Estimation of high-dimensional directed acyclic graphs with surrogate intervention

MIN JIN HA*

*Department of Biostatistics, The University of Texas MD Anderson Cancer Center,
1515 Holcombe Boulevard, Houston, TX 77030, USA*

WEI SUN

*Program in Biostatistics and Bioinformatics, Fred Hutchinson Cancer Research Center,
1100 Fairview Ave N, Seattle, WA 98109-1024 USA*

mjha@mdanderson.org

SUMMARY

Directed acyclic graphs (DAGs) have been used to describe causal relationships between variables. The standard method for determining such relations uses interventional data. For complex systems with high-dimensional data, however, such interventional data are often not available. Therefore, it is desirable to estimate causal structure from observational data without subjecting variables to interventions. Observational data can be used to estimate the skeleton of a DAG and the directions of a limited number of edges. We develop a Bayesian framework to estimate a DAG using surrogate interventional data, where the interventions are applied to a set of external variables, and thus such interventions are considered to be surrogate interventions on the variables of interest. Our work is motivated by expression quantitative trait locus (eQTL) studies, where the variables of interest are the expression of genes, the external variables are DNA variations, and interventions are applied to DNA variants during the process of a randomly selected DNA allele being passed to a child from either parent. Our method, surrogate intervention recovery of a DAG (*sirDAG*), first constructs a DAG skeleton using penalized regressions and the subsequent partial correlation tests, and then estimates the posterior probabilities of all the edge directions after incorporating DNA variant data. We demonstrate the utilities of *sirDAG* by simulation and an application to an eQTL study for 550 breast cancer patients.

Keywords: Directed acyclic graphs; eQTL; Surrogate intervention.

1. INTRODUCTION

A directed acyclic graph (DAG) is a useful tool for studying conditional independence relations among a set of random variables, where each vertex represents a random variable and each directed edge represents a conditional dependence relation, with a restriction of no directed cycles. We consider a DAG $\mathcal{G} = (V, E)$ for a set of random variables $Y_1, \dots, Y_p$, with vertices $V = \{1, \dots, p\}$ corresponding to variables $Y_1, \dots, Y_p$, and a collection of directed edges E . We assume that $Y = (Y_1, \dots, Y_p)^T \in \mathbb{R}^p$ follows a multivariate

*To whom correspondence should be addressed.

Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_Y)$. We denote the parent vertices of vertex i by pa_i and the corresponding random variables by Y_{pa_i} , and then can decompose the likelihood of $Y_1, \dots, Y_p$, based on the Markov property that Y_i is independent of all the remaining variables given its parents Y_{pa_i} :

$$f(Y_1, \dots, Y_p) = \prod_{i=1}^p f(Y_i | Y_{\text{pa}_i}). \quad (1.1)$$

Observational data provide the conditionally independent relations among the random variables. Since a set of conditional independence restrictions may be compatible with multiple DAGs, observational data can identify the *skeleton* of a DAG (i.e., all the edges in a DAG without directions) and the directions of a limited number of edges. The collection of all the DAGs corresponding to the same set of conditional independence restrictions constitute a *Markov equivalence* class, and thus observational data can only identify a Markov equivalence class, but cannot distinguish the DAGs within the Markov equivalence class. For example, the DAGs $Y_1 \rightarrow Y_2 \rightarrow Y_3$, $Y_1 \leftarrow Y_2 \leftarrow Y_3$, and $Y_1 \leftarrow Y_2 \rightarrow Y_3$ all represent the (conditional) independence/dependence that $Y_1 \sim Y_2$, $Y_2 \sim Y_3$, $Y_1 \sim Y_3$, and $Y_1 \perp Y_3 | Y_2$, where $A \sim B$ indicates that A and B are dependent, and $A \perp B | C$ indicates that A is independent of B given C . Therefore, these three DAGs constitute a Markov equivalence class and cannot be distinguished by observational data.

In this article, we relate the probabilistic DAG that describes conditional independences among genes to causal DAG representation of genes under assumptions of causal Markov and causal faithful assumptions (Spirtes and others, 2000). We also need to assume that there are no unmeasured confounders for relations among V . For example, in the causal DAG, $\mathcal{G}$, the structure, $Y_1 \leftarrow Y_2 \rightarrow Y_3$ can be interpreted, as Y_2 is the common cause of effects Y_1 and Y_3 . Extensive efforts have been devoted to estimate causally sufficient DAG structures using observational data. These efforts have produced, for example, the PC algorithm (named after its authors, Peter and Clark) (Spirtes and others, 2000), the PC-stable algorithm that resolves the order-dependency of the PC algorithm (Colombo and Maathuis, 2012), the greedy equivalence search algorithm (Chickering, 2003), and the penalized regression methods (Schmidt and others, 2007; Ha and others, 2016). In this article, we aim to estimate DAGs in another setting, a “surrogate experiment” (Bareinboim and Pearl, 2012). In a surrogate experiment, the variables of interest can be influenced by another group of variables that may be subject to interventions, and thus the interventions do not directly apply to the variables of interest. We refer to such interventions as surrogate interventions.

Our method is motivated by the problem of studying gene–gene networks using gene expression quantitative trait locus (eQTL) data. In an eQTL study, both gene expression and the genotypes of DNA variants are collected in a common set of samples. To simplify the discussion in this article, we assume the DNA variants are single nucleotide polymorphisms (SNPs), though other types of DNA variants, such as copy number variants or indels (small insertions or deletions), also can be used.

We use a DAG to describe the gene–gene network, where each vertex is the expression of a gene, and each edge indicates a causal relation between two genes. Interventions on SNP genotypes can influence gene expression, and thus are surrogate interventions. Direct intervention on an SNP genotype is often impossible. The randomization of the SNP alleles passing to daughter cells during meiosis, however, is analogous to a randomized experiment (Li and others, 2006). This is the so-called *Mendelian randomization* (MR), which has been applied to *instrumental variable analysis* (IVA), where genetic markers (SNPs) are instrumental variables to assess the effect of an exposure (e.g., gene expression) on an outcome (Burgess and others, 2017). The IVA for MR is designed to facilitate a randomized trial to control for confounding variables, except that the random allocation of the exposure is performed by nature, rather than by an investigator. The key assumption is that the SNPs have an effect on the exposure, but not on an outcome, except through the exposure. In this article, we have made assumption that there is no unobserved confoundings, and thus MR mainly helps us to justify the association between a SNP and a

gene has a causal interpretation $\text{SNP} \rightarrow \text{gene}$, which is consistent with biological principal that DNA can affect gene expression but gene expression can rarely influence DNA sequence.

