example, to study the relationships between phenotypic traits in humans, especially in the context of twin designs (e.g. [20,21]). More recently, it has been also employed in quantitative genetics mixed model analysis, and on gene-phenotype network reconstruction, as discussed below.

2. QTL information and the randomization of alleles

Thomas and Conti [22] have pointed out that genetically randomized experimental populations that segregate naturally occurring allelic variants can provide a basis for the inference of networks of causal associations among genetic loci, physiological phenotypes, and disease states. In particular, the randomization of alleles that occurs during meiosis provides a setting that is analogous to a randomized experimental design, such that causality can be inferred within the classical Fisherian statistical framework.

In this context, Schadt et al. [7] have proposed a multi-step procedure to infer causal relationships between two phenotypic traits and a common QTL. More specifically, they have tried to disentangle the causal path involving the expression of a particular gene, a cis-acting expression QTL (eQTL), and a complex trait (e.g. a disease trait), to determine if they are related to each other following a causal, reactive or independent model. Such models (denoted here as Models C, R and I, respectively) can be represented as in Figure 1, in which the variables g, y_1 and y_2 denote the cis-acting eQTL, the transcriptional activity of the gene, and the complex trait, respectively. Model C depicted in Figure 1a refers to the simplest causal relationship with respect to y_1, in which allelic variations in g change y_2 by changing the transcriptional activity y_1. Model R (Figure 1b) represents the simplest reactive model with respect to y_1, in which the expression y_1 is modulated by the trait y_2. Lastly, Model I (Figure 1c) represents a situation in which the QTL g controls y_1 and y_2 independently.

Schadt et al. [7] have proposed a likelihood-based causality model selection (LCMS) test that uses conditional correlation measures to determine which relationship among a trio of traits (a transcriptional trait, a complex phenotype, and a common QTL affecting both) is best supported by the data. Likelihoods associated with each of the models (causal, reactive and independent models) have been constructed and maximized with respect to the model parameters, and the AIC criterion has been used to select the model best supported by the data. More specifically, the joint probability distributions of the three models depicted in Figure 1 have been described as:

$$\begin{cases} \mathrm{M_C}: p(g, y_1, y_2) = p(g)p(y_1 \mid g)p(y_2 \mid y_1) \\ \mathrm{M_R}: p(g, y_1, y_2) = p(g)p(y_2 \mid g)p(y_1 \mid y_2) \\ \mathrm{M_I}: p(g, y_1, y_2) = p(g)p(y_1 \mid g)p(y_2 \mid g, y_1) \end{cases}$$

where y_1 and y_2 were assumed normally distributed about each genotypic mean at the common locus g. With those settings, model-specific likelihoods were obtained and standard maximum likelihood estimation methods have been employed.

Schadt et al. [7] have applied their methodology to a mouse genetical genomics study comprised of large-scale genotypic, gene-expression and complex-trait data to identify genes related to obesity, and have been able to identify known and new susceptibility genes for fat mass, and to successfully predict transcriptional response to perturbation in such genes. Their procedure, however, is restricted to simple gene-phenotypes networks, focusing on the identification of genes in the causal-reactive interval considering a trio of nodes comprising a common QTL affecting the expression of a specific gene and a complex trait. Evidently, gene and phenotype networks can be much more complex, as the causal-reactive genes may be also interacting in a broader network through an intricate cascade of genes and phenotypic traits.

More specifically with SEM, Li et al. [8] have presented a methodology to analyze multilocus, multitrait genetic data. Their method extends that of [7], not only by the number of loci and phenotypic traits studied, but

also by different possible causal relationships among them, such that it provides a better characterization of the genetic architecture underlying complex traits. For instance, even if only a single locus and two correlated traits are considered, it allows for alternative recursive effects between phenotypes (Figure 3), outside the causal-reactive interval explored by [7].

The method of [8] comprises a series of five steps. First, single locus genome scans are run for each individual phenotype using a LOD-based test. Next, conditional genome scans are performed using one trait as a covariate in the analysis of another trait. As the authors mention, the choice of which trait(s) to use as covariates can be performed extensively or, alternatively, it may be guided by known biological relationships among the traits. In this setting, traits that are known to be upstream in the causal pathways should be employed as conditioning variables. The comparison between results from unconditioned and conditioned scans can give a first insight into the causal relationships among the phenotypes. For example, in model (8) of Figure 3, g and y_1 are unconditionally independent; however, conditioning on y_2 will result in a nonzero partial correlation between them. By contrast, in model (9), g and y_1 are unconditionally correlated, and by conditioning on y_2 their dependence vanishes. When the QTL g and both traits y_1 and y_2 are causally connected, as in model (1)-(3), the raw and partial correlations between them will all be nonzero, but they will change in magnitude depending on the signs of the path coefficients [8]. A third step on Li et al.'s [8] procedure refers to the construction of an initial path model and its respective SEM representation. In the graphical SEM, each measured trait is represented as a node, including the QTL identified in steps 1 and 2. Edges should be directed from the QTL to the corresponding traits, and edges should be added also from conditioning traits to the responses whenever a significant difference in LOD scores (ΔLOD) is observed. After the path models are constructed, they are assessed in terms of goodness-of-fit by comparing the predicted and observed covariance matrices and by significance tests for individual path coefficients. Finally, an

