

# CAUSAL GRAPHICAL MODELS IN SYSTEMS GENETICS: A UNIFIED FRAMEWORK FOR JOINT INFERENCE OF CAUSAL NETWORK AND GENETIC ARCHITECTURE FOR CORRELATED PHENOTYPES<sup>1</sup>

BY ELIAS CHAIBUB NETO, MARK P. KELLER, ALAN D. ATTIE  
AND BRIAN S. YANDELL

*University of Wisconsin–Madison*

Causal inference approaches in systems genetics exploit quantitative trait loci (QTL) genotypes to infer causal relationships among phenotypes. The genetic architecture of each phenotype may be complex, and poorly estimated genetic architectures may compromise the inference of causal relationships among phenotypes. Existing methods assume QTLs are known or inferred without regard to the phenotype network structure. In this paper we develop a QTL-driven phenotype network method (QTLnet) to jointly infer a causal phenotype network and associated genetic architecture for sets of correlated phenotypes. Randomization of alleles during meiosis and the unidirectional influence of genotype on phenotype allow the inference of QTLs causal to phenotypes. Causal relationships among phenotypes can be inferred using these QTL nodes, enabling us to distinguish among phenotype networks that would otherwise be distribution equivalent. We jointly model phenotypes and QTLs using homogeneous conditional Gaussian regression models, and we derive a graphical criterion for distribution equivalence. We validate the QTLnet approach in a simulation study. Finally, we illustrate with simulated data and a real example how QTLnet can be used to infer both direct and indirect effects of QTLs and phenotypes that co-map to a genomic region.

**1. Introduction.** In the past few years it has been recognized that genetics can be used to establish causal relationships among phenotypes organized in networks [Schadt et al. (2005), Kulp and Jagalur (2006), Li et al. (2006), Chen, Emmert-Streib and Storey (2007), Liu, de la Fuente and Hoeschele (2008), Aten et al. (2008), Chaibub Neto et al. (2008)]. These approaches aim to generate a hypothesis about causal relationships among phenotypes involved in biological pathways underlying complex diseases such as diabetes. A key element in these methods is the identification of quantitative trait loci (QTLs) that are causal for each phenotype. The genetic architecture of each phenotype, which consists of the locations

---

Received December 2008; revised July 2009.

<sup>1</sup>Supported by CNPq Brazil (ECN); NIDDK Grants DK66369, DK58037 and DK06639 (ADA, MPK, BSY, ECN); and by NIGMS Grants PA02110 and GM069430-01A2 (BSY).

*Key words and phrases.* Causal graphical models, QTL mapping, joint inference of phenotype network and genetic architecture, systems genetics, homogeneous conditional Gaussian regression models, Markov chain Monte Carlo.

and effects of detectable QTLs, may be complex. Poorly estimated genetic architectures may compromise the inference of causal relationships among phenotypes. Existing methods that estimate QTLs from genome scans that ignore causal phenotypes bias the genetic architecture by incorrectly inferring QTLs that have indirect effects.

In this paper we propose a novel framework for the joint inference of phenotype network structure and genetic architecture (QTLnet). We model phenotypes and QTL genotypes jointly using homogeneous conditional Gaussian regression (HCGR) models [Lauritzen (1996)]. The genetic architecture for each phenotype is inferred conditional on the phenotype network. Because the phenotype network structure is itself unknown, the algorithm iterates between updating the network structure and genetic architecture using a Markov chain Monte Carlo (MCMC) approach. The posterior sample of network structures is summarized by Bayesian model averaging. To the best of our knowledge, no other proposed method explicitly uses an inferred network structure among phenotypes when performing QTL mapping. Tailoring QTL mapping to network structure avoids the false detection of QTLs with indirect effects and improves phenotype network structure inference.

We employ a causal inference framework with components of both randomized experiments and conditional probability. Randomization of alleles during meiosis and the unidirectional influence of genotype on phenotype allow the inference of causal QTLs for phenotypes. Causal relationships among phenotypes can be inferred using these QTL nodes, enabling us to distinguish between networks that would otherwise be distribution equivalent.

We are particularly interested in inferring causal networks relating sets of phenotypes mapping to coincident genomic regions. It is widely asserted that alleged “hot spots” may have a “master regulator” and that most co-mapping is due to indirect effects [Breitling et al. (2008)]. That is, such a hot spot QTL could influence a single phenotype that is upstream of many others in a causal network; ignoring the phenotype network would result in a perceived hot spot. One objective of our QTLnet method is to sort out the direct and indirect effects of QTLs and phenotypes in such situations.

In the next sections we develop in detail a framework for the joint inference of causal network and genetic architecture of correlated phenotypes. The core idea is to learn the structure of mixed Bayesian networks composed of phenotypes (continuous variables) and QTLs (discrete variables). We model the conditional distribution of phenotypes given the QTLs with homogeneous conditional Gaussian regression models described in Section 2. This allows us to justify formal inference of causal direction along the Bayesian networks in Section 3. That is, we can reduce the size of equivalence classes of Bayesian networks of phenotypes by using driving QTL. In Section 4 we show how a conditional LOD score can formally measure conditional dependence among phenotypes and QTLs and how QTL mapping can be embedded in a graphical models framework. Section 5 presents the MCMC approach for QTLnet. Simulation studies in Section 6 validate the QTLnet

approach and provide an explanation for some QTL hot spots. Section 7 uses real data to illustrate how QTLnet can be used to infer direct and indirect effects of QTLs and phenotypes that co-map to a genomic region. The Discussion puts this work in the context of open questions. Proofs of formal results are given in the Supplement [Chaibub Neto et al. (2009)].

**2. HCGR genetic model.** In this section we recast the genetical model for QTL studies as a homogeneous conditional Gaussian regression model that jointly models phenotypes and QTL genotypes. Conditional on the QTL genotypes and covariates, the phenotypes are distributed according to a multivariate normal distribution. The QTLs and covariates enter the HCGR model through the mean in a similar fashion to the seemingly unrelated regression model [Banerjee et al. (2008)]. However, the correlation structure among phenotypes is explicitly modeled according to the directed graph representation of the phenotype network. We derive the genetic model from a system of linear regression equations and show that it corresponds to a homogenous conditional Gaussian regression model.

In QTL studies, the observed data consist of phenotypic trait values,  $\mathbf{y}$ , and marker genotypes,  $\mathbf{m}$ , on  $n$  individuals derived from an inbred line cross. Following Sen and Churchill (2001), we condition on unobserved QTL genotypes,  $\mathbf{q}$ , to partition our model into genetic and recombination components, respectively relating phenotypes to QTLs and QTLs to observed markers across the genome,

$$p(\mathbf{y}, \mathbf{q} \mid \mathbf{m}) = p(\mathbf{y} \mid \mathbf{q}, \mathbf{m})p(\mathbf{q} \mid \mathbf{m}) = p(\mathbf{y} \mid \mathbf{q})p(\mathbf{q} \mid \mathbf{m}),$$

where the second equality follows from conditional independence,  $\mathbf{y} \perp\!\!\!\perp \mathbf{m} \mid \mathbf{q}$ . That is, given the QTL genotypes, the marker genotypes provide no additional information about the phenotypes. Estimation of the recombination model,  $p(\mathbf{q} \mid \mathbf{m})$ , is a well-solved problem [Broman et al. (2003)] and is not addressed in this paper.