Previous studies have used eQTL data to dissect the causal relations among three variables, including an eQTL, a gene expression trait, and a third variable, which can be a clinical phenotype (Schadt and others, 2005), another gene expression trait (Kulp and Jagalur, 2006; Chen and others, 2007), or the activity of a transcription factor (Sun and others, 2007). Neto and others (2008) first estimated the skeleton of a DAG, and then used eQTLs to orient the edges in the network. Then they extended this method to jointly estimate causal networks of gene expression traits and the underlying genetic architecture using Bayesian model averaging and a modified Metropolis–Hastings algorithm (Neto and others, 2010). Hageman and others (2011) jointly estimated gene–gene networks and eQTLs using a Bayesian method, while placing constraints on the network through a structural prior. Another approach for graphical model estimation is the use of structural equation models that permit both cyclic and acyclic graphs. Li and others (2006) employed a score-based model selection method. Logsdon and Mezey (2010) estimated network skeletons by applying an adaptive lasso regression for each gene expression trait, and then transformed the skeleton into a DAG or a directed cyclic graph, based on eQTL perturbations. Cai and others (2013) extended the work of Logsdon and Mezey (2010) by providing the initial parameter estimates for the adaptive lasso from penalized regressions using a lasso penalty.

Despite the success of previous work, using high-dimensional gene expression data with a limited sample size to estimate DAGs remains challenging. We propose a high-dimensional DAG estimation method, surrogate intervention recovery of a DAG (*sirDAG*) that integrates an upstream external dataset to learn causal structure. The *sirDAG* incorporates recent methodological developments on skeleton estimation and eQTL mapping. Specifically, we employ the PenPC method (Ha and others, 2016) to estimate the network skeleton, which uses penalized regression and shows substantial advantages over existing methods, such as the PC-stable algorithm for estimating high-dimensional DAGs (Colombo and Maathuis, 2012). We also employ a recently developed method to identify *cis*-eQTLs, using allele-specific expression (ASE) of RNA-seq data (Sun, 2012). A local or distant eQTL identified by standard eQTL analysis using total expression, rather than ASE, may impact gene expression indirectly by modifying the expression of other genes or biological processes. In contrast, a *cis*-eQTL identified from ASE must have a direct influence on gene expression (Doss and others, 2005). Through simulation studies under various settings, we demonstrate the improved performance of *sirDAG* from other existing methods, even when the small proportion of genes in the network have eQTLs. We illustrate the prognostic utility of our method using The Cancer Genome Atlas (TCGA) breast cancer data. The computational time of *sirDAG* is discussed in Section S1 of the supplementary material available at *Biostatistics* online.

2. METHODS

2.1. Notations

We study the causal relations of p variables $Y = (Y_1, \dots, Y_p)$ by a DAG $\mathcal{G} = (V, E)$. We assume that $Y \sim P_Y$, where the distribution, P_Y , is Markov and faithful to the causal DAG, $\mathcal{G}$. We assume Gaussian distribution on Y that is Markov and faithful to the true unknown underlying causal DAG (Spirtes and others, 2000). We further assume that we observe all the relevant variables, and that there are no hidden confounding variables for the relations among V . Let X be an additional set of variables, such that they are direct causes of the variables $Y_1, \dots, Y_p$ and subject to interventions. Let $\mathcal{I} \subseteq \{Y_1, \dots, Y_p\}$ be the set of vertices that are associated with at least one variable in X . For example, in eQTL studies, $\mathcal{I}$ is the set of genes with at least one eQTL. For any $Y_i \in \mathcal{I}$, $X^{(i)} = (X_1^{(i)}, \dots, X_{q_i}^{(i)})^T \in \mathbb{R}^{q_i}$ denotes the q_i variables that directly influence Y_i ($X_j^{(i)} \rightarrow Y_i$ for all $j = 1, \dots, q_i$). Let n be the sample size and denote the $n \times p$ observed data matrix of Y by $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_p)$ that is centered for each column to have mean 0.

Following the notations in [Bareinboim and Pearl \(2012\)](#), we denote the interventional values on variables $X^{(i)}$ by $\hat{\mathbf{x}}^{(i)} = (\hat{\mathbf{x}}_1^{(i)}, \dots, \hat{\mathbf{x}}_{q_i}^{(i)})$. The sufficient conditions for interventions on $X^{(1)}, \dots, X^{(p)}$ to be surrogate experiments require that for each $i \in V$, $X^{(i)}$ has no direct effect on variables Y_j for $j \in V \setminus \{i\}$ ([Pearl, 2000](#)). In this study, we assume there is no hidden variable, and thus this assumption is testable. We did not pursue such testing because we restrict the eQTLs to be cis-eQTLs, and cis-regulatory elements in human genome are rarely shared across genes.

2.2. Background: inductive causation algorithm

We briefly review the framework of the inductive causation algorithm (ICA). Using observational data, the Markov equivalence class of a DAG is identifiable. Two DAGs are Markov equivalent, if and only if they have the same *skeleton* and the same *v-structure* ([Andersson and others, 1997](#)). The skeleton of a DAG $\mathcal{G}$ is an undirected graph that is obtained by removing the directions of all the edges in the DAG. We denote the skeleton of $\mathcal{G}$ by $\mathcal{G}^u = (V, E^u)$, where $Y_i - Y_j \in E^u \Leftrightarrow Y_i \rightarrow Y_j \in E$ or $Y_j \rightarrow Y_i \in E$. A v-structure is an ordered triplet of vertices (Y_i, Y_j, Y_k) , such that $\mathcal{G}$ contains the directed edges $Y_i \rightarrow Y_k$ and $Y_j \rightarrow Y_k$, and Y_i and Y_j are not adjacent in $\mathcal{G}$. Given a Markov equivalence class, a directed edge is *compelled* if this edge exists in every DAG in the equivalence class; it is *reversible* otherwise. The edges participating in v-structures of a DAG are compelled edges. A Markov equivalence class can be represented by a *completed partially directed acyclic graph* (CPDAG), which is a partially directed acyclic graph (PDAG) consisting of directed edges for all compelled edges in the equivalence class, and undirected edges for all reversible edges in the equivalence class ([Chickering, 2002](#)).

The ICA to estimate a CPDAG from observational data consists of three steps: (1) estimation of the skeleton $\mathcal{G}^u$, usually by conditional independence testing, (2) v-structure identification, and (3) completion of the PDAG obtained from (1) and (2) to obtain the CPDAG ([Pearl, 2009](#)). Both Steps (2) and (3) are deterministic. More specifically, at Step (2), a triplet (Y_i, Y_k, Y_j) is assigned a v-structure $Y_i \rightarrow Y_k \leftarrow Y_j$, if $Y_i - Y_k \in E^u$, $Y_k - Y_j \in E^u$, $Y_i - Y_j \notin E^u$, and Y_k is not included in the conditioning set that makes Y_i and Y_j independent. The completion in Step (3) is to orient the undirected edges as much as possible, with restrictions of no directed cycles and no extra v-structures, which can be done by applying the rules in [Meek \(1995\)](#).