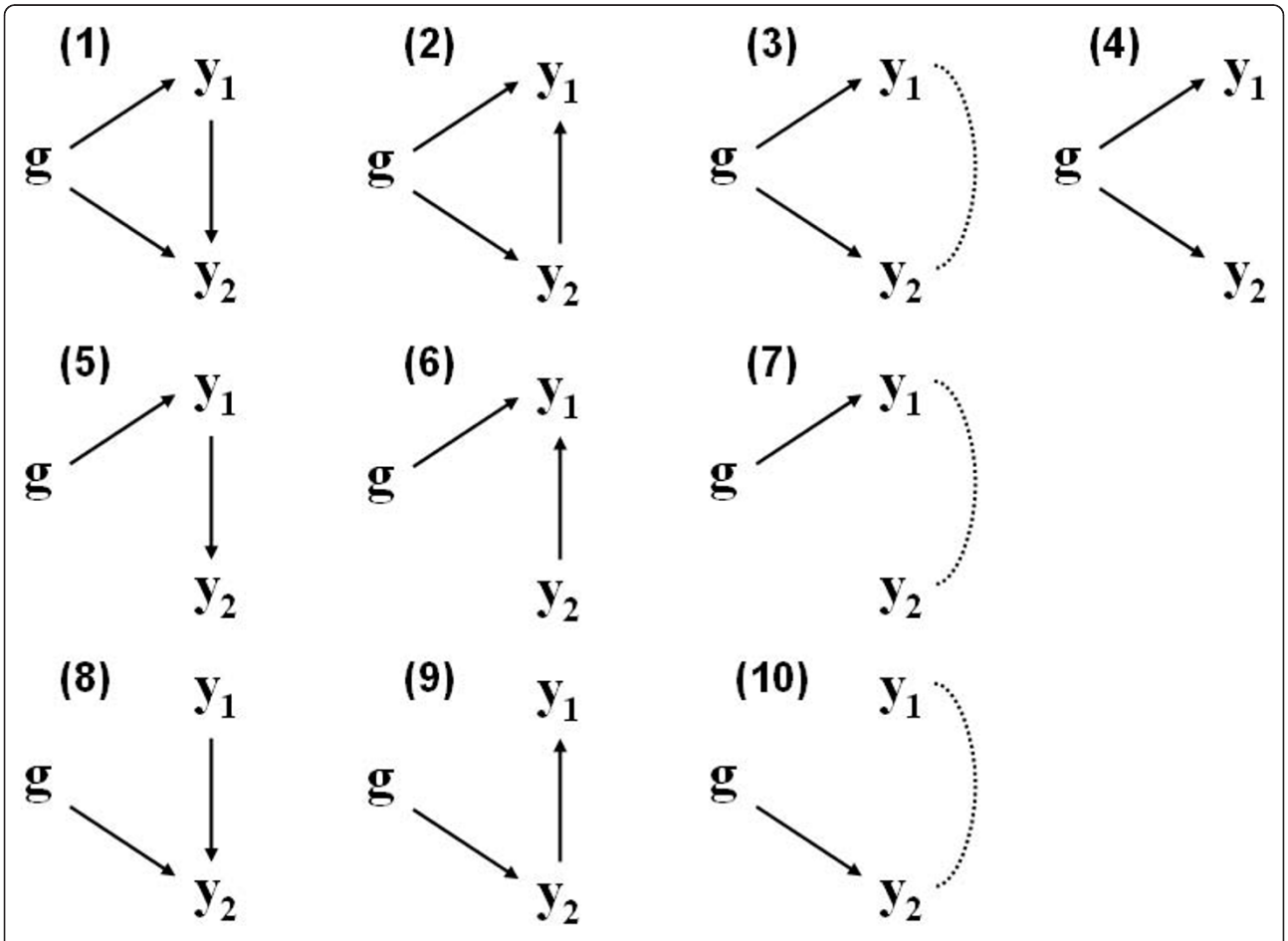

Figure 3 Causal relationships among a QTL (g) and two correlated phenotypes (y_1 and y_2). Arrows indicate the direction of causal effects and dotted lines represent unresolved associations between the two phenotypes (adapted from [8]).

additional step is performed to refine the model, by proposing and assessing alternative models, which are generated by adding or removing edges in the initial model, or by reversing the causal direction of an edge. The authors use a LRT approach to compare such models, but they also suggest that alternative model criteria could be used, such as the AIC or variations thereof, or predictive ability assessed through some cross-validation strategy. Steps 4 and 5 of model refinement and assessment may be also carried iteratively.

Li et al. [8] have carried out the genome scans with tests on every 2 cM using a permutation approach, followed by the SEM component of the analysis. They have applied the methodology proposed to the analysis of body weight and weights of the inguinal, gonadal, peritoneal, and mesenteric fat pads of a SM × NZB intercross population with 260 females and 253 male mice raised on an atherogenic diet, and concluded that SEM provide an insightful descriptive approach to the genetic analysis of multiple traits, allowing the characterization of pleiotropic and heterogeneous genetic effects of multiple loci on multiple traits, as well as the physiological interactions among traits.

Another application of SEM for phenotype causal network inference has been presented by [9], who propose a methodology to search for a set of sparser structures within a putative directed network of causal regulatory relationships among gene expression levels and eQTL in genetical genomics studies. Their method encompasses three steps. First, eQTL mapping techniques are used to identify chromosomal regions modulating the expression of genes. Secondly, regulator-target pairs are identified, such that a directed network can be obtained. Finally, sparser optimal networks are sought within the initial directed network using a SEM approach. Liu et al. [9] have applied their methodology to a genetical genomics data on yeast containing information on expression levels of 4589 genes and genotypes for 2956 markers on 112 haploid offspring originating from a cross between a laboratory and a wild strain. They have detected a number of *cis*- and *trans*-acting eQTL and regulator-target pairs, from which a directed network comprising 28K+ regulator-target pairs was constructed. Based on a partition of this initial network, which comprises 168 genes involved in a cycle genes and all genes connected to the cycle genes by up to three edges and all the eQTL associated with these genes, a SEM analysis has been performed for its sparsification. The preliminary subnetwork had 265 genes, 241 QTL, 832 edges connecting genes, and 640 edges connecting eQTL to genes. The resulting SEM network contained 475 edges connecting genes, and 468 edges connecting eQTL to genes. Some additional analyses have been performed to check for lists of genes with specific biological functions that were enriched on this network, revealing for example that 41.6% of the genes are involved in catalytic activity, and other 18% are involved in hydrolase activity.

Also using QTL information to orient edges connecting phenotypes, Chaibub Neto et al. [11] have proposed a methodology comprised of two main steps. First, an association network is constructed using either an undirected dependency graph (UDG; [4]) or a skeleton derived from the PC algorithm of Spirtes et al. [23]. Second, LOD score tests are used to determine causal direction for every edge that connects a pair of phenotypes, conditional on QTL affecting the phenotypes. They have assessed the performance of their methodology in simulations studies, showing that it can recover network edges and infer their causal direction correctly at a high rate. However, although their method can be applied to human studies and outbred populations, it depends heavily on the availability of reliable information regarding QTL affecting the phenotypic traits of interest. Nonetheless, as discussed by [12], traditional QTL mapping approaches are based on single-trait analyses, in which the network structure among phenotypes is not taken into account. Such single-trait analyses may detect QTL that directly affect each phenotype, as well as QTL with indirect effects, which directly affect phenotypes upstream to the specific phenotype being analyzed. For example, consider the causal graph depicted in Figure 4a, consisting of five phenotypes (y_1-y_5) and three QTL (q_1-q_5). The outputs of single-trait analyses under this scenario are given in Figure 4b. Now, when a multi-trait QTL analysis is performed according to the actual phenotype causal network, detecting indirect-effect QTL is avoided by simply performing mapping analysis of each phenotype conditional on their parents (i.e., upstream phenotypes). For example, in Figure 4a, if a QTL analysis for phenotype y_3 is performed conditionally on trait y_2, only QTL q_3 will be detected because y_3 is conditionally orthogonal to q_1 and q_2, the two QTL with indirect effects (through y_1 and y_2) on y_3.