Let  $i = 1, \dots, n$  and  $t = 1, \dots, T$  index individuals and phenotype traits, respectively. Let  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)'$  represent all phenotypic trait values,  $\mathbf{y}_i = (y_{1i}, \dots, y_{Ti})'$  represent the measurements of the  $T$  phenotype traits for individual  $i$ , and let  $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \dots, \varepsilon_{Ti})'$  represent the associated independent normal error terms. We assume that individual  $i$  and trait  $t$  have the following phenotype model:

$$(2.1) \quad y_{ti} = \mu_{ti}^* + \sum_{v \in \text{pa}(y_t)} \beta_{tv} y_{vi} + \varepsilon_{ti}, \quad \varepsilon_{ti} \sim N(0, \sigma_t^2),$$

where  $\mu_{ti}^* = \mu_t + \mathbf{X}_{ti}\boldsymbol{\theta}_t$ , where  $\mu_t$  is the overall mean for trait  $t$ ,  $\boldsymbol{\theta}_t$  is a column vector of all genetic effects constituting the genetic architecture of trait  $t$  plus any additional additive or interactive covariates, and  $\mathbf{X}_{ti}$  represents the row vector of genetic effects predictors derived from the QTL genotypes along with any covariates. The notation  $\text{pa}(y_t)$  represents the set of parent phenotype nodes of  $y_t$ , that is, the set of phenotype nodes that directly affect  $y_t$ . Genetic effects may follow Cockerham's genetic model, but need not be restricted to this form [Zeng, Wang and Zou (2005)].

The Jacobian transformation from  $\mathbf{e}_i \rightarrow \mathbf{y}_i$  allows us to represent the joint density of the phenotype traits conditional on the respective genetic architectures as multivariate normal with the following mean vector and covariance matrix.

**RESULT 1.** The conditional joint distribution of the phenotype traits organized according to the set of structural equations defined in (2.1) is  $\mathbf{y}_i \mid \boldsymbol{\mu}_i^*, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 \sim N_T(\boldsymbol{\Omega}^{-1} \boldsymbol{\gamma}_i, \boldsymbol{\Omega}^{-1})$ , where  $\boldsymbol{\mu}_i^* = (\mu_{1i}^*, \dots, \mu_{Ti}^*)'$ ,  $\boldsymbol{\beta} = \{\beta_{tv} : v \in \text{pa}(y_t), t = 1, \dots, T\}$ ,  $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_T^2)'$ ,  $\boldsymbol{\Omega}$  is the concentration matrix with entries given by

$$\omega_{tv} = \begin{cases} \frac{1}{\sigma_t^2} + \sum_s \frac{\beta_{st}^2}{\sigma_s^2} \mathbb{1}_{\{t \rightarrow s\}}, & \text{for } t = v, \\ -\frac{\beta_{vt}}{\sigma_v^2} \mathbb{1}_{\{t \rightarrow v\}} - \frac{\beta_{tv}}{\sigma_t^2} \mathbb{1}_{\{v \rightarrow t\}} + \sum_s \frac{\beta_{sv} \beta_{st}}{\sigma_v^2} \mathbb{1}_{\{v \rightarrow s, t \rightarrow s\}}, & \text{for } t \neq v, \end{cases}$$

$\boldsymbol{\gamma}_i$  is a vector with entries  $\frac{\mu_{ti}^*}{\sigma_t^2} - \sum_{s \neq t} \frac{\beta_{st} \mu_{si}^*}{\sigma_s^2} \mathbb{1}_{\{t \rightarrow s\}}$ , and  $\mathbb{1}_{\{t \rightarrow s\}}$  is the indicator function that trait  $t$  affects trait  $s$ .

**REMARKS.** (1) The model allows different genetic architectures for each phenotype. (2) The covariance structure depends exclusively in the relationships among phenotypes since  $\boldsymbol{\Omega}$  depends only on the partial regression coefficients relating phenotypes ( $\beta$ 's) and variances of error terms ( $\sigma^2$ 's), and not on the genetic architectures defined by the  $\boldsymbol{\theta}$ 's. (3) When the correlation between two phenotypes arises exclusively because of a pleiotropic QTL, conditioning on the QTL genotypes makes the phenotypes independent; thus, the concentration matrix of the conditional model does not depend on the genetic architecture. (4) This model can represent acyclic and cyclic networks. However, we focus on acyclic networks in this paper.

We now show that our model corresponds to a homogeneous conditional Gaussian regression model. The conditional Gaussian (CG) parametric family models the covariation of discrete and continuous random variables. Continuous random variables conditional on discrete variables are multivariate normal [Lauritzen (1996)]. The joint distribution of the vectors of discrete ( $\mathbf{q}_i$ ) and continuous ( $\mathbf{y}_i$ ) variables have a density  $f$  such that

$$(2.2) \quad \log f(\mathbf{q}_i, \mathbf{y}_i) = g(\mathbf{q}_i) + \mathbf{h}'(\mathbf{q}_i) \mathbf{y}_i - \mathbf{y}_i' \mathbf{K}(\mathbf{q}_i) \mathbf{y}_i / 2,$$

where  $g(\mathbf{q}_i)$  is a scalar,  $\mathbf{h}(\mathbf{q}_i)$  is a vector and  $\mathbf{K}(\mathbf{q}_i)$  is a positive definite matrix. The density  $f$  depends on observed markers  $\mathbf{m}_i$  as

$$(2.3) \quad \log f(\mathbf{q}_i, \mathbf{y}_i) = \log p(\mathbf{y}_i, \mathbf{q}_i \mid \mathbf{m}_i),$$

where  $g(\mathbf{q}_i) = \log p(\mathbf{q}_i \mid \mathbf{m}_i) - \frac{1}{2}(T \log 2\pi - \log \det(\boldsymbol{\Omega}) + \sum_{t=1}^T \mu_{ti}^{*2} / \sigma_t^2)$ . Observe that the linear coefficients  $\mathbf{h}(\mathbf{q}_i) = \boldsymbol{\gamma}_i$  depend on  $\mathbf{q}_i$  through  $\mathbf{X}_{ti}$ , while the concentration matrix  $\mathbf{K}(\mathbf{q}_i) = \boldsymbol{\Omega}$  does not. Thus, our model is a homogeneous CG model

[Lauritzen (1996), page 160]. Furthermore, since our genetic model was derived from a set of regression equations with normal errors, our model is in the homogeneous conditional Gaussian regression parametric family.

**3. A causal framework for systems genetics.** This section formalizes our causal inference framework for systems genetics that combines ideas from randomized experiments with a purely probabilistic approach to causal inference. We argue that while causal claims about the relationship between QTLs and phenotypes are justified by randomization of alleles during meiosis and the unidirectional influence of genotype on phenotype, causal claims about the relationships between phenotypes follow from conditional probability. In a nutshell, by adding QTL nodes to phenotype networks, we can distinguish between phenotype networks that would, otherwise, be distribution equivalent.

In order to formalize our approach, we first show that adding causal QTL nodes can break Markov-equivalence among phenotype networks by creating new conditional independence relationships among nodes. Second, we note that two models in the HCGR parametric family are distribution equivalent if and only if they are Markov equivalent. The last result together with Theorem 1 (see below) provide a simple graphical criterion to determine whether two DAGs belonging to the HCGR parametric family are distribution equivalent.

The analogy between the randomization of alleles during meiosis and a randomized experimental design was first pointed out by Li et al. (2006). Causality can be unambiguously inferred from a randomized experiment for two reasons [Dawid (2007)]: (1) the treatment to an experimental unit (genotype) precedes measured outcomes (phenotypes); and (2) random allocation of treatments to experimental units guarantees that other common causes are averaged out. The central dogma of molecular biology [Crick (1958)] suggests that QTL genotypes generally precede phenotypes. For the types of Eukaryotic data which we analyze (including clinical traits gene expression and metabolites), causality must go from QTL to phenotype. Furthermore, random allocation of QTL genotypes eliminates confounding from other genetic and environmental effects.