In our proposed method, we directly use the surrogate intervention for the edge orientation of the skeleton estimated by the Step (1). For the estimation of the skeleton, the PC algorithm ([Spirtes and others, 2000](#); [Kalisch and Bühlmann, 2007](#)) starts from a completely connected graph where all pairs of genes are connected, and iteratively removes edges by testing zero partial correlations, starting from zero-order (marginal) up to $n - 3$ order, with sample size n with a significance level α . The PC algorithm requires an exhaustive search, especially for edges that are present in the skeleton; the corresponding pair of variables are tested for the conditional independences given all possible subsets of the other variables. Thus, estimating high-dimensional DAGs without prior knowledge on their structures is a challenging problem. The PenPC algorithm ([Ha and others, 2016](#)) for DAG skeleton estimation employs a neighborhood selection method to choose a Markov blanket of each vertex, and then a greatly reduced number of partial-correlation tests are performed to remove false positive edges between co-parents of v-structures. The PenPC algorithm improves computational efficiency over the PC algorithm in the following areas: (1) PenPC reduces the search space within the selected Markov blanket for the subsequent partial-correlation tests; (2) the neighborhood selection approach that converts the GGM into a more tractable node-wise multiple regression model enables parallel computing, and the computational time can be substantially reduced, if multicore processors are available; and (3) for the second stage of PenPC, we further reduced the number of partial correlation tests to declare an edge in the skeleton by removing redundant conditioning sets for testing. Moreover, for the neighborhood selection, we employed the log penalty $p(|b|; \lambda, \tau) = \lambda \log(|b| + \tau)$ ([Mazumder and others, 2011](#)), which significantly improves the

accuracy of the Markov blanket search for high-dimensional problems (Sun and others, 2010), and we provided theoretical justifications of the estimation consistency of high-dimensional DAG skeletons (Ha and others, 2016).

2.3. Edge orientation given surrogate experiments

Now we consider orienting the edges in the skeleton $\mathcal{G}^u$. To simplify the notations, we assume each variable Y_k is causally affected by a set of external variables $X^{(k)}$, while $X^{(k)}$ can be an empty set. Given the external variables, the factorization of the joint probability density becomes

$$f(Y_1, \dots, Y_p) = \prod_{k=1}^p f(Y_k | Y_{\text{pa}_k}, \hat{\mathbf{x}}^{(k)}), \quad (2.1)$$

where $\hat{\mathbf{x}}^{(k)}$ includes q_k fixed intervention values. We will orient edges based on this augmented likelihood function. Each conditional distribution in the product term of equation (2.2) can be expressed by the following linear regressions: given $\mathcal{G}$,

$$\mathbf{y}_k = \mathbf{y}_{\text{pa}_k} \boldsymbol{\beta}_k + \hat{\mathbf{x}}^{(k)} \boldsymbol{\alpha}_k + \boldsymbol{\epsilon}_k, \quad k = 1, 2, \dots, p, \quad (2.2)$$

where $\mathbf{y}_{\text{pa}_k}$ is $n \times |\text{pa}_k|$ sub-matrix of $\mathbf{Y}$ corresponding to the parent set of Y_k from $\mathcal{G}$, $\boldsymbol{\beta}_k$ is the corresponding regression coefficients, $\boldsymbol{\alpha}_k$ is the regression coefficients for the intervention values $\hat{\mathbf{x}}^{(k)}$, and $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \sigma_k^2 \mathbf{I})$. We assume linearity for relations between genes and eQTLs, because both PC-stable and PenPC algorithms are based on linearity assumption between a node and its parents, and the linear relations between gene expressions and eQTLs are often assumed as reasonable approximations in eQTL studies (Kendzioriski and Wang, 2006; Ritchie and others, 2015).

To orient the skeleton $\mathcal{G}^u = (V, E^u)$, we aim to estimate the posterior probability of $Y_i \rightarrow Y_j$ versus $Y_j \rightarrow Y_i$ for an undirected edge $Y_i - Y_j$ in E^u by Bayesian model averaging (Hoeting and others, 1999). Suppose $G_t = (V, E_t)$ is a DAG that can be constructed from the skeleton $\mathcal{G}^u$ by adding directions. Let pa_k^t be the set of parents of Y_k in graph G_t , which can be an empty set if Y_k has no parent in graph G_t . Given G_t , let $\boldsymbol{\theta}^t$ be the all parameters including $\boldsymbol{\theta}_k^t = (\boldsymbol{\beta}_k^t, \boldsymbol{\alpha}_k, \sigma_k^2)$ for all $k \in V$ in the model (2.2). Denoting $\mathbf{D}$ as the data including $\mathbf{Y}$ and $\{\hat{\mathbf{x}}^{(1)}, \dots, \hat{\mathbf{x}}^{(p)}\}$, and letting $I(\cdot)$ be an indicator function,

$$\begin{aligned} P(Y_i \rightarrow Y_j | \mathbf{D}) &= \sum_{t=1}^T P(Y_i \rightarrow Y_j | G_t, \mathbf{D}) P(G_t | \mathbf{D}) \\ &= \sum_{t=1}^T I(Y_i \rightarrow Y_j \in E_t | G_t) P(G_t | \mathbf{D}), \end{aligned} \quad (2.3)$$

where $G_t = (V, E_t)$ for $t = 1, \dots, T$ are all possible DAGs, given the skeleton $\mathcal{G}^u$. Given G_t , $P(Y_i \rightarrow Y_j | G_t, \mathbf{D})$ is either 0 or 1, thus $P(Y_i \rightarrow Y_j | G_t, \mathbf{D}) = I(Y_i \rightarrow Y_j \in E_t | G_t)$. The posterior of a DAG, G_t is computed by

$$P(G_t | \mathbf{D}) = \frac{P(\mathbf{D} | G_t) P(G_t)}{\sum_{t=1}^T P(\mathbf{D} | G_t) P(G_t)} = \frac{P(\mathbf{D} | G_t)}{\sum_{t=1}^T P(\mathbf{D} | G_t)},$$

assuming all the G_t 's have the same prior probability. This is reasonable, since they all have the same skeleton. The marginal probability can be expressed as the following integral:

$$P(\mathbf{D}|G_t) = \int P(\mathbf{D}|G_t, \boldsymbol{\theta}^t) P(\boldsymbol{\theta}^t) d\boldsymbol{\theta}^t. \quad (2.4)$$