Hence, traditional QTL mapping approaches that ignore the phenotype network result in poorly estimated genetic architecture of phenotypes, which may hamper correct inferences regarding causal relationships among phenotypes. In view of this drawback of traditional QTL analyses and phenotype network reconstruction methods, Chaibub et al. [12] have suggested a methodology that simultaneously infers a causal phenotype network and its associated genetic architecture. Their approach is based on jointly modeling phenotypes and QTL using homogeneous conditional Gaussian regression models and a graphical criterion for model equivalence. The concept of randomization of alleles during meiosis and the unidirectional relationship from genotype to phenotype are used to infer causal effects of QTL on

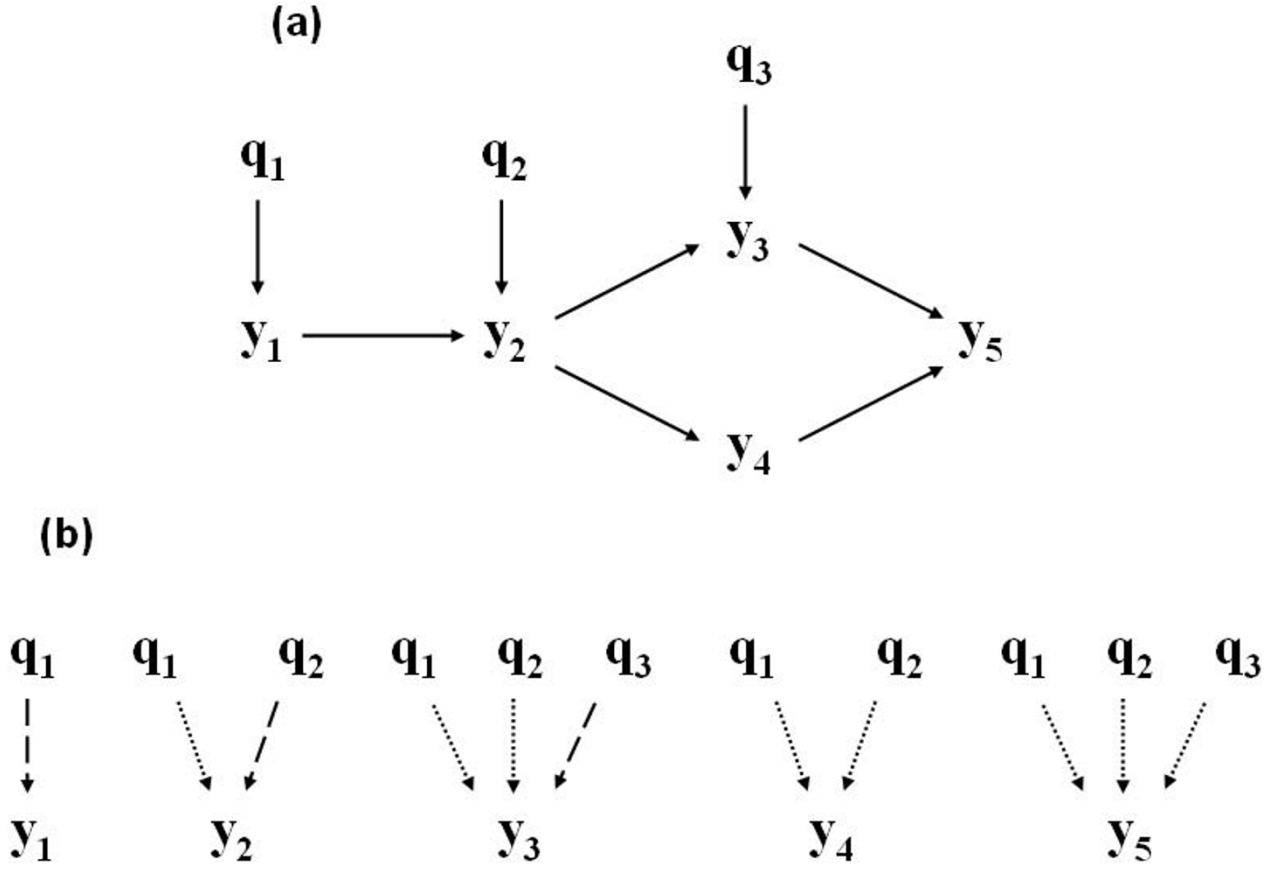

Figure 4 Example of network with five phenotypes and three QTL (Panel a), and the expected output of single-trait QTL analyses for such phenotypes (Panel b). On Panel b, dashed and pointed arrows represent direct and indirect effects of QTLs on phenotypes, respectively.

phenotypes. Subsequently, causal relationships among phenotypes are inferred using the QTL nodes, which might make it possible to distinguish among phenotype networks that would otherwise be distribution equivalent.

3. Inferring causal phenotype networks with no genomic information

All phenotype network reconstruction approaches discussed so far rely on information regarding QTL affecting the phenotypes, or on the availability of genetic marker information for the joint inference regarding phenotype network and genetic architecture. Such QTL are used as parent nodes on putative networks, facilitating inferences on the remainder of the network, either on the construction of preliminary undirected graphs or on the establishment of causal relationships.

However, SEM have also been used to study relationships among phenotypic traits in the context of classical quantitative genetics and animal breeding, even if molecular marker or QTL information is not available. A methodology to insert SEM within a mixed effects model applied to quantitative genetics has been described by [13], and since then applied by many researchers working with different species and phenotypic traits. Some details regarding this methodology and examples of application are described below.

SEM embedded within a quantitative genetics mixed model

A SEM with a specific causal structure and random additive genetic effects can be written as [13,24]:

$$\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i,$$

where $\mathbf{y}_i$ is a $(t \times 1)$ vector of phenotypic records on subject i; $\mathbf{\Lambda}$ is a $(t \times t)$ matrix of structural coefficients describing the chosen causal structure; $\mathbf{X}_i\boldsymbol{\beta}$ represents the effects of exogenous covariates as linear regressions, in which the matrix $\mathbf{X}_i$ contains the covariates and $\boldsymbol{\beta}$ is a vector of 'fixed' regression coefficients; $\mathbf{u}_i$ and $\mathbf{e}_i$ are $(t \times 1)$ vectors of random additive genetic effects and model residuals, respectively, which are both associated with the i^{th} subject. Furthermore, $\mathbf{u}_i$ and $\mathbf{e}_i$ are assumed to be distributed as $\begin{bmatrix}\mathbf{u}_i \\ \mathbf{e}_i\end{bmatrix} \sim N\left\{\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_0\end{bmatrix}\right\}$,

where $\mathbf{G}_0$ and $\mathbf{\Psi}_0$ are the additive genetic and residual covariance matrices, respectively.

The model for n animals can be described as $\mathbf{y} = (\mathbf{\Lambda}\otimes\mathbf{I}_n)\mathbf{y} + \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$, with:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim N\left\{\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0\otimes\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Psi}_0\otimes\mathbf{I}_n\end{bmatrix}\right\},$$

where $\mathbf{y}$, $\mathbf{u}$ and $\mathbf{e}$ are, respectively, vectors of phenotypic records, additive genetic effects and model residuals sorted by trait and subject within trait, and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating effects in β and $\mathbf{u}$ and $\mathbf{y}$. This model may be rewritten as $[\mathbf{I}_{tn}\text{-}(\mathbf{\Lambda}\otimes\mathbf{I}_n)]\ \mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$, so that an equivalent reduced model can be obtained as [13]:

$$\begin{aligned}\mathbf{y} &= \left[\mathbf{I}_{tn} - (\mathbf{\Lambda}\otimes\mathbf{I}_n)\right]^{-1}\mathbf{X\beta}\\ &\quad + \left[\mathbf{I}_{tn} - (\mathbf{\Lambda}\otimes\mathbf{I}_n)\right]^{-1}\mathbf{Zu} + \left[\mathbf{I}_{tn} - (\mathbf{\Lambda}\otimes\mathbf{I}_n)\right]^{-1}\mathbf{e}.\end{aligned}$$