Causal relationships among phenotypes require the additional assumption of conditional independence. Suppose a QTL,  $Q$ , and two phenotypes mapping to  $Q$ ,  $Y_1$  and  $Y_2$ , have true causal relationship  $Q \rightarrow Y_1 \rightarrow Y_2$ . That is,  $Y_2$  is independent of  $Q$  given  $Y_1$ . The randomization of genotypes in  $Q$  leads to a randomization of  $Y_1$ , thus averaging out confounding affects. However, precedence of the ‘randomized’  $Y_1$  before the ‘outcome’  $Y_2$  cannot in general be determined a priori. Conditional independence is the key to determine causal order among phenotypes [Schadt et al. (2005), Li et al. (2006), Chen, Emmert-Streib and Storey (2007), Liu, de la Fuente and Hoeschele (2008), Chaibub Neto et al. (2008)].

In the remainder of this section we present some graphical model definitions and results needed in the formalization of our causal graphical models framework. A path is any unbroken, nonintersecting sequence of edges in a graph, which may go along or against the direction of the arrows.

DEFINITION 1 (d-separation). A path  $p$  is said to be d-separated (or blocked) by a set of nodes  $Z$  if and only if

1.  $p$  contains a chain  $i \rightarrow m \rightarrow j$  or a fork  $i \leftarrow m \rightarrow j$  such that the middle node  $m$  is in  $Z$ , or
2.  $p$  contains an inverted fork (or collider)  $i \rightarrow m \leftarrow j$  such that the middle node  $m$  is not in  $Z$  and such that no descendant of  $m$  is in  $Z$ .

A set  $Z$  is said to d-separate  $X$  from  $Y$  if and only if  $Z$  blocks every path from a node in  $X$  to a node in  $Y$ .  $X$  and  $Y$  are d-connected if they are not d-separated [Pearl (1988, 2000)].

Two graphs are Markov equivalent (or faithful indistinguishable) if they have the same set of d-separation relations [Spirtes, Glymour and Scheines (2000)]. The skeleton of a causal graph is the undirected graph obtained by replacing its arrows by undirected edges. A v-structure is composed by two converging arrows whose tails are not connected by an arrow.

THEOREM 1 (Detecting Markov equivalence). *Two directed acyclic graphs (DAGs) are Markov equivalent if and only if they have the same skeletons and the same set of v-structures [Verma and Pearl (1990)].*

Two models are likelihood equivalent if  $f(y | M_1) = f(y | M_2)$  for any data set  $y$ , where  $f(y | M)$  represent the prior predictive density of the data,  $y$ , conditional on model  $M$  [Heckerman, Geiger and Chickering (1995)]. In this paper we extend the definition of likelihood equivalence to predictive densities obtained by plugging in maximum likelihood estimates in the respective sampling models. A closely related concept, distribution equivalence, states that two models are distribution equivalent if one is a reparametrization of the other. While likelihood equivalence is defined in terms of predictive densities (prior predictive density or sampling model evaluated on the maximum likelihood estimates), distribution equivalence is defined in terms of the sampling model directly. Because of the invariance property of maximum likelihood estimates, distribution and likelihood equivalence are equivalent concepts in the frequentist setting. This is also true in the Bayesian setting with proper priors invariant to model reparameterizations.

Suppose that for each pair of connected phenotypes in a graph there exists at least one QTL affecting one but not the other phenotype. Denote this new graph with QTLs included by the “extended graph.” The next result shows that, in this particular situation, we can distinguish between causal models belonging to a Markov equivalent class of phenotype networks.

RESULT 2. Consider a class of Markov equivalent DAGs  $\mathcal{G}$ . Let  $Y_1$  and  $Y_2$  be any two adjacent nodes in the graphs in  $\mathcal{G}$ . Assume that for each such pair there exists at least one variable,  $Q$ , directly affecting  $Y_1$  but not  $Y_2$ . Let  $\mathcal{G}_E$  represent the class of extended graphs. Then the graphs in  $\mathcal{G}_E$  are not Markov equivalent.

As an illustrative example consider the following class of Markov equivalent models:  $\mathcal{G} = \{Y_1 \rightarrow Y_2 \rightarrow Y_3, Y_1 \leftarrow Y_2 \leftarrow Y_3, Y_1 \leftarrow Y_2 \rightarrow Y_3\}$ . These causal models are Markov equivalent because they have the same set of conditional independence relations, namely,  $Y_1 \perp\!\!\!\perp Y_3 \mid Y_2$ . In accordance with Theorem 1, the three models have the same skeleton,  $Y_1 - Y_2 - Y_3$ , and the same set of v-structures (no v-structures). Now consider one QTL,  $Q$ , affecting  $Y_2$  but not  $Y_1$  and  $Y_3$ . Then  $\mathcal{G}_E$  is composed by

$$\begin{array}{ccc} \begin{array}{c} Q \\ \downarrow \\ Y_1 \rightarrow Y_2 \rightarrow Y_3 \end{array} & \begin{array}{c} Q \\ \downarrow \\ Y_1 \leftarrow Y_2 \leftarrow Y_3 \end{array} & \begin{array}{c} Q \\ \downarrow \\ Y_1 \leftarrow Y_2 \rightarrow Y_3 \end{array} \end{array}$$

Observe that these models still have the same skeleton but different sets of v-structures:  $Y_1 \rightarrow Y_2 \leftarrow Q$ ,  $Q \rightarrow Y_2 \leftarrow Y_3$  and  $\emptyset$ , respectively. The next result guarantees that for the HCGR parametric family Markov equivalence implies distribution equivalence and vice-versa.

**RESULT 3.** For the HCGR parametric family, two DAGs are distribution equivalent if and only if they are Markov equivalent.

It follows from Results 2 and 3 that by extending the phenotype network to include QTLs we are able to reduce the size of and equivalence class of graphs (possibly to a single network). Furthermore, if we consider Theorem 1 and Result 3 together, we have the following:

**RESULT 4.** For the HCGR parametric family, two DAGs are distribution equivalent if and only if they have the same skeletons and same sets of v-structures.

Result 4 provides a simple graphical criterion to determine whether two HCGR models are distribution equivalent. This allows us to determine distribution equivalence by inspection of graph structures without the need to go through algebraic manipulations of joint probability distributions as in Chaibub Neto et al. (2008).

**4. QTL mapping and phenotype network structure.** In this section we show that the conditional LOD score can be used as a formal measure of conditional independence relationships between phenotypes and QTLs. Even though in this paper we restrict our attention to HCGR models, conditional LOD profiling is a general framework for the detection of conditional independencies between continuous and discrete random variables and does not depend on the particular parametric family adopted in the modeling. Contrary to partial correlations, the conditional LOD score does not require the assumption of multinormality of the data in order to formally test for independence [recall that only in the Gaussian

case, a zero (partial) correlation implies statistical (conditional) independence], and it can handle interactive covariates.

The conditional LOD score is defined as

$$(4.1) \quad \begin{aligned} \text{LOD}(y, q | x) &= \text{LOD}(y, q, x) - \text{LOD}(y, x) \\ &= \log_{10} \left\{ \frac{f(y | q, x)}{f(y)} \right\} - \log_{10} \left\{ \frac{f(y | x)}{f(y)} \right\}, \end{aligned}$$

where  $f(\cdot)$  represents a predictive density (a maximized likelihood or the prior predictive density in a Bayesian setting). It follows directly from this definition that

$$(4.2) \quad \text{LOD}(y, q | x) = 0 \quad \Leftrightarrow \quad f(y | q, x) = f(y | x) \quad \Leftrightarrow \quad Y \perp\!\!\!\perp Q | X.$$

Therefore, we can use conditional LOD scores as a formal measure of independence between continuous ( $Y$ ) and discrete ( $Q$ ) random variables, conditional on any set of variables  $X$ , that could be either continuous, discrete or both.