Instead of using the marginal probability in (2.4), we replace $P(\mathbf{D}|G_t)$ by $P(\mathbf{D}|G_t, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum a posteriori (MAP) estimate under uniform priors, $p(\boldsymbol{\theta})$ (i.e., maximum likelihood estimate). Therefore,

$$\begin{aligned} P(Y_i \rightarrow Y_j | \mathbf{D}) &= \frac{\sum_{t=1}^T I(Y_i \rightarrow Y_j \in E_t | G_t) P(\mathbf{D}|G_t, \hat{\boldsymbol{\theta}}^t)}{\sum_{t=1}^T P(\mathbf{D}|G_t, \hat{\boldsymbol{\theta}}^t)} \\ &= \frac{\sum_{t=1}^T I(Y_i \rightarrow Y_j \in E_t) \prod_{k=1}^p L_n(Y_k | Y_{\text{pa}_k^t}, \hat{\mathbf{x}}^{(k)}, \hat{\boldsymbol{\theta}}_k^t)}{\sum_{t=1}^T \prod_{k=1}^p L_n(Y_k | Y_{\text{pa}_k^t}, \hat{\mathbf{x}}^{(k)}, \hat{\boldsymbol{\theta}}_k^t)}, \end{aligned} \quad (2.5)$$

where

$$L_n(Y_k | Y_{\text{pa}_k^t}, \hat{\mathbf{x}}^{(k)}, \hat{\boldsymbol{\theta}}_k^t) = \exp \left\{ -\frac{n \log \hat{\sigma}_k^2}{2} - \frac{\|\mathbf{y}_k - \mathbf{y}_{\text{pa}_k^t} \hat{\boldsymbol{\beta}}_k^t - \hat{\mathbf{x}}^{(k)} \hat{\boldsymbol{\alpha}}_k\|^2}{2 \hat{\sigma}_k^2} \right\}.$$

As an alternative approach, we consider using a conjugate normal-inverse gamma prior for the parameters $\boldsymbol{\theta}_k^t = (\boldsymbol{\beta}_k^t, \boldsymbol{\alpha}_k, \sigma_k^2)$. The model in (2.2) is re-expressed as: for each node k ,

$$\mathbf{y}_k | G_t, \mathbf{b}_k, \sigma_k^2 \sim N(\mathbf{z}_k^t \mathbf{b}_k^t, \sigma_k^2 \mathbf{I}),$$

where $\mathbf{z}_k^t = (\mathbf{y}_{\text{pa}_k^t}, \hat{\mathbf{x}}^{(k)})$ and $\mathbf{b}_k^t = (\boldsymbol{\beta}_k^t, \boldsymbol{\alpha}_k^T)^T$. Let $\kappa_k = 1/\sigma_k^2$. We consider normal-gamma priors: for each k ,

$$\begin{aligned} \mathbf{b}_k^t | g, \sigma_k^2 &\sim N(\mathbf{0}, g \sigma_k^2 (\mathbf{z}_k^{tT} \mathbf{z}_k^t)^{-1}) \\ \kappa_k &\sim \text{gamma}(\nu_k/2, \delta_k/2), \end{aligned} \quad (2.6)$$

where g , ν_k , and δ_k are fixed positive values. The variance structure of $\mathbf{b}_k^t$ is defined according to Zellner's g-prior. A larger g provides more uncertainty that the coefficients are zero, and we employ the unit information prior by setting $g = n$. The marginal probability in (2.4) under the normal-inverse gamma priors can be derived as

$$\begin{aligned} P(\mathbf{D}|G_t) &= \prod_{k=1}^p P(\mathbf{y}_k | G_t) = \prod_{k=1}^p \int \int P(\mathbf{y}_k | G_t, \mathbf{b}_k, \sigma_k^2) P(\mathbf{b}_k | \sigma_k^2) P(\sigma_k^2) d\mathbf{b}_k d\sigma_k^2 \\ &= \prod_{k=1}^p (\pi)^{-n/2} \frac{\Gamma(\frac{\nu_k+n}{2})}{\Gamma(\frac{\nu_k}{2})} (1+g)^{-p_k^t/2} \frac{\delta_k^{\frac{\nu_k}{2}}}{(SSR_k^t + \delta_k)^{\frac{\nu_k+n}{2}}}, \end{aligned}$$

where $SSR_k^t = \mathbf{y}_k^T (\mathbf{I} - \frac{g}{g+1} \mathbf{z}_k^t (\mathbf{z}_k^{tT} \mathbf{z}_k^t)^{-1} \mathbf{z}_k^{tT}) \mathbf{y}_k$ and $p_k^t = |\mathbf{z}_k^t|$.

We call our methods, which use MAP and normal-gamma priors, `sirDAG-MAP` and `sirDAG-NG`, respectively. Using the marginal likelihood, $p(\mathbf{D}|G_t)$ in `sirDAG-NG` is the standard Bayesian approach for model selection and penalize the likelihood with respect to the increasing complexity of models. In our framework, however, we assume that the skeleton is sparse (i.e., the number of parents for each node is low), and a limited number of eQTLs with strong gene signals are used for inferring the directions of the skeleton. In simulation studies (Section 2.4), we compared the performance of those two approaches and found no significant difference, even when there were multiple eQTLs.

Finally, we use posterior probabilities of the two possible directions for an edge in $\mathcal{G}^u$. We select the direction $Y_i \rightarrow Y_j$, if the ratio $P(Y_i \rightarrow Y_j|D)/P(Y_i \leftarrow Y_j|D) > c$ for a cutoff $c \geq 1$, with a constraint of no directed cycle. Let $g_{ij} = P(Y_i \rightarrow Y_j|\mathbf{D}) = 1 - P(Y_j \rightarrow Y_i|\mathbf{D})$. The values $1 - g_{ij} = g_{ji}$ can be considered as estimates of the local false discovery rate (FDR) (Efron and Tibshirani, 2002), if the direction $Y_i \rightarrow Y_j$ is called a discovery. Given a desired FDR level $\eta \in (0, 1)$, the set of directed edges, $\mathcal{X}_{\phi_\eta} = \{Y_i \rightarrow Y_j : g_{ij} > \phi_\eta\}$, for $\phi_\eta = c/(1 + c)$ and $c \geq 1$ are called discoveries. The expected FDR for a given ϕ_η is

$$\eta = \frac{\sum_{(i,j)} g_{ji} I(g_{ij} > \phi_\eta)}{\sum_{(i,j)} I(g_{ij} > \phi_\eta)}.$$

2.4. Sampling DAGs

The model averaging across all possible T DAGs in equation (2.3) is not computationally feasible for a network skeleton with a large number of edges. For example, if there are 10 undirected edges in a skeleton, all possible DAGs have maximum number of 2^{10} directed structures without the acyclic assumption. Therefore, when there are a large number of undirected edges in the skeleton, instead of evaluating all the possible DAGs, we evaluate a large number of high-likelihood DAGs. To construct a high-likelihood DAG from a skeleton, we sequentially orient undirected edges one-by-one, starting from edges including eQTLs by posteriors evaluated under local graphs. Let the current PDAG for V be G_t . For an undirected edge $Y_i - Y_j$ in G_t , a directed local graph for $Y_i - Y_j$ is defined by $G_t(Y_i \rightarrow Y_j) : [X^{(i)} \cup Y_{\text{pa}_i^t}] \rightarrow Y_i \rightarrow Y_j \leftarrow [X^{(j)} \cup Y_{\text{pa}_j^t}]$. Denoting $\mathbf{D}_{ij}^t$ as the data including $\{\mathbf{y}_k : k \in \text{pa}_i^t \cup \text{pa}_j^t \cup \{i, j\}\}$ and $\{\hat{\mathbf{x}}^{(i)}, \hat{\mathbf{x}}^{(j)}\}$. The posteriors for edge directions are computed by

$$P(Y_i \rightarrow Y_j|\mathbf{D}_{ij}^t) = \frac{P(\mathbf{D}_{ij}^t|G_t(Y_i \rightarrow Y_j))}{P(\mathbf{D}_{ij}^t|G_t(Y_i \rightarrow Y_j)) + P(\mathbf{D}_{ij}^t|G_t(Y_i \leftarrow Y_j))},$$

where the priors are assumed to be $P(G_t(Y_i \rightarrow Y_j)) = P(G_t(Y_i \leftarrow Y_j)) = 0.5$. If two nodes are connected by an undirected edge, and at least one of these two genes has eQTLs, we refer to such an undirected edge as a *starting edge*. We identify a high-likelihood DAG in two steps:

Step 1. Set G_t as the skeleton. For each of the starting edges, update the direction of $Y_i - Y_j$ in G_t by randomly selecting a direction $Y_i \rightarrow Y_j$ or $Y_i \leftarrow Y_j$ with posterior probabilities $P(Y_i \rightarrow Y_j|\mathbf{D}_{ij}^t)$ and $P(Y_i \leftarrow Y_j|\mathbf{D}_{ij}^t)$, subject to acyclic constraints among the directions of all starting edges. Note that pa_i^t and pa_j^t are null sets, as G_t is the skeleton in the first step.

Step 2. Sample one of the nodes that have eQTLs. Orient the edges for the neighboring nodes (i.e., connected by undirected edges), take for example V_1 of the selected node, based on their posterior probabilities of $P(Y_i \rightarrow Y_j|\mathbf{D}_{ij}^t)$ and $P(Y_i \leftarrow Y_j|\mathbf{D}_{ij}^t)$, subject to acyclic constraint. Then set V_2 as the neighboring nodes of V_1 , and then do the same procedure. Repeat the same procedure until V_t is an empty set. This procedure is performed for each connected component of the skeleton.

Fig. 1. An example of a skeleton for $\{Y_1, Y_2, Y_3, Y_4\}$ with an eQTL $X_1^{(2)}$ for Y_2 .

As an illustrative example, a skeleton among 4 nodes $\{Y_1, Y_2, Y_3, Y_4\}$ with one eQTL, $X_1^{(2)}$ for Y_2 is displayed in Figure 1. The starting edges from the skeleton are, $Y_1 - Y_2$ and $Y_2 - Y_3$. In Step (1), we randomly sample the direction of each starting edge by their posteriors under the local graphs, $Y_1 \leftrightarrow Y_2 \leftarrow X_1^{(2)}$ and $X_1^{(2)} \rightarrow Y_2 \leftrightarrow Y_3$. If the orientation for the starting edges is $Y_1 \rightarrow Y_2 \rightarrow Y_3$, in Step (2), we orient edges $Y_1 - Y_4$ under the local graph $Y_1 \leftrightarrow Y_4$, and $Y_3 - Y_4$ under the local graph $Y_2 \rightarrow Y_3 \leftrightarrow Y_4$. The orientation $Y_3 \rightarrow Y_4 \rightarrow Y_1$ is not allowed as the acyclic constraint. For computing $P(\mathbf{D}_{ij}^t | G_t(Y_i \rightarrow Y_j))$, MAP and normal-gamma priors are used for *sirDAG*-MAP and *sirDAG*-NG, respectively. The QTL-directed dependency graph (QDG) (Neto and others, 2008) method uses an logarithm of the odds score to compare the two possible orientations based on the local graphs $G_t(Y_i \rightarrow Y_j)$ versus $G_t(Y_i \leftarrow Y_j)$, and iteratively updates the edge orientation until convergence without acyclic restriction. Thus, their algorithm is only dependent on local graphs, not the global graph structure. Although their local-graph based approach is similar to the procedure of sampling DAGs of *sirDAG*-MAP, the posterior $P(Y_i \rightarrow Y_j | \mathbf{D})$ from our framework is computed by averaging over the high-likelihood DAGs, weighted by their posterior model probability $P(G_t | \mathbf{D})$ (i.e., the global model posterior). The randomness of our DAG generation procedure comes from sampling directions of the starting edges in Step 1, weighted by their local posterior probabilities; the orientation in Step 2 can be different, depending on the node we start from, as we sequentially orient the edges by moving to the neighboring nodes based on local posteriors. In other words, using local posteriors may provide different DAGs, if we use a different node to start. Note that the number of DAGs that we can generate may be limited for the acyclic constraints. From a simulation study (Section S1), however, we found that *sirDAG* with different T provided stable estimated networks.

2.5. Simulation

We evaluate the performance of the *sirDAG* under various simulation settings by generating DAGs with different levels of sparsity, and single or multiple surrogate interventional variables for a gene. We simulate random DAGs following the Erdős and Rényi (1960) (ER) model, with a connection probability p_E , which is the probability that one vertex is connected to any other vertex and control for the sparsity of the graphs, and with a Barabási and Albert (1999) (BA) model with one edge to add in each time step (Ha and others, 2016). Denote the total number of eQTLs across p genes by $q = \sum_{i=1}^p q_i$, and the number of genes that have at least one eQTL by $|Z| = m$. We generate random DAGs, as well as gene expression and eQTL genotype datasets as follows:

1. Simulate the $n \times q$ eQTL genotype data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_q)$ that includes q random vectors distributed as $\text{multinomial}(n, (1 - p_m)^2, 2p_m(1 - p_m), p_m^2)$, where p_m is a predetermined minor allele frequency, and $X_i = 0, 1$, or 2 with probabilities $(1 - p_m)^2$, $2p_m(1 - p_m)$, and p_m^2 , respectively.

2. Construct a $p \times q$ matrix $\mathbf{A} = (\alpha_1^T, \dots, \alpha_p^T)^T$ and a $p \times p$ matrix $\mathbf{B} = (\beta_1^T, \dots, \beta_p^T)^T$ where all elements are zero. The former is regression coefficients of eQTL effect sizes and the latter is regression coefficients for gene–gene associations.
3. Generate a random graph from the ER model with the parameter p_E , or the BA model with one edge to add at each time. The graph structure determines the zero structure of $\mathbf{B}$, and the nonzero elements are replaced by independent realizations of random variable $2Z$, where $Z \sim \text{Exp}(\lambda = 5)$, an exponential distribution with parameter $\lambda = 5$. The sign of each nonzero element is randomly assigned with probability 0.5.
4. For randomly chosen m number of rows of $\mathbf{A}$, we assign the q eQTLs by randomly selecting nonzero elements and filling in values by realizations of random variable $2Z$, where $Z \sim \text{Exp}(\lambda(q))$ with parameter $\lambda(q)$ that depends on the number of total eQTLs across all p genes. The sign of each nonzero element is randomly assigned with probability 0.5.
5. Generate $n \times p$ gene expression data matrix, $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_p)$ by

$$\mathbf{y}_j = \mathbf{Y}\beta_j + \mathbf{X}\alpha_j + \epsilon_j,$$

where $\epsilon_j \sim \mathcal{N}(0, I_n)$. We simulate $\mathbf{y}_j$ for $j = 1, \dots, p$ sequentially, and after simulating each $\mathbf{y}_j$, we scale it so that it has variance 1 before simulating $\mathbf{y}_{j+1}$.