The resulting sampling distribution of $\mathbf{y}$ given the location parameters and the residual covariance matrix is:

$$\begin{aligned}p(\mathbf{y}\mid\mathbf{\Lambda},\mathbf{\beta},\mathbf{u},\mathbf{\Psi}_0) \sim N\Big\{&[\mathbf{I}_{tn} - (\mathbf{\Lambda}\otimes\mathbf{I}_n)]^{-1}(\mathbf{X\beta} + \mathbf{Zu}),\\ &[\mathbf{I}_{tn} - (\mathbf{\Lambda}\otimes\mathbf{I}_n)]^{-1}\mathbf{\Psi}[\mathbf{I}_{tn} - (\mathbf{\Lambda}\otimes\mathbf{I}_n)]^{'-1}\Big\}\end{aligned}$$

where $\mathbf{\Psi} = \mathbf{\Psi}_0\otimes\mathbf{I}_n$.

By reducing the SEM, the location and dispersion parameters are transformed into parameters of a standard MTM [24,25], as indicated below:

$$\begin{aligned}\mathbf{y}_i &= (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{X}_i\mathbf{\beta}\\ &\quad + (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{u}_i + (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{e}_i\\ &= \mathbf{\mu}_i^* + \mathbf{u}_i^* + \mathbf{e}_i^*,\end{aligned}$$

where $\mathbf{\mu}_i^* = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{X}_i\mathbf{\beta}$, $\mathbf{u}_i^* = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{u}_i$, and $\mathbf{e}_i^* = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{e}_i$. In addition, the joint distribution of $\mathbf{u}_i^*$ and $\mathbf{e}_i^*$ is:

$$\begin{bmatrix}\mathbf{u}_i^*\\ \mathbf{e}_i^*\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0^* & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_0^*\end{bmatrix}\right),$$

with $\mathbf{G}_0^* = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{G}_0(\mathbf{I}_t - \mathbf{\Lambda})^{'-1}$ and $\mathbf{R}_0^* = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{\Psi}_0(\mathbf{I}_t - \mathbf{\Lambda})^{'-1}$.

Here, $\mathbf{\mu}_i^*$, $\mathbf{u}_i^*$, $\mathbf{e}_i^*$, $\mathbf{G}_0^*$ and $\mathbf{R}_0^*$ are respectively the vectors of fixed effects, additive genetic effects, model residuals, and the genetic and residual covariance matrices of an MTM. Hence, it is seen that SEM and MTM are equivalent models, i.e.:

$$\begin{aligned}&N\left(\mathbf{\mu}_i^* + \mathbf{u}_i^*, \mathbf{R}_0^*\right)\\ &\quad = N\left((\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{X}_i\mathbf{\beta} + (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{u}_i, (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{\Psi}_0(\mathbf{I}_t - \mathbf{\Lambda})^{'-1}\right)\end{aligned}$$

However, an MTM is just-identified [24], such that changes in parametric values necessarily result in some change in the joint distribution of $\mathbf{y}$. Conversely, SEM carries extra parameters in $\mathbf{\Lambda}$, resulting in an unidentifiable likelihood function. Nevertheless, it is possible to introduce constraints in SEM to achieve parameter identifiability [24]. A constraint which is typically sufficient is coercing the residual covariance matrix $\mathbf{\Psi}_0$ to be diagonal, as in the examples discussed below. After defining the causal structure and achieving parameter identifiability, one may apply standard statistical methodologies (e.g., [26]) to make inferences about model parameters.

SEM models have been used to study simultaneous and recursive relationships between phenotypes in various species and breeds, such as dairy goats [27], Landrace and Yorkshire pigs [25], Holstein (e.g., [15,28-30]) and Norwegian Red (e.g., [14,31,32]) cattle. The phenotypic traits studied span from production (e.g. milk yield in dairy cattle and body weight in pigs) to reproductive (e.g. gestation length and calving ease in dairy cattle, and litter size in pigs) and health-related traits (somatic cell score and mastitis incidence in dairy cattle). In addition, some extensions of the methodology proposed by [13] have been suggested, such as threshold models with structural coefficients functioning at the level of liabilities ([15,28,33]), and models with heterogeneous structural coefficients, such as time- and yield-dependent coefficients (e.g., [15,29,31]). Some details on these applications of SEM in animal breeding and quantitative genetics are provided below.

De los Campos et al. [14,27] have presented the first applications of SEM to study recursive or simultaneous effects between traits within a quantitative genetics mixed effects models. De los Campos et al. [14] have compared four SEM specifications to study relationships between somatic cell score (SCS) and milk yield (MY) in first-lactation Norwegian Red cows using a sire model. Model parameters are estimated using maximum likelihood and the models are compared via BIC. Results indicated a recursive effect from SCS on MY, providing evidence that the negative association between MY and SCS is more likely to be due to an effect of infection (measured indirectly by the SCS) on production than to the opposite direction (i.e., a dilution effect). These results are corroborated by de los Campos et al. [27], who have studied the relationship between MY and SCS in dairy goats. The data consist of repeated measurements in each half of the

udder of the animals. Again, a negative effect of SCS on MY has been observed and the evidence in favor of a dilution effect is not strong. In addition, the authors have found simultaneity of effects between SCS from the left and right halves of the udder.

Also working with MY and SCS data in dairy cattle, Wu et al. [31] have extended the simultaneous and recursive model of [13] to accommodate possible population heterogeneity. A Bayesian analysis via Markov chain Monte Carlo (MCMC) methods has been employed on test-day data of first-lactation Norwegian Red cows. Once more results suggest large negative direct effects from SCS to MY and small reciprocal effects in the opposite direction. In addition, estimated effects between MY and SCS are larger in the first 60 d of lactation than in the subsequent period, and also appear to be yield-dependent, larger in higher producing cows than in lower producing cows.

Another study concerning the relationships between MY and SCS has been conducted by Jamrozik et al. [30] with Canadian Holstein data. The authors have considered multiple-trait random regression animal models with heterogeneous (across lactations and days in milk intervals) simultaneous and recursive links between phenotypes, which are implemented using Bayesian methods via Gibbs sampling. However, in this case, model comparisons based on Bayes factors indicated superiority of simultaneous models over recursive parameterizations.

To infer simultaneous and recursive relationships between binary and Gaussian characters, Wu et al. [33] have proposed a Gaussian-threshold model within the general framework of SEM, and used such a methodology to study the relationships between clinical mastitis (CM) and MY in Norwegian Red cows. The first 180 d of lactation were arbitrarily divided into three periods of 60 days each, in order to investigate how these relationships evolve in the course of lactation. The recursive model shows negative within-period effects from (liability to) CM to MY in all three lactation periods, and positive between-period effects from MY to (liability to) CM in the following period. The results suggest unfavorable effects of production on liability to mastitis, and dynamic relationships between mastitis and test-day MY in the course of lactation.