Furthermore, the conditional LOD score can be used to formally test for conditional independence in the presence of interacting covariates (denoted by  $X \cdot Q$ ) since

$$(4.3) \quad \text{LOD}(y, q | x, x \cdot q) = \log_{10} \left\{ \frac{f(y | q, x, x \cdot q)}{f(y)} \right\} - \log_{10} \left\{ \frac{f(y | x)}{f(y)} \right\} = 0$$

if and only if  $Y \perp\!\!\!\perp \{Q, X \cdot Q\} | X$ . This is a very desirable property since, in general, testing for conditional independence in the presence of interactions is not straightforward. For example, Andrei and Kendziorski (2009) point that in the presence of interactions, there is no one-to-one correspondence between zero partial correlations and conditional independencies, even when we assume normality of the full conditional distributions.

Traditional QTL mapping focuses on single trait analysis, where the network structure among the phenotypes is not taken into consideration in the analysis. Thus, single-trait analysis may detect QTLs that directly affect the phenotype under investigation, as well as QTLs with indirect effects, affecting phenotypes upstream to the phenotype under study. Consider, for example, the causal graph in Figure 1. The outputs of single trait analysis when Figure 1 represents the true network are given in Figure 2.

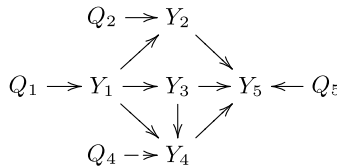


FIG. 1. Example network with five phenotypes and four QTLs.



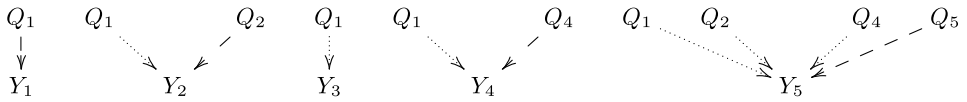


FIG. 2. Output of a single trait QTL mapping analysis for the phenotypes in Figure 1. Dashed and pointed arrows represent direct and indirect QTL/phenotype causal relationships, respectively.

Now let's consider QTL mapping according to the phenotype network structure. When the phenotype structure corresponds to the true causal model, we avoid detecting indirect QTLs by simply performing mapping analysis of the phenotypes conditional on their parents. For example, in Figure 1, if we perform a mapping analysis of  $Y_5$  conditional on  $Y_2, Y_3$  and  $Y_4$ , we do not detect  $Q_1, Q_2$  and  $Q_4$  because  $Y_5 \perp\!\!\!\perp Q_1 \mid Y_2, Y_3, Y_4, Y_5 \perp\!\!\!\perp Q_2 \mid Y_2, Y_3, Y_4$  and  $Y_5 \perp\!\!\!\perp Q_4 \mid Y_2, Y_3, Y_4$ . We only detect  $Q_5$  since  $Y_5 \not\perp\!\!\!\perp Q_5 \mid Y_2, Y_3, Y_4$ .

Now consider Figure 3(a). If we perform a mapping analysis of  $Y_5$  conditional on  $Y_3$  and  $Y_4$ , we still detect  $Q_1$  and  $Q_2$  as QTLs for  $Y_5$ , since failing to condition on  $Y_2$  leaves the paths  $Q_1 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_5$  and  $Q_2 \rightarrow Y_2 \rightarrow Y_5$  in Figure 1 open. In other words,  $Q_1$  and  $Q_2$  are d-connected to  $Y_5$  conditional on  $(Y_3, Y_4)$  in the true causal graph. Mistakenly inferring that a QTL has a direct effect when in reality it indirectly affects the phenotype is problematic, but not a serious concern.

On the other hand, if we map an upstream phenotype conditional on downstream phenotypes, we could infer that downstream QTLs are causal. This would be a serious problem, as it would dramatically reverse the causal flow. Consider, for example, Figure 3(b). If we perform a mapping analysis of  $Y_4$  conditional on  $Y_1, Y_3$  and  $Y_5$ , we incorrectly detect  $Q_5$  as a QTL for  $Y_4$  because in the true network the paths  $Y_4 \rightarrow Y_5 \leftarrow Q_5$  and  $Y_4 \leftarrow Y_3 \rightarrow Y_5 \leftarrow Q_5$  in Figure 1 are open when we condition on  $Y_5$ . That is, if we perform mapping analysis of a phenotype conditional on phenotypes located downstream in the true network, we induce dependencies between the upstream phenotype and QTLs affecting downstream phenotypes, and we erroneously conclude that a downstream QTL affects an upstream phenotype. However, a model with reversed causal relationships among phenotypes and incorrectly having downstream QTLs detected as direct QTLs for the upstream node will generally have a lower marginal likelihood score than the

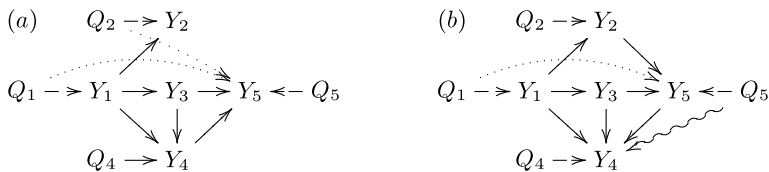


FIG. 3. QTL mapping tailored to the network structure. (a) and (b) display the results of QTL mapping according to slightly altered network structures from Figure 1. Dashed, pointed and wiggled arrows represent, respectively, direct, indirect and incorrect QTL/phenotype causal relationships.

model with the correct causal order for the phenotypes and correct genetic architecture. Therefore, in practice, our model selection procedure protects against this type of mistake.

**5. QTLnet algorithm.** In this section we propose a statistical framework (QTLnet) for the joint inference of phenotype network structure and genetic architecture in systems genetics studies. Work to date in genetical network reconstruction has treated the problems of QTL inference and phenotype network reconstruction separately, generally performing genetic architecture inference first, and then using QTLs to help in the determination of the phenotype network structure [Chaibub Neto et al. (2008), Zhu et al. (2008)]. As indicated in the previous section, such strategy can incorporate QTLs with indirect effects into the genetic architecture of phenotypes.

The great challenge in the reconstruction of networks is that the graph space grows super-exponentially with the number of nodes, so that exhaustive searches are impractical even for small networks, and heuristic approaches are needed to efficiently traverse the graph space. The Metropolis–Hastings algorithm below integrates the sampling of network structures [Madigan and York (1995), Husmeier (2003)] and QTL mapping.