We fix the data dimension for the simulation study by $p = 700$ and $n = 300$ to match the dimension with the TCGA breast cancer data in Section 2.6. We consider eight methods in total for comparison: (1) ICA from PC-stable and PenPC algorithms, (2) the QTL-directed dependency graph (QDG) method (Neto and others, 2008), (3) sirDAG-MAP with skeletons from PC-stable algorithm, (4) sirDAG-MAP and sirDAG-NG with skeletons from PenPC algorithm, and (5) sirDAG-MAP and sirDAG-NG with skeletons from PenPC algorithm when randomly selected $m = 100$ genes have eQTLs. QDG employs a two-step approach to first estimate the skeleton using a PC-stable algorithm, and then orient the edges using eQTL data. A limitation of the QDG algorithm is that all genes in the network should have at least one eQTL ($p = m$), while sirDAG allows $p > m$. To accommodate the limitation of the QDG algorithm, and compare sirDAG and QDG, we generate the simulation datasets under $p = m$. Then in the methods (5), we randomly select $m = 100$ genes among $p = 700$ to have eQTLs. The PC-stable algorithm includes one tuning parameter, α , which is the significance level for the partial correlation tests. The PenPC algorithm has two steps for neighborhood selection, and a modified PC-stable algorithm. For the first step, the tuning parameters are selected by extended Bayesian Information criterion (Ha and others, 2016), and then the extra tuning parameter for the second step is α , which is the same tuning parameter in the PC-stable algorithm. For sirDAG-NG, we have hyper-parameters, v_k and δ_k in priors (2.6). Because we standardize the datasets before analyzing, we set $v_k = \delta_k = 1$ for all $k = 1, \dots, p$. For sirDAG, we decide the directions of edges based on the posterior probabilities by the ratio $P(Y_i \rightarrow Y_j | \mathbf{D}) / P(Y_i \leftarrow Y_j | \mathbf{D}) > c$, where c is a predetermined cutoff, and we set $c = 1$.

To evaluate the structural difference between an estimated graph and the true graph, we use the structural Hamming distance (SHD) that directly compares the structure of the learned network, including PDAG and CPDAG, and the original network that also include PDAG and CPDAG (Tsamardinos and others, 2006). SHD is defined by the number of the following operators required to make the two graphs match: add or delete an undirected edge, and add, remove, or reverse the orientation of an edge [Algorithm 4 in (Tsamardinos and others, 2006)]. For a true edge, $Y_1 \rightarrow Y_2$ as an example, $Y_1 - Y_2$, $Y_1 \rightarrow Y_2$, $Y_1 \leftarrow Y_2$, and no edge between Y_1 and Y_2 provide SHD values of 1 (add), 0, 1 (reverse), 1 (delete), respectively. Figure 2 displays the SHD between the estimated graphs and the true DAG as α values, that is the tuning parameter of PC-stable and PenPC algorithms and controls for the sparsity of skeletons. We consider four scenarios in Figure 2: (a) ER model with $p_E = 2/(p - 1)$ and $p = q = m = 700$ (all genes have single

Fig. 2. Structural Hamming distance (SHD) according to the significance level (α) for various simulation settings for $p = 700$, $n = 300$: (a) ER model with $p_E = 2/(p - 1)$, (b) ER model with $p_E = 4/(p - 1)$, (c) ER model with $p_E = 2/(p - 1)$ and multiple eQTLs ($q = 4000$ eQTLs are randomly assigned to the p nodes), and (d) BA model.

eQTL); (b) ER model with $p_E = 4/(p - 1)$ and $p = q = m = 700$; (c) ER model with $p_E = 2/(p - 1)$ and $p = m = 700$ and $q = 4000$ (all genes have at least one eQTL, i.e., multiple eQTLs); and (d) BA model and $p = q = m = 700$. We set $\lambda(q = 700) = 2$ and $\lambda(q = 4000) = 10$ in the data generation Step 4. Our simulation results are based on averages over 100 replications of the data generation Steps 1–5.

Comparison between PC-stable and PenPC. The CPDAG obtained from ICA with PenPC had lower SHD across the α values, except that PenPC showed slightly higher SHD in the case of BA model with $\alpha = 0.001$.

Comparison between QDG and sirDAG. Since the QDG algorithm uses the PC-stable algorithm for the skeleton estimation, we adjusted sirDAG to be comparable to QDG by using PC-stable for skeletons.

Across all simulation settings, the SHD values are significantly decreased by using our *sirDAG*, compared to using QDG. The reduction in SHD is highest when the networks are dense (Figure 2-b).

Comparison between sirDAG-MAP and sirDAG-NG. We have two different types of methods, *sirDAG-MAP* and *sirDAG-NG*, by the prior specifications. Across all simulation settings, the two methods showed similar performances, although *sirDAG-NG* showed slightly higher SHD than *sirDAG-MAP* when we used only $m = 100$ genes with eQTLs in the dense graph case (Figure 2b).

Comparison between QDG and sirDAG with $m = 100$. When we used the eQTLs for only $m = 100$ genes, the estimated graph was still more accurate than that achieved by other methods, including QDG, across all α values.

Overall, *sirDAG* with the PenPC algorithm showed the best performance across the various simulation settings, and *sirDAG-MAP* and *sirDAG-NG* showed similar accuracy in structure estimation. The main advantage of using *sirDAG* over the QDG algorithm is the applicability when only a subset of the genes in a network have eQTLs with accuracy gains. We also evaluated the performance of *sirDAG* when eQTLs for a gene are in linkage disequilibrium (LD), and our methods showed better performance than ICA algorithms (Section S1 of the supplementary material available at *Biostatistics* online).

2.6. Application to TCGA breast cancer data

Our method is applied to a breast cancer study in The Cancer Genome Atlas (TCGA). We use RNA-seq data from tumor tissues obtained from 550 female Caucasian patients, corresponding to the pathways PI3K/AKT, P53, cell cycle, apoptosis, mTOR, MAPK, RAS, and ERBB, which we selected based on literatures, TCGA breast cancer study (*Network and others*, 2012) and pan-cancer study (*Akbani and others*, 2014). We selected 764 genes that are included in at least one of the pathways using the KEGG database <http://www.genome.jp/kegg>. The expression of each gene within each sample is measured by the total read count (TReC), and we use the log-transformed TReC (logTReC) in this study. A 550×764 residual data matrix is obtained after removing the effects of several important covariates by linear regression: the 75th percentile of logTReC (which captures read depth), plate, institution, age, and six genotype principal components. Our goal is to estimate a network among those 764 genes that are included in the eight important pathways in cancer progression.