A related application of Bayesian linear-threshold SEM has been presented by König et al. [28], who have studied the relationships between claw disorders and test-day MY in Holstein cows in eastern Germany. Four different claw disorders (digital dermatitis, sole ulcer, wall disorder, and interdigital hyperplasia) have been scored as binary traits and analyzed separately. Recursive models at the phenotypic level consider a progressive path of lagged relationships describing the influence of test-day milk yield (MY1) on claw disorders and the effect of the disorder on milk production level at the following test day (MY2). As expected, positive structural coefficients have been estimated for the gradient of disease with respect to MY1, and negative coefficients have been obtained for the rate of change in MY2 with respect to the previous claw disorder.

Other applications of Gaussian-threshold SEM with heterogeneous structural coefficients have been presented by de Maturana et al. [15,29] to explore biological relationships between gestation length (GL), calving difficulty (CD), and perinatal mortality (or stillbirth; SB) in dairy cattle. An acyclic model has been assumed, where recursive effects exist from the GL phenotype to the liabilities (latent variables) to CD and SB and from the liability to CD to that of SB considering four periods regarding GL. The results indicate that gestations ~274 days long (three days shorter than the average) lead to the lowest CD and SB levels, and confirm the existence of an intermediate optimum of GL with respect to these traits.

Working with health and fertility traits in dairy cows, Heringstad et al. [32] have employed trivariate recursive Gaussian-threshold models to analyze two fertility traits (calving to first insemination - CFI, and nonreturn rate within 56 d after first insemination - NR56) together with a disease trait, either clinical mastitis (CM), ketosis (KET) or retained placenta. The estimated structural coefficients of the recursive models indicated that presence of KET or retained placenta lengthens CFI, whereas causal effects from CM to fertility are negligible. Recursive effects of disease on NR56, and of CFI on NR56, are all close to zero. The authors conclude that selection against disease is expected to slightly improve fertility (shorter CFI and higher NR56) as a correlated response and vice versa.

Finally, Varona et al. [25] have presented an analysis of litter size and average piglet weight at birth in Landrace and Yorkshire using a standard two-trait mixed model (SMM) and a recursive mixed model (RMM). On the one hand, in Landrace, results in terms of posterior predictive model checking support a model without any form of recursion or, alternatively, a SMM with diagonal covariance matrices for all random effects considered, i.e. additive genetic, permanent and temporary environmental effects. On the other hand, in Yorkshire, the same criterion favors a model with recursion at the level of temporary environmental effects only, or, in terms of the SMM, the association between traits is shown to be exclusively due to an environmental (negative) correlation. In concluding remarks the authors suggest that the choice between a SMM or a RMM should be guided by the availability of software, by ease of interpretation, or by the need to test a particular theory or hypothesis that

may be better formulated under one parameterization and not the other.

Recovering recursive causal structures

To fit a SEM, the matrix $\mathbf{\Lambda}$ of coefficients defining the causal structure must be specified. In all applications of SEM in quantitative genetics so far, the causal structure was assumed known *a priori* (e.g., [15,32]), or just a few putative structures selected using some prior knowledge were compared (e.g., [14,27,25,33]). However, it may be argued that even without information on QTL it may be possible to infer (at least partially) the causal relationships among phenotypic traits using data-driven algorithms that search for a causal structure.

For example, there are algorithms that use the notion of d-separation [3] to explore the space of causal hypotheses so as to arrive to a causal structure (or a class of observationally equivalent causal structures) that is capable of generating the observed pattern of conditional probabilistic independencies between variables. As an example, here we describe how such search can be performed for the model $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{e}_i$.

A recursive causal structure can be represented by a Directed Acyclic Graph (DAG), which is a set of variables (or nodes) connected by directed edges (arrows). Pairs of connected nodes represent direct causal relationships. A path in the causal structure is a sequence of connected variables. Unconditionally, flows of dependence between variables in the extremes of paths may take place, unless there is a collider (variable with arrows converging at it, like c in $a \rightarrow c \leftarrow b$) in the path. Colliders block the flow of dependency in a path, which makes a and b independent in the structure above. Conditioning on a variable that is not in the extremes of the path switches its status regarding the flow of dependence through it, i.e. if the variable is a collider it allows the flow, whereas if it is a non-collider it blocks the flow. Two variables a and b in a DAG are said to be d-separated conditionally on a subset $\mathbf{S}$ of remaining variables if there are no path between a and b such that all its nodes allow the flow of dependence (i.e., no path between a and b in a DAG such that all the colliders or its descendants are in $\mathbf{S}$ and no non-colliders are in $\mathbf{S}$). Under some assumptions, d-separations in the causal structure of a SEM result in conditional independencies in the joint probability distribution of $\mathbf{y}$. This is used to guide the selection of a causal structure or a class of equivalent causal structures (different causal structures that result in joint distributions presenting the same set of conditional independences) that is compatible with the joint distribution of the data [3,23].

Methodologies such as the IC algorithm [3,34] have been developed to explore the connection between recursive causal structures and joint distributions and recover underlying DAG structures (or a class of observationally equivalent structures). Based on a given correlation matrix, this algorithm performs a list of queries about conditional independencies between variables. Assuming that such independencies reflect d-separations in the underlying DAG, the algorithm returns a partially oriented graph as output, which generally results on an important constraint on the initial causal hypothesis space that could be used to fit the SEM. Partially oriented graphs are graphs with directed and undirected edges representing a class of equivalent causal structures.

Considering a set V of random variables, the IC algorithm can be described by the following steps:

1. For each pair of variables a and b in V, search for a set of variables S_{ab} such that a is independent of b given S_{ab}. If a and b are dependent for every possible conditioning set, connect a and b with an undirected edge. This step results in an undirected graph U. Connected variables in U are called adjacent.
2. For each pair of non-adjacent variables a and b with a common adjacent variable c in U (i.e., a - c - b), search for a set S_{ab} that contains c such that a is independent of b given S_{ab}. If this set does not exist, then add arrowheads pointing at c ($a \rightarrow c \leftarrow b$). If this set exists, then continue.
3. In the resulting partially-oriented graph, orient as many undirected edges as possible in such a way that it does not result in new colliders or in cycles.

The goal of the first step of the algorithm is to obtain a graph that specifies pairs of traits that are directly connected by an edge, because variables that are adjacent in the underlying causal structure are not d-separated (hence they are not probabilistically independent) given any possible set of variables. The second step aims to orient edges by searching for unshielded colliders (structures where a collider is directly caused by two non-adjacent variables). Non-adjacent parents of a collider variable are d-separated given at least one set of variables, but not if conditioned to any set of variables that contains the collider. The observational consequence of this is the probabilistic dependence between the non-adjacent parents conditionally on every possible set of variables that contains the common child. The third step performs every further edge orienting that does not result in a new collider or in a cycle. Additional constraining of the output may be achieved by incorporating background knowledge like time precedence or other prior beliefs [4,23].