Let  $\mathcal{M}$  represent the structure of a phenotype network composed of  $T$  nodes. The posterior probability of a specified structure is given by

$$(5.1) \quad p(\mathcal{M} | \mathbf{y}, \mathbf{q}) = \frac{p(\mathbf{y} | \mathbf{q}, \mathcal{M})p(\mathcal{M})}{\sum_{\mathcal{M}} p(\mathbf{y} | \mathbf{q}, \mathcal{M})p(\mathcal{M})},$$

where the marginal likelihood

$$(5.2) \quad p(\mathbf{y} | \mathbf{q}, \mathcal{M}) = \int_{\Gamma} p(\mathbf{y} | \mathbf{q}, \Gamma, \mathcal{M})p(\Gamma | \mathcal{M})d\Gamma$$

is obtained by integrating the product of the prior and likelihood of the HCGR model with respect to all parameters  $\Gamma$  in the model. Assuming that the phenotype network is a DAG, the likelihood function factors according to  $\mathcal{M}$  as

$$(5.3) \quad p(\mathbf{y}_i | \mathbf{q}_i, \Gamma, \mathcal{M}) = \prod_t p(y_{ti} | \mathbf{q}_{ti}, \text{pa}(y_t)),$$

where

$$(5.4) \quad p(y_{ti} | \mathbf{q}_{ti}, \text{pa}(y_t)) = N\left(\mu_{ti}^* + \sum_{y_k \in \text{pa}(y_t)} \beta_{tk}y_{ki}, \sigma_t^2\right)$$

and the problem factors out as a series of linear regression models. (Note that QTL genotypes  $\mathbf{q}_{ti}$  enter the model through  $\mu_{ti}^*$ .)

We estimate the posterior probability in (5.1) using a Metropolis–Hastings algorithm detailed in Section 1 of the Supplement. The M–H proposals, which make

single changes (add or drop an edge, or change causal direction), require remapping of any phenotypes that have altered sets of parent nodes. The accept/reject calculation involves estimation of the marginal likelihood conditional on the parent nodes and newly mapped QTL(s).

Because the graph space grows rapidly with the number of phenotype nodes, the network structure with the highest posterior probability may still have a very low probability. Therefore, instead of selecting the network structure with the highest posterior probability, we perform Bayesian model averaging [Hoeting et al. (1999)] for the causal relationships between phenotypes and infer an averaged network. Explicitly, let  $\Delta_{uv}$  represent a causal relationship between phenotypes  $u$  and  $v$ , that is,  $\Delta_{uv} = \{Y_u \rightarrow Y_v, Y_u \leftarrow Y_v, Y_u \not\rightarrow Y_v \text{ and } Y_u \not\leftarrow Y_v\}$ . Then

$$\begin{aligned} p(\Delta_{uv} | \mathbf{y}) &= \sum_k p(\Delta_{uv} | \mathcal{M}_k, \mathbf{y}, \mathbf{q}) p(\mathcal{M}_k | \mathbf{y}, \mathbf{q}) \\ (5.5) \qquad \qquad &= \sum_k \mathbb{1}\{\Delta_{uv} \in \mathcal{M}_k\} p(\mathcal{M}_k | \mathbf{y}, \mathbf{q}). \end{aligned}$$

The averaged network is constructed by putting together all causal relationships with maximum posterior probability or with posterior probability above a predetermined threshold.

**6. Simulations.** In this section we evaluate the performance of the QTLnet approach in simulation studies of a causal network with five phenotypes and four causal QTLs. We consider situations with strong or weak causal signals, leading respectively to high or low phenotype correlations. We show that important features of the causal network can be recovered. Further, this simulation illustrates how an alleged hot spot could be explained by sorting out direct and indirect effects of the QTLs.

We generated 1000 data sets according to Figure 1. Each simulated cross object [Broman et al. (2003)] had 5 phenotypes simulated for an F2 population with 500 individuals. The genome had 5 chromosomes of length 100 cM with 10 equally spaced markers per chromosome. We simulated one QTL per phenotype, except for phenotype  $Y_3$  with no QTLs. The QTLs  $Q_t$ ,  $t = 1, 2, 4, 5$ , were unlinked and placed at the middle marker on chromosome  $t$ .

Each simulated cross object had different sampled parameter value combinations for each realization. In the strong signal simulation, we sampled the additive and dominance effects according to  $U[0.5, 1]$  and  $U[0, 0.5]$ , respectively. The partial regression coefficients for the phenotypes were sampled according to  $\beta_{uv} \sim 0.5U[-1.5, -0.5] + 0.5U[0.5, 1.5]$ . In the weak signal simulation, we generated data sets with additive and dominance effects from  $U[0, 0.5]$  and  $U[0, 0.25]$ , respectively, and partial regression coefficients sampled with  $\beta_{uv} \sim U[-0.5, 0.5]$ . The residual phenotypic variance was fixed at 1 in both settings.

TABLE 1

Frequencies of QTL detection for both unconditional (top half) and conditional (bottom half) QTL mapping according with the true phenotype network structure in Figure 1. Results, for each simulation, based on 1000 simulated data sets described in the text. The expected architecture is the set of  $d$ -connected QTLs for the phenotype conditioning with respect to the network in Figure 1

Phenotypes	Strong signal				Weak signal				Expected architecture
	$Q_1$	$Q_2$	$Q_4$	$Q_5$	$Q_1$	$Q_2$	$Q_4$	$Q_5$	
$Y_1$	0.997	0.000	0.000	0.000	0.431	0.000	0.000	0.000	$\{Q_1\}$
$Y_2$	0.884	0.930	0.000	0.000	0.001	0.384	0.000	0.000	$\{Q_1, Q_2\}$
$Y_3$	0.941	0.000	0.000	0.000	0.003	0.000	0.000	0.000	$\{Q_1\}$
$Y_4$	0.603	0.000	0.690	0.000	0.003	0.000	0.370	0.000	$\{Q_1, Q_4\}$
$Y_5$	0.637	0.321	0.321	0.340	0.000	0.000	0.001	0.336	$\{Q_1, Q_2, Q_4, Q_5\}$
$Y_2   Y_1$	0.001	0.999	0.000	0.000	0.000	0.424	0.000	0.000	$\{Q_2\}$
$Y_3   Y_1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	$\emptyset$
$Y_4   Y_1, Y_3$	0.000	0.000	0.999	0.000	0.000	0.000	0.422	0.000	$\{Q_4\}$
$Y_5   Y_2, Y_3, Y_4$	0.000	0.000	0.000	0.999	0.000	0.000	0.000	0.415	$\{Q_5\}$

We first show the accuracy of the mapping analysis in our simulated data sets. We used interval mapping with a LOD score threshold of 5 to detect significant QTLs. Table 1 shows the results of both unconditional and QTL mapping according to the phenotype network in Figure 1. In the strong signal setting, the unconditional mapping often detected indirect QTLs, but the mapping of phenotypes conditional on their parent nodes increased detection of the true genetic architectures. In the weak signal simulation, the unconditional mapping did not detect indirect QTLs in most cases, but we still observe improvement in detection of the correct genetic architecture when we condition on the parents.

The expected architecture contains the  $d$ -connected QTLs when conditioning (or not) on other phenotypes as indicated in the first column of Table 1. For instance,  $Q_1$  and  $Q_2$  are  $d$ -connected to  $Y_2$ , but only  $Q_2$  is  $d$ -connected to  $Y_2$  when properly conditioning on  $Y_1$ . Supplementary Tables S1, S2, S3, S4 and S5 show the simulation results for all possible conditional mapping combinations.

For each simulated data set we applied the QTLnet algorithm using simple interval mapping for QTL detection. The ratio of marginal likelihoods in the Metropolis–Hastings algorithm was computed using the BIC asymptotic approximation to the Bayes factor (equation 1.1 on the Supplement). We adopted uniform priors over network structures. We ran each Markov chain for 30,000 iterations, sampled a structure every 10 iterations, and discarded the first 300 (burnin) network structures producing posterior samples of size 2700. Posterior probabilities for each causal relationship were computed via Bayesian model averaging.

Table 2 shows the frequency, out of the 1000 simulations, the true model was the most probable, second most probable, etc. The results show that in the strong

TABLE 2  
*Frequencies that the posterior probability of the true model was the highest, second highest, etc.  
 Results based on 1000 simulated data sets described in the text*

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	≥13th
Strong	842	100	21	11	3	4	3	2	1	1	3	1	8
Weak	21	33	18	19	19	16	17	15	8	12	13	5	804

signal setting the true model got the highest posterior probability in most of the simulations. In the weak signal setting the range of rankings was very wide-spread.