Applying the PenPC algorithm, we obtained the GGM (Step 1) with 13 061 undirected edges and the skeleton (Step 2) with 2255 undirected edges, using the p-value cutoff $\alpha = 0.001$. The degree of the skeleton ranged from 1 to 14 and NCK2 and RXRA genes had the highest degrees. Then we oriented the undirected edges in the skeleton using *cis*-eQTLs for 76 genes (10%), where we kept the most significant *cis*-eQTL per gene. The *cis*-eQTLs are identified by eQTL mapping using both TReC and ASE (*Sun*, 2012). After applying *sirDAG-MAP*, we oriented 2092 edges (93%). We evaluated the prognostic utility of the estimated network in selecting gene signatures. Using gene-specific causal modules, which consist of a gene and its parent set, based on the network from *sirDAG-MAP*, we identified which genes had a significant effect on patients' survival times. The unsupervised causal structural learning by *sirDAG-MAP* provides a PDAG. The main idea is that when we model a gene in relation to patient survival times, we adjust for its parent genes, which are obtained from the PDAG. If the DAG is known, the parent genes for all genes are obviously determined. Using the unique gene modules for each gene, we can find the effect of a gene on the survival time by including the expression of its parent genes in a Cox proportional hazards model. However, the remaining undirected edges in the PDAG generate uncertainty in determining the parent genes, because the undirected edges imply either direction. To obtain reasonable estimates of the causal effect, while accounting for such uncertainty, similar to the method of *Maathuis and others* (2009), we estimate the lower bound of the effect size of the gene on survival time. This lower bound is calculated by switching the directions for the undirected edges with constraints on no extra v-structure and directed cycle, and obtaining the minimum absolute value of the coefficients. In particular, if all neighboring genes

of i are directed, then we have an obvious module for gene i , and the module will be used to compute the prognostic effect of the gene using the Cox proportional hazards model, with the genes in the module as covariates. For genes with neighboring genes that have undirected edges to the gene, we compute multiple candidate modules for the gene in the equivalent class of the PDAG; this results in multiple prognostic effects for the gene, denoted by $\{\theta_i^1, \dots, \theta_i^m\}$ for gene i and m , DAGs that are in the Markov equivalence class of the PDAG. We use θ_i^k , where $k = \arg \min_j |\theta_j^i|$ for the prognostic gene ranking. The detailed algorithm for constructing multiple modules for a gene is described in [Maathuis and others \(2009\)](#).

The resulting PDAG from *sirDAG-MAP* is displayed in Figure 3, where the nodes are weighted and colored by the effect sizes, $|\theta_i^k|$ and signs, $\text{sign}(\theta_i^k)$. To evaluate the performance of the edge orientation, we compared *sirDAG-MAP*, which exploits the external *cis*-eQTLs, with the ICA using the PenPC algorithm (Section 2.2). The network obtained from *sirDAG* produced stronger effect sizes for 38 genes than that obtained from the ICA. Our method identified the collagen gene family, including types 1, 4, and 6, which have been shown to play important roles in breast cancer progression by altering the extracellular matrix architecture and composition (Barsky and others, 1983; Kauppila and others, 1998; Burnier and others, 2011; Fang and others, 2014). Table 1 shows the top 10 genes identified by *sirDAG-MAP*, the effect sizes, concordance indices (c-index) (Therneau and Grambsch, 2013) to test data using 10-fold cross-validation, and the parent sets from *sirDAG-MAP* and ICA. The signs of the effect sizes were consistently estimated between *sirDAG-MAP* and ICA, except for the top gene, COL4A2, which showed 1147.03% of the relative increase in effect size from *sirDAG-MAP* compared to ICA. Moreover, the c-indices of the *sirDAG-MAP* and the ICA for the gene COL4A2 were respectively 0.58 and 0.51. The superior prognostic power of our method for the gene COL4A2 comes from the different choices of directions for the neighboring genes THBS2 and COL4A1, which were selected as parents in the network from *sirDAG*, while they were selected as children of COL4A2 in the network from CPDAG. Similarly, the COL1A1 gene showed a 1230.65% increase in effect size by selecting the COL1A2 gene as the parent gene using *sirDAG-MAP*, while the signs were the same. Overall, *sirDAG-MAP* showed better prognostic power than ICA across the top genes, based on the c-indices evaluated from test data (Table 1).

3. DISCUSSION

Estimation of a DAG based on observational data is a challenging problem, because the conditional independence relations implied by the distribution satisfying the Markov property may represent several DAGs. We have developed a method to estimate the DAG when there is an additional set of variables, which are subject to interventions and are direct causes of the variables in the DAG. Simulation studies demonstrate the satisfactory performances of our method. We apply our method to construct a regulatory network from high-dimensional gene expression data, where we use genotype data of DNA polymorphisms as surrogate interventional data. Using the regulatory gene modules based on the graph structure using *sirDAG*, we showed the prognostic performance of our method.

Our method is based on the assumptions that gene expressions Y follows multivariate Gaussian distribution with no unobserved confounding variables. The faithfulness (general than Gaussian assumption) and no hidden confounders are fundamental assumptions, which are required for the identification of causal DAGs. Both assumptions are infeasible to check, however, we evaluated our approach in the framework of prognostic modeling using real data. For the external variables X , we impose no distributional assumption, but each DNA variant on a gene has no direct effect on other genes for a sufficient condition for admitting X as surrogate variables. This assumption on X implies that regulatory site-sharing between genes is not allowed, although genetic variants may shape multiple phenotypes (pleiotropy) (Tong and others, 2017).

Our method provides a flexible modeling framework by incorporating other types of DNA data, including methylation, copy number, and mutation as surrogate interventions for gene expressions. With the appropriately selected surrogate interventions using the biological hierarchy and filtering steps, *sirDAG* can be extended to other integrative modeling frameworks incorporating genomic, epigenomic, transcriptomic, and proteomic data.

4. SOFTWARE

Software in the form of R code, together with a sample input data set and complete documentation, are available at <https://github.com/MinJinHa/sirDAG>.