The decisions about declaring pairs of variables as conditionally dependent or not are based on partial correlations inferred from a sample, which involves some degree of uncertainty. To account for that, decisions

may be made by testing null hypotheses of vanishing partial correlations or, in a Bayesian approach, using highest posterior density (HPD) intervals for the partial correlations.

The IC algorithm was developed based on the connection between causal structure and joint distribution, which requires some assumptions [23]. Maybe the strongest assumption refers to causal sufficiency: it is assumed that every variable that influences two or more variables within the set of studied variables is already within this set. In other words, it is assumed that there are no hidden causes of two or more variables. Considering that residuals in a SEM account for the sum of the effects of the parents of each trait that are not included in the model predictor, the consequence of the causal sufficiency assumption is the absence of sources of residual covariance among traits, i.e. residual covariance matrices must be diagonal [3]. However, as mentioned earlier, this model constraint (i.e., $\mathbf{\Psi}_0$ to be diagonal) is already adopted in recent applications of SEM in animal breeding in order to achieve model identifiability. Therefore, the assumptions of the IC algorithms are not stronger than the assumptions considered in recent application of SEM in quantitative genetics. In those applications, not only covariance matrices of random variables are assumed to be structured (usually diagonal), but the causal structure itself is assumed to be known.

Causal structure search within a quantitative genetics mixed models context

Valente et al. [16] have adopted a SEM setting with a diagonal residual covariance matrix, as in [14,15,32]. Within this construction, a recursive causal structure that is compatible with the joint probability distribution of the data may be searched using the IC algorithm. In the formulation described in the section above, model residuals are regarded as independent, and recursive effects are used to model (interpret) patterns of covariability between observable variables. However, in a mixed SEM (as presented by [13]) with independent residuals, associations between observed traits are explained not only by causal links between them, but also by genetic reasons. Therefore, the unobserved correlated genetic effects considered in this context may confound the causal structure search if one tries to perform it based on the joint distribution of the phenotypes.

Take as an example the causal structure depicted in Figure 5, where there are recursive relationships among phenotypes y_1 through y_5, with uncorrelated residuals (e_1,...,e_5) and correlated additive genetic effects (u_1,...,u_5). The connection between the causal structure among phenotypes and their joint probability distribution does not hold in a model where genetic effects are uncontrolled

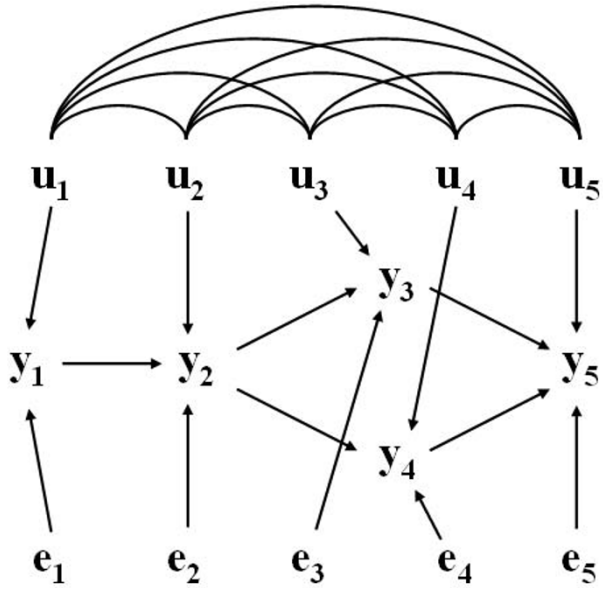

Figure 5 Example of network involving five phenotypic (observable) traits, and their corresponding additive genetic (*u*'s) and residual (*e*'s) effects. The arcs connecting *u*'s represents genetic correlations (adapted from [16]).

hidden variables. For example, given such causal structure y_1 would be expected to be independent of y_3 given y_2, but this may not hold because of the correlation between u_1 and u_3.

Nonetheless, as indicated by [16], genetic relationship information between individuals gives a means of "controlling" for this confounder. Within this context, Valente et al. [16] have proposed an approach to search for acyclic causal structures in which d-separations are reflected as conditional independencies on the distribution of phenotypes after taking into account the additive genetic effects (i.e., the distribution of the phenotypes conditionally on the genetic effects). Given the model settings presented above, i.e., a SEM that accounts for additive genetic effects, the covariance matrix of the phenotypic vector $\mathbf{y}_i$ can be expressed as:

$$Var(\mathbf{y}_i) = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{G}_0(\mathbf{I}_t - \mathbf{\Lambda})^{'-1} + (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{\Psi}_0(\mathbf{I}_t - \mathbf{\Lambda})^{'-1}.$$

Note that $(\mathbf{I}_t - \mathbf{\Lambda})^{-1}\ \mathbf{G}_0\ (\mathbf{I}_t - \mathbf{\Lambda})^{-1}$ and $(\mathbf{I}_t - \mathbf{\Lambda})^{-1}\ \mathbf{\Psi}_0\ (\mathbf{I}_t - \mathbf{\Lambda})^{-1}$ are the covariance matrices of additive genetic effects ($\mathbf{G}_0^*$) and of residuals ($\mathbf{R}_0^*$) obtained from a standard multiple trait mixed model that accounts for covariance between genetic effects and residuals from different traits, but not for causal relationships between phenotypes [13,25]. The covariance matrix of $\mathbf{y}_i$ can be

then rewritten as $Var(\mathbf{y}_i) = \mathbf{G}_0^* + \mathbf{R}_0^*$, and the covariance matrix between traits conditionally on the additive genetic effects can be represented as $Var(\mathbf{y}_i \mid \mathbf{u}_i) = (\mathbf{I}_t - \mathbf{\Lambda})^{-1}\mathbf{\Psi}_0(\mathbf{I}_t - \mathbf{\Lambda})^{'-1} = \mathbf{R}_0^*$. Therefore, estimates of $\mathbf{R}_0^*$ can be used to select a causal structure among phenotypes.

In Valente et al. [16], the (co)variance matrix $\mathbf{R}_0^*$ is inferred using Bayesian MCMC methods, in which samples are drawn from the posterior distribution of $\mathbf{R}_0^*$. These samples are used then to obtain measures of uncertainty about this matrix, while accounting for uncertainty of all other parameters included in the reduced MTM. In summary, the overall statistical approach proposed by [16] consists of three stages:

1. A Bayesian MTM is fitted, and posterior samples of $\mathbf{R}_0^*$ are obtained.
2. The IC algorithm is applied to the posterior samples of $\mathbf{R}_0^*$ to make the statistical decisions required. Specifically, for each query about the statistical independence between variables *a* and *b* given a set of variables *S* and, implicitly, the genetic effects:
   a) Obtain the posterior distribution of residual partial correlation $\rho_{a,b|S}$. These partial correlations are functions of $\mathbf{R}_0^*$. Therefore their posterior distribution can be obtained by computing the correlation at each sample drawn from the posterior distribution of $\mathbf{R}_0^*$.
   b) Compute the 95% HPD interval for the posterior distribution of $\rho_{a,b|S}$.
   c) If the HPD interval contains 0, declare $\rho_{a,b|S}$ as null. Otherwise, declare *a* and *b* as conditionally dependent.
3. Lastly, a SEM using the selected causal structure (or one member within the class of observationally equivalent structures retrieved by the IC algorithm) is fitted, as in [13], such that causal relationships (i.e., recursive effects) can be estimated.