Table 3 shows the proportion of times that each possible causal direction ( $Y_u \rightarrow Y_v, Y_u \leftarrow Y_v$ ) or no causal relation ( $\{Y_u \not\rightarrow Y_v, Y_u \not\leftarrow Y_v\}$ ) had the highest posterior probability for all pairs of phenotypes. The results show that in the strong signal simulations, the correct causal relationships were recovered with high probability. The results are weaker but in the correct direction in the weak signal setting.

Interestingly, single trait analysis with strong signal showed that  $Y_5$  mapped to  $Q_1$  more frequently than to  $Q_5$  (Table 1). This result can be understood using a path analysis [Wright (1934)] argument. In path analysis, we decompose the correlation between two variables among all paths connecting the two variables in a graph. Let  $\mathcal{D}_{uv}$  represent the set of all direct and indirect directed paths connecting  $u$  and  $v$  (a directed path is a path with all arrows pointing in the same direction).

TABLE 3  
*Frequencies that each possible causal relationship had the highest posterior probability (computed via Bayesian model averaging). Results based on 1000 simulated data sets described in the text*

Phenotypes	Strong signal			Weak signal		
	$\rightarrow$	$\leftarrow$	$\not\rightarrow, \not\leftarrow$	$\rightarrow$	$\leftarrow$	$\not\rightarrow, \not\leftarrow$
(1, 2)	0.996	0.002	0.002	0.594	0.177	0.229
(1, 3)	0.990	0.002	0.008	0.471	0.263	0.266
(1, 4)	0.990	0.001	0.009	0.541	0.196	0.263
(1, 5)	0.054	0.002	0.944	0.028	0.005	0.967
(2, 3)	0.016	0.022	0.962	0.018	0.017	0.965
(2, 4)	0.037	0.012	0.951	0.018	0.015	0.967
(2, 5)	0.997	0.003	0.000	0.712	0.075	0.213
(3, 4)	0.967	0.031	0.002	0.482	0.253	0.265
(3, 5)	0.997	0.003	0.000	0.653	0.116	0.231
(4, 5)	0.996	0.004	0.000	0.670	0.115	0.215

Then the correlation between these nodes can be decomposed as

$$(6.1) \quad \begin{aligned} \text{cor}(y_u, y_v) &= \sum_{P \in \mathcal{D}_{uv}} \phi_{p_2 u} \phi_{p_3 p_2} \cdots \phi_{v p_{m-1}} \\ &= \left\{ \frac{\text{var}(y_u)}{\text{var}(y_v)} \right\}^{1/2} \sum_{P \in \mathcal{D}_{uv}} \beta_{p_2 u} \beta_{p_3 p_2} \cdots \beta_{v p_{m-1}}, \end{aligned}$$

where  $\phi_{ij} = \beta_{ij} \{\text{var}(y_j)/\text{var}(y_i)\}^{1/2}$  is a standardized path coefficient. Assuming intra-locus additivity and encoding the genotypes as 0, 1 and 2 (for the sake of easy computation), we have from equation (6.1) that

$$\begin{aligned} \text{cor}(y_5, q_1) &= \beta_{1,q_1} (\beta_{52} \beta_{21} + \beta_{53} \beta_{31} + \beta_{54} \beta_{41} + \beta_{54} \beta_{43} \beta_{31}) \left\{ \frac{\text{var}(q_1)}{\text{var}(y_5)} \right\}^{1/2}, \\ \text{cor}(y_5, q_5) &= \beta_{1,q_5} \left\{ \frac{\text{var}(q_5)}{\text{var}(y_5)} \right\}^{1/2}. \end{aligned}$$

We therefore see that if the partial regression coefficients between phenotypes are high, and the QTL effects  $\beta_{1,q_1}$  and  $\beta_{5,q_5}$  and QTL variances are close (as in the strong signal simulation), then  $\text{cor}(y_5, q_1)$  will be higher than  $\text{cor}(y_5, q_5)$  and  $Y_5$  will map to  $Q_1$  with stronger signal than to  $Q_5$ .

This result suggests a possible scenario for the appearance of eQTL hot spots when phenotypes are highly correlated. A set of correlated phenotypes may be better modeled in a causal network with one upstream phenotype that in turn has a causal QTL. Ignoring the phenotype network can result in an apparent hot spot for the correlated phenotypes. Here, all phenotypes detect QTL  $Q_1$  with high probability when mapped unconditionally (Table 1). In the weak signal setting, the phenotypes map mostly to their respective QTLs and do not show evidence for a hot spot. No hot spot was found in additional simulations having strong QTL/phenotype relationships and weak phenotype/phenotype relations (results not shown). Thus, our QTLnet approach can effectively explain a hot spot found with unconditional mapping when phenotypes show strong causal structure.

**7. Network inference for a liver hot spot.** In this section we illustrate the application of QTLnet to a subset of gene expression data derived from a F2 intercross between inbred lines *C3H/HeJ* and *C57BL/6J* [Ghazalpour et al. (2006), Wang et al. (2006)]. The data set is composed of genotype data on 1,065 markers and liver expression data on the 3421 available transcripts from 135 female mice. Interval mapping indicates that 14 transcripts map to the same region on chromosome 2 with a LOD score higher than 5.3 (permutation  $p$ -value  $< 0.001$ ). Only one transcript, *Pscdbp*, is located on chromosome 2 near the hot spot locus. The 14 transcripts show a strong correlation structure, and the correlation structure adjusting for the peak marker on chromosome 2, *rs3707138*, is still strong (see Supplementary Table S6). This co-mapping suggest all transcripts are under

the regulation of a common factor. Causal relationships among phenotypes could explain the strong correlation structure that we observe, although other possibilities are environmental factors or a latent factor that is not included.

We applied the QTLnet algorithm on the 129 mice that had no missing transcript data using Haley–Knott (1992) regression (and assuming genotyping error rate of 0.01) for the detection of QTLs conditional on the network structure, and we used the BIC approximation to estimate the marginal likelihood ratio in the Metropolis–Hastings algorithm (equation 1.1 on the Supplement). We adopted uniform priors over network structures. We ran a Markov chain for 1,000,000 iterations and sampled a network structure every 100 iterations, discarding the first 1000 and basing inference on a posterior sample with 9000 network structures. Diagnostic plots and measures (see Section 3 on the Supplement) support the convergence of the Markov chain.

We performed Bayesian model averaging and, for each of 91 possible pairs  $(Y_u, Y_v)$ , we obtained the posterior probabilities of  $Y_u \rightarrow Y_v$ ,  $Y_u \leftarrow Y_v$  and of no direct causal connection. The results are shown in Supplementary Table S7. Figure 4 shows a model-averaged network.