Table 1. Top 10 cancer genes (BRCA) we identified within the data set of 764 genes that are included in the 8 major pathways, ranked by the sirDAG-MAP model: effects on survival time, C-index (SD) to test data using 10-fold cross-validation and the parent genes, from sirDAG-MAP and ICA, and relative increase in effect size resulting from sirDAG-MAP compared to ICA

	Gene	sirDAG-MAP			ICA			$\frac{ a - b }{ b } \times 100$
		Effect ^a	C-index (SD)	Parents	Effect ^b	C-index (SD)	Parents	
1	COL4A2	-1.19	0.58 (0.036)	THBS2, COL4A1, FLNA	0.1	0.51 (0.038)	FLNA	1147.03%
2	LAMB1	-0.68	0.60 (0.082)	LAMA1, SEPT4, LAMA4	-0.68	0.60 (0.082)	LAMA1, SEPT4, LAMA4	0%
3	CDC20	-0.66	0.58 (0.016)	MKNK1, ORC1, PLK1	-0.74	0.53 (0.085)	MKNK1, HDAC1, CDK7, PLK1	-11.8%
4	COL6A2	0.61	0.57 (0.08)	COL6A1	0.61	0.57 (0.08)	COL6A1	0%
5	COL1A1	-0.61	0.52 (0.08)	COL1A2	-0.05	0.50 (0.09)		1230.65%
6	PKMYT1	-0.6	0.64 (0.04)	E2F2, RB1, PLK1	-0.73	0.54 (0.01)	E2F2, HSP90AB1, TSC2, RB1, SESN3, CHEK1, PLK1	-18.29%
7	CTSB	0.59	0.61 (0.08)	RRAGB, FNI, PIK3R3, BCL2A1, CTSS	0.61	0.58 (0.02)	CTSZ, FNI, PIK3R3, BCL2A1, CTSS	-3.8%
8	ORC6	0.58	0.62 (0.01)	CDC45, RBL2, FLNB, SIAH1	0.58	0.62 (0.01)	CDC45, RBL2, FLNB, SIAH1	0%
9	MAP3K12	0.52	0.58 (0.095)	LAMB2, SYNGAP1	0.49	0.52 (0.067)	EIF4B, PLA2G4C, MAPK8IP1, CACNA2D4, LAMB2, SYNGAP1	6.92%
10	BUB1B	0.51	0.57 (0.047)	CCNB2	0.26	0.52 (0.041)	BRCA1, ESPL1	99.65%

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

REFERENCES

- AKBANI, R., NG, P. K. S., WERNER, H. M. J., SHAHMORADGOLI, M., ZHANG, F., JU, Z., LIU, W., YANG, J.-Y., YOSHIHARA, K., LI, J. and others. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communications* **5**, 3887.
- ANDERSSON, S. A., MADIGAN, D. AND PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* **25**, 505–541.
- BARABÁSI, A.-L. AND ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.
- BAREINBOIM, E. AND PEARL, J. (2012). Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (pp. 113–120). AUAI Press.
- BARSKY, S. H., TOGO, S., GARBISA, S. AND LIOTTA, L. A. (1983). Type IV collagenase immunoreactivity in invasive breast carcinoma. *The Lancet* **321**, 296–297.
- BURGESS, S., SMALL, D. S. AND THOMPSON, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26**, 2333–2355.
- BURNIER, J. V., WANG, N., MICHEL, R. P., HASSANAIN, M., LI, S., LU, Y., METRAKOS, P., ANTECKA, E., BURNIER, M. N., PONTON, A. and others. (2011). Type IV collagen-initiated signals provide survival and growth cues required for liver metastasis. *Oncogene* **30**, 3766.
- CAI, X., BAZERQUE, J. A. AND GIANNAKIS, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology* **9**, e1003068.
- CHEN, L. S., EMMERT-STREIB, F., STOREY, J. D. and others. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology* **8**, R219.
- CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *The Journal of Machine Learning Research* **2**, 445–498.
- CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research* **3**, 507–554.
- COLOMBO, D. AND MAATHUIS, M. H. (2012). A modification of the PC algorithm yielding order-independent skeletons. *CoRR*, abs/1211.3295.
- DOSS, S., SCHADT, E. E., DRAKE, T. A. AND LUSIS, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research* **15**, 681–691.
- EFRON, B. AND TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.
- ERDŐS, P. AND RÉNYI, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61.
- FANG, M., YUAN, J., PENG, C. AND LI, Y. (2014). Collagen as a double-edged sword in tumor progression. *Tumor Biology* **35**, 2871–2882.
- HA, M. J., SUN, W. AND XIE, J. (2016). PenPC: a two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* **72**, 146–155.

- HAGEMAN, R. S., LEDUC, M. S., KORSTANJE, R., PAIGEN, B. AND CHURCHILL, G. A. (2011). A Bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics* **187**, 1163–1170.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. E. AND VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **1**, 382–401.
- KALISCH, M. AND BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research* **8**, 613–636.
- KAUPPILA, S., STENBÄCK, F., RISTELI, J., JUKKOLA, A. AND RISTELI, L. (1998). Aberrant type I and type III collagen gene expression in human breast cancer in vivo. *The Journal of Pathology* **186**, 262–268.
- KENDZIORSKI, C. AND WANG, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome* **17**, 509–517.
- KULP, D. C. AND JAGALUR, M. (2006). Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7**, 125.
- 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.
- LOGSDON, B. A. AND MEZEY, J. (2010). Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology* **6**, e1001014.
- MAATHUIS, M. H., KALISCH, M. AND BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* **37**, 3133–3164.
- MAZUMDER, R., FRIEDMAN, J. H. AND HASTIE, T. (2011). Sparsenet: coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106**, 1125–1138.
- MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., pp. 403–410.
- NETO, E. C., FERRARA, C. T., ATTIE, A. D. AND YANDELL, B. S. (2008). Inferring causal phenotype networks from segregating populations. *Genetics* **179**, 1089–1100.
- NETO, E. C., KELLER, M. P., ATTIE, A. D. AND YANDELL, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics* **4**, 320.
- CANCER GENOME ATLAS NETWORK. (2012). Comprehensive molecular portraits of human breast tumors. *Nature* **490**, 61.
- PEARL, J. (2000). *Causality: Models, Reasoning and Inference*, Volume 29. Cambridge University Press.
- PEARL, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.
- RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. AND SMYTH, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47.
- SCHADT, E. E., LAMB, J., YANG, X., ZHU, J., EDWARDS, S., GUHATHAKURTA, D., SIEBERTS, S. K., MONKS, S., REITMAN, M., ZHANG, C. and others. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717.
- SCHMIDT, M., NICULESCU-MIZIL, A. AND MURPHY, K. (2007). Learning graphical model structure using L1-regularization paths. Proceedings of the 22nd national conference on Artificial intelligence-Volume 2 (pp. 1278–1283). AAAI Press.
- SPIRITES, P., GLYMOUR, C. N. AND SCHEINES, R. (2000). *Causation, Prediction and Search*, Volume 81. MIT Press, Cambridge, MA.

- SUN, W. (2012). A statistical framework for eqtl mapping using RNA-seq data. *Biometrics* **68**, 1–11.
- SUN, W., IBRAHIM, J. G. AND ZOU, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* **185**, 349–359.
- SUN, W., YU, T. AND LI, K.-C. (2007). Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics* **23**, 2290–2297.
- THERNEAU, T. M. AND GRAMBSCH, P. M. (2013). Modeling survival data: extending the Cox model. Springer Science & Business Media.
- TONG, P., MONAHAN, J. AND PRENDERGAST, J. G. D. (2017). Shared regulatory sites are abundant in the human genome and shed light on genome evolution and disease pleiotropy. *PLoS Genetics* **13**, e1006673.
- TSAMARDINOS, I., BROWN, L. E. AND ALIFERIS, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* **65**, 31–78.

[Received December 5, 2017; revised November 18, 2018; accepted for publication November 25, 2018]