Valente et al. [16] have validated their methodology using simulated data with different causal structures and sample sizes, showing that it can indeed recover the underlying causal structure among phenotypic traits. A first application of such methodology with real data has been presented by Valente et al. [35], who have studied relationships among five traits (birth weight, weight at 35 days of age, age at sexual maturity, average egg weight, and rate of lay) in meat-type quail. The data include 854 females phenotyped for all five traits, and a pedigree file with a total of 10,680 birds. The posterior distributions of the partial correlations obtained are not very sharp, such that different HPD interval contents have been used for the statistical decisions, namely 0.7, 0.75, 0.8, 0.85, 0.9, 0.95 probabilities. Some null partial correlations have been detected; however the structures returned are completely undirected (Figure 6). In this application the edges were oriented based on time sequence information regarding the expression of each trait.

Conclusions

Structural equation models are able to express causality among traits. However, one may fit a SEM with causal structures that do not express the actual causal relationship among traits. The inference of the causal structure is a much harder task than just describing data by a stochastic model. As discussed in this review, using the IC algorithm and related techniques involves accepting specific assumptions, from which the causal sufficiency seems to be the strongest one. In this regard, applying the IC algorithm may be regarded as a causal structure inference only if one is willing to accept the causal assumptions. Otherwise, the application of such algorithms can be viewed simply as a causal structure selection for SEM constructed with diagonal residual covariance matrices.

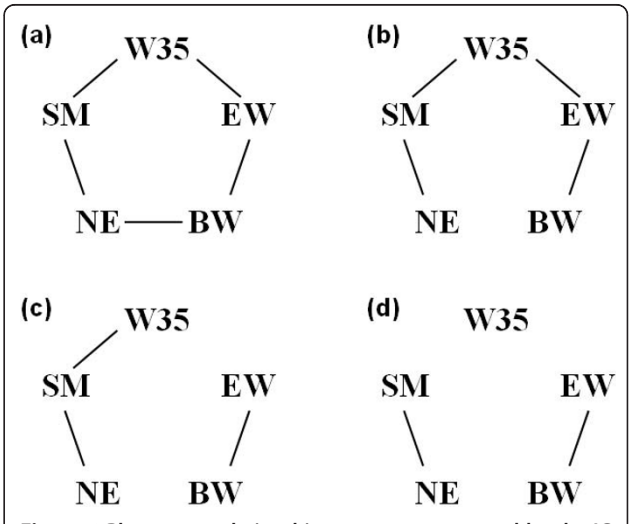

Figure 6 Phenotype relationships structure recovered by the IC algorithm within a mixed model approach as described by [16], **applied to data on meat-type quail**. Edges connecting two traits represent non-null partial correlations, as determined by Highest Posterior Density (HPD) intervals with different contents: Panel (a): 0.7, Panel (b): 0.75, 0.8, 0.85, Panel (c): 0.9, and Panel (d): 0.95; the five traits considered are BW: Birth weight, W35: Weight at 35d, SM: Age at sexual maturity (1st egg), EW: Average egg weight, and NE: Rate of lay (number of eggs).

Nonetheless, the latter applications may still produce interesting and useful results such as the generation of causality hypotheses for further research and investigation. Such hypotheses can then be supported or dismissed by additional data collected from other studies, or they might be tested experimentally through controlled interventions. In genetics, for example, a putative causal mutation could be ultimately tested using gene knockout or knockdown methodologies. However, quite often, randomized experiments are not an alternative due to logistic or ethical constrains, and one is restricted to the analysis of observational studies. In this context, SEM and causal search tools like the IC algorithm are handy. Moreover, in genetics and genomics studies, causal inference is aided by the concept of Mendelian randomization [22], in which allelic variants are randomized to zygotes during meiosis and eventually passed on from parents to offspring, analogously to a randomized experimental design. Applying SEM-related methodologies to QTL analysis and gene mapping with multiple traits not only allows inference regarding causal relationships among phenotypes, but it also enhances detection power and precision of estimates, with the additional advantage of a distinction between direct and indirect genetic effects of QTL on each trait [12].

In addition to DNA polymorphism information and knowledge about genes or QTL that can be used as parent nodes in phenotype network reconstruction, the joint analysis of multilayer large-scale "omics" data such as transcriptome, metabolome and proteome can certainly provide added information and enhance the ability to infer causal phenotype relationships, although it also brings another level of statistical, computational and data mining challenge [36]. Moreover, structural and functional data such as gene sequence, gene localization, transcription binding sites, gene ontology, and metabolic pathway among others can also be used post hoc to verify and test putative gene and phenotype networks [36]. Such data can be used also as a priori information to aid network inference, the same way it has already been used in other "omics" applications such as microarray data [37].

SEM have also been used in the context of quantitative genetics analysis of multiple phenotypic traits when QTL or genomic information is not available [13], allowing a different interpretation of relationships among traits relative to standard multiple trait models traditionally used in animal breeding, where all relationships are represented by symmetric linear associations among traits. As discussed previously, in all applications of SEM in animal breeding so far, the causal structure was assumed known or just a few putative structures were compared. More recently, Valente et al. [16] have proposed a methodology that allows searching for recursive causal structures in the context of mixed models and quantitative genetics. Their approach involves a first step of data adjustment for genetic effects, which otherwise act as confounders of causal effects between phenotypic traits. In Valente et al. [16,35], a classical infinitesimal additive genetic model involving a relationship matrix **A** constructed from pedigree information has been considered for such task. As an alternative, if high density molecular marker data is available (e.g., SNP genotypes), more efficient genetic merit prediction approaches can be employed such as Bayesian regression techniques [38] or kernel methods [39]. This is a topic which deserves further investigation to assess the impact of better estimation of genetic effects on the ability to uncover causal links between phenotypes.

Some other areas related to phenotype network inference that would also warrant additional research refers to the development of (parametric or non-parametric) methods to deal with non-Gaussian traits, as well as search algorithms and software suitable to handle huge number (on the level of thousands) of variables. Lastly, and specifically in the context of animal and plant breeding, extra research is required to study how knowledge regarding causal effects between traits could be explored for the development of more efficient breeding programs and agricultural production enterprises.

In summary, SEM provide a flexible and insightful approach for the genetic analysis of multiple traits, allowing the characterization of pleiotropic and heterogeneous genetic effects of multiple loci on multiple traits, as well as causal relationships among phenotypes, which can be used to predict behavior of complex systems, e.g. biological pathways underlying disease traits. More specifically with livestock, SEM can be used to infer phenotype networks in the genetic analysis of quantitative traits, such that the effect of external interventions can be better predicted. This may foster the development of more efficient breeding programs and optimal decision-making strategies regarding farm management practices.

Acknowledgements

Dr. Guilherme Rosa would like to acknowledge support from the Wisconsin Agricultural Experiment Station and by Vilas Associate Award from the Graduate School of the University of Wisconsin.