This network suggests a key role of *Il16* in the regulation of the other transcripts in the network. *Il16* is upstream to all other transcripts, and is the only one directly

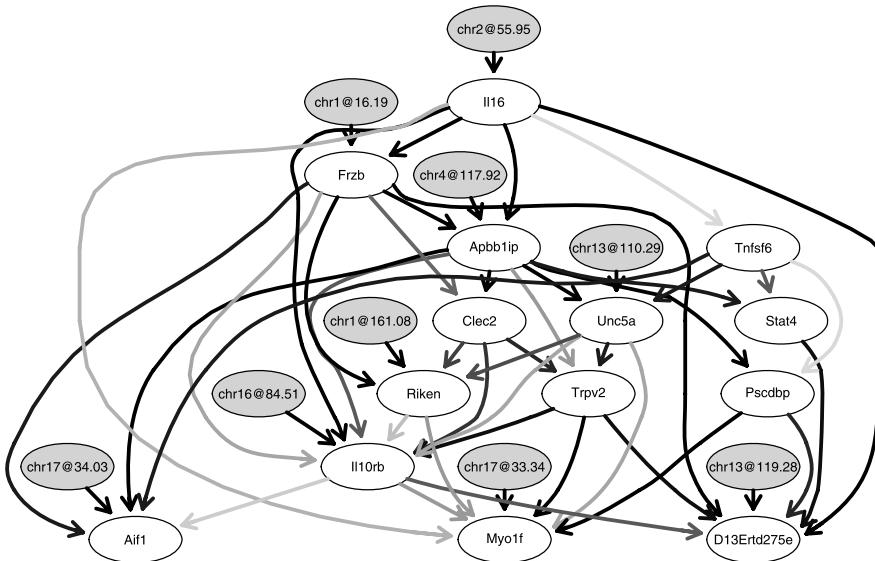


FIG. 4. Model-averaged posterior network. Arrow darkness is proportional to the posterior probability of the causal relation computed via Bayesian model averaging. For each pair of phenotypes, the figure displays the causal relationship (presence or absence of an arrow) with highest posterior probability. Light grey nodes represent QTLs and show their chromosome number and position in centimorgans. *Riken* represents the *riken* gene 6530401C20Rik.

mapping to the locus of chromosome 2. We would have expected the *cis* transcript, *Pscdbp*, to be the upstream phenotype in this network. However, the data suggests *Pscdbp* is causal to only two other transcripts and that some other genetic factor on chromosome 2 may be driving this pathway. This estimated QTLnet causal network provides new hypotheses that could be tested in future mouse experiments.

**8. Discussion.** We have developed a statistical framework for causal inference in systems genetics. Causal relationships between QTLs and phenotypes are justified by the randomization of alleles during meiosis together with the unidirectional influence of genotypes on phenotypes. Causal relationships between phenotypes follows from breakage of distribution equivalence due to QTL nodes augmenting the phenotype network. We have proposed a novel approach to jointly infer genetic architecture and causal phenotype network structure using HCGR models. We argue in this paper that failing to properly account for phenotype network structure for mapping analysis can yield QTLs with indirect effects in the genetic architecture, which can decrease the power to detect the correct causal relationships between phenotypes.

Current literature in systems genetics [Chaibub Neto et al. (2008), Zhu et al. (2008)] has considered the problems of genetic architecture and phenotype network structure inference separately. Chaibub Neto et al. (2008) used the PC-algorithm [Spirtes, Glymour and Scheines (2000)] to first infer the skeleton of the phenotype network and then use QTLs to determine the directions of the edges in the phenotype network. Zhu et al. (2008) reconstructed networks from a consensus of Bayesian phenotype networks with a prior distribution based on causal tests of Schadt et al. (2005). Their prior was computed with QTLs determined by single trait analysis.

Liu, de la Fuente and Hoeschele (2008) presented an approach based in structural equation models (and applicable to species where sequence information is available) that partially accounts for the phenotype network structure when selecting the QTLs to be incorporated in the network. They perform eQTL mapping using *cis*, *cis-trans* and *trans*-regulation [Doss et al. (2005), Kulp and Jagalur (2006)] and then use local structural models to identify regulator-target pairs for each eQTL. The identified relationships are then used to construct an encompassing directed network (EDN) with nodes composed by transcripts and eQTLs and arrows from (1) eQTLs to *cis*-regulated target transcripts; (2) *cis*-regulated transcripts to *cis-trans*-regulated target transcripts; and (3) *trans*-regulator transcripts to target transcripts, and from *trans*-eQTL to target transcripts. The EDN defines a network search space for network inference with model selection based on penalized likelihood scores and an adaptation of Occam's window [Madigan and Raftery (1994)]. Their local structural models, which fit at most two candidate regulators per target transcript, can include indirect eQTLs in the genetic architecture of target transcripts when there are multiple paths connecting a *cis*-regulator to a *cis-trans*-target transcript. In other words, some transcripts identified as *cis*-regulated targets may actually be *cis-trans*.



Winrow et al. (2009) rely on a (nonhomogeneous) conditional Gaussian model and employ the standard Metropolis–Hastings (M–H) algorithm (add, remove or delete edges) to search over the space of DAG structures. It differs from our approach in how QTL detection is coupled with the M–H proposal. QTLnet constructs M–H proposals for edges connecting phenotypes and detects QTLs conditional on the network structure. On the other hand, Winrow et al. first detect QTLs and then construct M–H proposals for both phenotypes and QTL nodes. In Section 4 of the Supplement we present a simulation study comparing our approach to Winrow’s strategy. QTLnet inferred the proper genetic architecture and phenotype network structure at a higher rate than Winrow’s approach with high signal-to-noise ratio. In the weak signal setting, QTLnet picks up direct causal relationship at a higher rate, but Winrow’s is better when there is no direct causal link; on average, they were comparable. Although Winrow’s strategy also avoids the detection of indirect QTLs, it is prone to miss direct QTLs when the causal effects among phenotypes are strong, and can perform poorly depending on the structure of the phenotype network (see Supplement for details).

Current interest in the eQTL literature centers on understanding the relationships among expression traits that co-map to a genomic region. It is often suggested that these eQTL “hot spots” result from a master regulator affecting the expression levels of many other genes [see Breitling et al. (2008)]. A path analysis argument suggests that if the correlation structure between the phenotypes is strong because of a strong causal signal, a well-defined hot spot pattern will likely appear when we perform single trait analysis. Our simulations and real data example suggest that this is the situation where the QTLnet algorithm is expected to be most fruitful.

The QTLnet approach is based on a Metropolis–Hastings algorithm that at each step proposes a slightly modified phenotype network and fits the genetic architecture conditional on this proposed network. Conditioning on the phenotype network structure should generally lead to a better inferred genetic architecture. Likewise, a better inferred genetic architecture should lead to a better inferred phenotype structure (models with better inferred genetic architectures should have higher marginal likelihood scores. A poorly inferred genetic architecture may compromise the marginal likelihood of a network with phenotype structure close to the true network).

Because the proposal mechanism of the Metropolis–Hastings algorithm is based in small modification of the last accepted network (addition, deletion or reversion of a single edge), the mixing of the Markov chain is generally slow and it is necessary to run long chains and use big thinning windows in order to achieve good mixing. This is a bottleneck to the scalability of this approach. We therefore plan to investigate more efficient versions of the Metropolis–Hastings algorithm for network structure inference. In particular, a new and more extensive edge reversal move proposed by Grzegorzczuk and Husmeier (2008) and an approach based in a Markov blanket decomposition of the network [Riggelsen (2005)].

Our method is currently implemented with R code. The analysis of the real data example took over 20 hours in a 64 bit Intel(R) Core(TM) 2 Quad 2.66 GHz machine. We can handle up to 20 phenotype nodes, at this point. The run time for the code, written in R, should be substantially improved as we optimize code, converting key functions to C (under development). Nonetheless, because the number of DAGs increases super-exponentially with the number of phenotype nodes, scaling up the proposed approach to large networks will likely be a very challenging task. Current R code is available from the authors upon request.

One of the most attractive features of a Bayesian framework is its ability to formally incorporate prior information in the analysis. Given the complexity of biological processes and the many limitations associated with the partial pictures provided by any of the “omic” data sets now available, incorporation of external information is highly desirable. We are currently working in the development of priors for network structures.