Author details

[1]Department of Animal Sciences, University of Wisconsin - Madison, Madison, WI 53706, USA. [2]Department of Biostatistics & Medical Informatics, University of Wisconsin - Madison, Madison, WI 53706, USA. [3]Federal University of Minas Gerais, Belo Horizonte, MG 30123, Brazil. [4]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35201, USA. [5]Department of Dairy Science, University of Wisconsin - Madison, Madison, WI 53706, USA.

Authors' contributions

GJMR and BDV wrote the manuscript; GC, XLW, DG and MAS provided critical insights and helped revising the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 18 October 2010 Accepted: 10 February 2011
Published: 10 February 2011

References

1. Henderson CR, Quaas RL: **Multiple trait evaluation using relatives' records.** *J Anim Sci* 1976, **43**:1188-1197.
2. Mrode R: *Linear Models for the Prediction of Animal Breeding Values.* 2 edition. New York, NY: CAB Int; 2005.
3. Pearl J: *Causality: Models, Reasoning and Inference.* 2 edition. Cambridge, UK: Cambridge University Press; 2009.
4. Shipley B: *Cause and Correlation in Biology* Cambridge, UK: Cambridge University Press; 2002.
5. Wright S: **Correlation and causation.** *J Agric Res* 1921, **201**:557-585.
6. Haavelmo T: **The statistical implications of a system of simultaneous equations.** *Econometrica* 1943, **11**:1-12.
7. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nat Genet* 2005, **37**:710-717.
8. Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA: **Structural model analysis of multiple quantitative traits.** *PLoS Genet* 2006, **2**:e114.
9. Liu B, De La Fuente A, Hoeschele I: **Gene network inference via structural equation modeling in genetical genomics experiments.** *Genetics* 2008, **178**:1763-1776.
10. Aten JE, Fuller TF, Lusis AJ, Horvath S: **Using genetic markers to orient the edges in quantitative trait networks: The NEO software.** *BMC Systems Biology* 2008, **2**:34.
11. Chaibub Neto E, Ferrara TC, Attie AD, Yandell BS: **Inferring causal phenotype networks from segregating populations.** *Genetics* 2008, **179**:1089-1100.
12. Chaibub Neto E, Keller MP, Attie AD, Yandell BS: **Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes.** *Ann Appl Stat* 2010, **4**:320-339.
13. Gianola D, Sorensen D: **Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes.** *Genetics* 2004, **167**:1407-1424.
14. de los Campos G, Gianola D, Heringstad B: **A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows.** *J Dairy Sci* 2006, **89**:4445-4455.
15. de Maturana EL, Wu X-L, Gianola D, Weigel KA, Rosa GJM: **Exploring biological relationships between calving traits in primiparous cattle with a Bayesian recursive model.** *Genetics* 2009, **181**:277-287.
16. Valente BD, Rosa GJM, de los Campos G, Gianola D, Silva MA: **Searching for recursive causal structures in multivariate quantitative genetics mixed models.** *Genetics* 2010, **185**:633-644.
17. Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *2nd International Symposium on Information Theory.* Edited by: Petrov BN, Csaki F. Publishing House of the Hungarian Academy of Sciences, Budapest; 1973:267-291.
18. Schwarz G: **Estimating the dimension of a model.** *Ann Stat* 1978, **6**:461-464.
19. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis.* 2 edition. Boca Raton, Florida: Chapman & Hall/CRC; 2004.
20. Duffy DL, Martin NG: **Inferring the direction of causation in cross-sectional twin data: Theoretical and empirical considerations.** *Genet Epidemiol* 1994, **11**:483-502.
21. Posthuma D, de Geus EJC, Neale MC, Hlshoff Pol HE, Baaré WEC, Kahn RS, Boomsma D: **Multivariate genetic analysis of brain structure in an extended twin design.** *Behavior Genet* 2000, **30**:311-319.
22. Thomas DC, Conti DV: **Commentary: The concept of 'Mendelian randomization'.** *Int J Epidemiol* 2004, **33**:21-25.
23. Spirtes P, Glymour C, Scheines R: *Causation, Prediction and Search.* 2 edition. Cambridge, MA: MIT Press; 2000.
24. Wu X-L, Heringstad B, Gianola D: **Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications.** *J Anim Breed Genet* 2010, **127**:3-15.
25. Varona L, Sorensen D, Thompson R: **Analysis of litter size and average litter weight in pigs using recursive model.** *Genetics* 2007, **177**:1791-1799.
26. Sorensen D, Gianola D: *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics* New York: Springer-Verlag; 2002.
27. de los Campos G, Gianola D, Boettcher P, Moroni P: **A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats.** *J Anim Sci* 2006, **84**:2934-2941.
28. König S, Wu X-L, Gianola D, Heringstad B, Simianer H: **Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models.** *J Dairy Sci* 2008, **91**:395-406.
29. de Maturana EL, de los Campos G, Wu X-L, Gianola D, Weigel KA, Rosa GJM: **Modeling relationships between calving traits: a comparison between standard and recursive mixed models.** *Genet Sel Evol* 2010, **42**:1.
30. Jamrozik J, Bohmanova J, Schaeffer LR: **Relationships between milk yield and somatic cell score in Canadian Holsteins from simultaneous and recursive random regression models.** *J Dairy Sci* 2010, **93**:1216-1233.
31. Wu X-L, Heringstad B, Chang YM, de los Campos G, Gianola D: **Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models.** *J Dairy Sci* 2007, **90**:3508-3521.
32. Heringstad B, Wu X-L, Gianola D: **Inferring relationships between health and fertility in Norwegian red cows using recursive models.** *J Dairy Sci* 2009, **92**:1778-1784.
33. Wu X-L, Heringstad B, Gianola D: **Exploration of lagged relationships between mastitis and milk yield in dairy cows using a Bayesian structural equation Gaussian-threshold model.** *Genet Sel Evol* 2008, **40**:333-357.
34. Verma T, Pearl P: **Equivalence and synthesis of causal models.** In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence. Volume 6.* Cambridge, MA; 1990:220-227, Reprinted in Uncertainty in Artificial Intelligence, 6: 255:268, Elsevier, Amsterdam.
35. Valente BD, Rosa GJM, Silva MA, Teixeira RB, Torres RA: **Busca por estruturas causais recursivas acíclicas envolvendo cinco características produtivas e reprodutivas de codornas de corte.** *III Congresso Brasileiro e IV Simpósio Internacional de Coturnicultura* Lavras, MG, Brazil; 2010.
36. Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM: **Defining gene and QTL networks.** *Curr Opin Plant Biol* 2009, **12**:241-246.
37. Rosa GJM, Vazquez AI: **Integrating biological information into the statistical analysis and design of microarray experiments.** *Animal* 2010, **4**:165-172.
38. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, **183**:347-363.
39. de los Campos G, Gianola D, Rosa GJM: **Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation.** *J Anim Sci* 2009, **87**:1883-1887.

doi:10.1186/1297-9686-43-6
Cite this article as: Rosa *et al.*: Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution* 2011 **43**:6.