The QTLnet approach can be seen as a method to infer causal Bayesian networks composed of phenotype and QTL nodes. Standard Bayesian networks provide a compact representation of the conditional dependency and independencies associated with a joint probability distribution. The main criticism of a causal interpretation of such networks is that different structures may be likelihood equivalent while representing totally different causal process. In other words, we can only infer a class of likelihood equivalent networks. We have formally shown how to break likelihood equivalence by incorporating causal QTLs.

We have focussed on experimental crosses with inbred founders, as the recombination model and genetic architecture are relatively straightforward. However, this approach might be extended to outbred populations with some additional work. The genotypic effects are random, and the problem needs to be recast in terms of variance components.

In this paper we only consider directed acyclic graphs. We point out, however, that the HCGR parametric family accommodates cyclic networks. The QTLnet approach can only infer causality among phenotypes that are consistent with the assumption of no latent variables and no measurement error. However, these complications can impact network reconstruction. We are currently investigating extensions of the proposed framework along these lines.

**Acknowledgments.** We would like to thank Alina Andrei, the associate editor and two anonymous referees for their helpful comments and suggestions.

#### SUPPLEMENTARY MATERIAL

**Supplement to “Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes”** (DOI: [10.1214/09-AOAS288SUPP](https://doi.org/10.1214/09-AOAS288SUPP); .pdf). The Supplement article presents: (1) the Metropolis–Hastings algorithm for QTLnet; (2) sup-

plementary tables for the simulations and real data example; (3) convergence diagnostics for the real data example; (4) comparison with Winrow et al. (2009); and (5) the proofs of Results 1, 2, 3 and 4.

## REFERENCES

- ANDREI, A. and KENDZIORSKI, C. (2009). An efficient method for identifying statistical interactors in graphical models. *Biostatistics* **10** 706–718.
- ATEN, J. E., FULLER, T. F., LUSIS, A. J. and HORVATH, S. (2008). Using genetic markers to orient the edges in quantitative trait networks: The NEO software. *BMC Sys. Biol.* **2** 34.
- BANERJEE, S., YANDELL, B. S. and YI, N. (2008). Bayesian QTL mapping for multiple traits. *Genetics* **179** 2275–2289.
- BREITLING, R., LI, Y., TESSON, B. M., FU, J., WU, C., WILTSHIRE, T., GERRITS, A., BYSTRYKH, L. V., DE HAAN, G., SU, A. I. and JANSEN, R. C. (2008). Genetical genomics: Spotlight on QTL hotspots. *PLoS Genet.* **4** e1000232.
- BROMAN, K., WU, H., SEN, S. and CHURCHILL, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19** 889–890.
- CHAIBUB NETO, E., FERRARA, C., ATTIE, A. D. and YANDELL, B. S. (2008). Inferring causal phenotype networks from segregating populations. *Genetics* **179** 1089–1100.
- CHAIBUB NETO, E., KELLER, M. P., ATTIE, A. D. and YANDELL, B. S. (2009). Supplement to “Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes.” DOI: 10.1214/09-AOAS288SUPP.
- CHEN, L. S., EMMERT-STREIB, F. and STOREY, J. D. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology* **8** R219.
- CRICK, F. H. C. (1958). On Protein Synthesis. *Symp. Soc. Exp. Biol.* **XII** 139–163.
- DAWID, P. (2007). Fundamentals of statistical causality. Research Report 279, Dept. Statistical Science, Univ. College London.
- DOSS, S., SCHADT, E. E., DRAKE, T. A., LUSIS, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research* **15** 681–691.
- GHAZALPOUR, A., DOSS, S., ZHANG, B., WANG, S., PLAISIER, C., CASTELLANOS, R., BROZELL, A., SCHADT, E. E., DRAKE, T. A., LUSIS, A. J. and HORVATH, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics* **2** e130.
- GRZEGORCZYK, M. and HUSMEIER, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning* **71** 265–305.
- HALEY, C. and KNOTT, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69** 315–324.
- HECKERMAN, D., GEIGER, D. and CHICKERING, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20** 197–243.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial (with discussion and rejoinder by authors). *Statist. Sci.* **14** 382–417. [MR1765176](#)
- HUSMEIER, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* **19** 2271–2282.
- KULP, D. C. and JAGALUR, M. (2006). Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7** 125.
- LAURITZEN, S. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. [MR1419991](#)
- LI, R., TSAIH, S. W., SHOCKLEY, K., STYLIANOU, I. M., WERGEDAL, J., PAIGEN, B. and CHURCHILL, G. A. (2006). Structural model analysis of multiple quantitative traits. *PLoS Genetics* **2** e114.

- LIU, B., DE LA FUENTE, A. and HOESCHELE, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* **178** 1763–1776.
- MADIGAN, D. and RAFTERY, J. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.
- MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Int. Stat. Rev.* **63** 215–232.
- PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Kaufmann, San Mateo, CA. MR0965765
- PEARL, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge Univ. Press, New York. MR1744773
- RIGGELSEN, C. (2005). MCMC learning of Bayesian network models by Markov blanket decomposition. In *Lecture Notes in Computer Science* 329–340. Springer, Berlin.
- SCHADT, E. E., LAMB, J., YANG, X., ZHU, J., EDWARDS, S., GUHATHAKURTA, D., SIEBERTS, S. K., MONKS, S., REITMAN, M., ZHANG, C., LUM, P. Y., LEONARDSON, A., THIERINGER, R., METZGER, J. M., YANG, L., CASTLE, J., ZHU, H., KASH, S. F., DRAKE, T. A., SACHS, A. and LUSIS, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37** 710–717.
- SEN, S. and CHURCHILL, G. A. (2001). A statistical framework for quantitative trait mapping. *Genetics* **159** 371–387.
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675
- VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Readings in Uncertain Reasoning* (G. Shafer and J. Pearl, eds.). Kaufmann, Boston. MR1091985
- WANG, S., YEHA, N., SCHADT, E. E., WANG, H., DRAKE, T. A. and LUSIS, A. J. (2006). Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics* **2** e15.
- WINROW, C. J., WILLIAMS, D. L., KASARSKIS, A., MILLSTEIN, J., LAPOSKY, A. D., YANG, H. S., MRAZEK, K., ZHOU, L., OWENS, J. R., RADZICKI, D., PREUSS, F., SCHADT, E. E., SHIMOMURA, K., VITATERNA, M. H., ZHANG, C., KOBLAN, K. S., RENGER, J. J. and TUREK, F. W. (2009). Uncovering the genetic landscape for multiple sleep-wake traits. *PLoS ONE* **4** e5161.
- WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5** 161–215.
- ZENG, Z. B., WANG, T. and ZOU, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* **169** 1711–1725.
- ZHU, J., ZHANG, B., SMITH, E. N., DREES, B., BREM, R. B., KRUGLYAK, L., BUMGARNER, R. E. and SCHADT, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics* **40** 854–861.

E. CHAIBUB NETO  
 B. S. YANDELL  
 DEPARTMENT OF STATISTICS  
 UNIVERSITY OF WISCONSIN-MADISON  
 1300 UNIVERSITY AVENUE  
 MADISON, WISCONSIN 53706  
 USA  
 E-MAIL: chaibub@stat.wisc.edu  
 byandell@wisc.edu

M. P. KELLER  
 A. D. ATTIE  
 DEPARTMENT OF BIOCHEMISTRY  
 UNIVERSITY OF WISCONSIN-MADISON  
 433 BABCOCK DRIVE  
 MADISON, WISCONSIN 53706  
 USA  
 E-MAIL: mkeller3@wisc.edu  
 adattie@wisc.edu