BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–9

Cyclic Regulatory Network Reconstruction from Genetic Perturbations

Benjamin A Logsdon [1] and Jason G Mezey [1,2,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853
[2]Department of Genetic Medicine, Weill Cornell Medical College, NY, NY 10021

ABSTRACT

Motivation: Discovery of novel regulatory relationships from the analysis of genome-wide expression data is a common goal of expression analysis. Among the methods applied to this problem are algorithms that make use of directed probabilistic graphs. To successfully infer regulatory relationships, these algorithms require perturbations of the network, produced by either experiments or naturally occurring genetic variation. While this critical point is well appreciated, there has been little theoretical work concerning the type of perturbations which provide the maximum resolution for arbitrary regulatory relationships.

Results: We derive the sufficient set of independent perturbations that provide maximum resolution for inferring directed cyclic networks. We also present the algorithm EXPLoRE (EXpression Penalized Likelihood Regulation Extractor), a method which makes use of our sufficient conditions to infer the direction of regulatory relationships in sparse networks. EXPLoRE works by incorporating perturbations of expression arising from *cis*-expression Quantitative Trait Loci (*cis*-eQTL) and, by identifying the signature of each *cis*-eQTL propagated through the regulatory network, the algorithm reconstructs directed relationships based on an inferred undirected graph. We demonstrate that this method can identify the correct regulatory relationships for both cyclic and non-cyclic cases. Using simulations, we also demonstrate that EXPLoRE performs well for sample sizes that are typical for genome-wide expression and genotype data. We also analyze expression and genotype data for individuals from the International HapMap Project and extract a putative regulatory network involving genes with roles in apoptosis, cell-signaling, and cancer.

Availability: The MATLAB code used for EXPLoRE is available at http://mezeylab.cb.bscb.cornell.edu/Software.aspx.

Contact: jgm45@cornell.edu

1 INTRODUCTION

Network analysis is commonly applied to genome-wide gene expression data to infer the set of regulatory relationships among genes (Chen *et al.*, 2008; Emilsson *et al.*, 2008). Probabilistic graphical techniques, which model genes as nodes and the conditional dependencies among genes as edges, are among the most frequently applied methods for this purpose. A diversity of such approaches have been proposed including Bayesian networks (Friedman *et al.*, 2000; Peer *et al.*, 2001), undirected networks (Margolin *et al.*, 2006; Schafer and Strimmer, 2005), and directed cyclic networks (Li *et al.*, 2006; Liu *et al.*, 2008; Neto *et al.*, 2008). The popularity of these methods derives, in part, from the structure of these models that is well suited to algorithm development. This is especially true for undirected networks (Kraemer *et al.*, 2009). In addition, the network representation of these models can be used to construct specific biological hypotheses about the processes governing the activity of genes in a system (Friedman *et al.*, 2000). As an example of this latter property, genes connected by an edge may indicate (at least) one of the genes is regulated by the other.

In graphical network inference, a theoretical principle that is now well appreciated (Schadt *et al.*, 2005; Rockman, 2008; Neto *et al.*, 2008; Liu *et al.*, 2008) is that 'perturbations' of the network can be leveraged to reduce the set of possible networks that can equivalently explain gene expression. Since equivalent models can indicate conflicting regulatory relationships, perturbations are often necessary to extract regulatory relationships with any confidence. Perturbations can be experimental or natural. The latter, for example, can be caused by genetic polymorphisms in a population, which alter the expression of genes across a population sample. These are considered expression quantitative trait loci (eQTL) (Rockman, 2008). Efficient algorithms have been proposed that can correctly resolve a sparse undirected graph (Margolin *et al.*, 2006; Kraemer *et al.*, 2009; Anjum *et al.*, 2009; Schafer and Strimmer, 2005). While useful, undirected graphs do not indicate the direction of regulation in a network. Such information is highly desirable when the goal is to infer which genes regulate which other genes and to develop specific biological hypotheses concerning the outcome of an experiment, e.g. if a gene is manipulated, what are the expected downstream effects?

For all directed graphical modeling, approaches for learning networks fall into one of two categories: 1) search through network space, and for each network, compute a score based on the fit of the network given the data (Friedman *et al.*, 2000; Schadt *et al.*, 2005; Li *et al.*, 2006; Liu *et al.*, 2008), or 2) identify all conditional independencies directly from the data, then use these conditional independencies to reconstruct a network (Spirtes *et al.*, 2001; Kalisch and Buhlmann, 2007; Richardson, 1996; Neto *et al.*, 2008). For the former, searches with reasonable coverage of the model space are not computationally feasible for larger networks.

*to whom correspondence should be addressed

While efficient searches for larger networks are possible with the latter, given a sparse graph, these methods tend to be very sensitive to sampling variation for tests of conditional dependence (Neto *et al.*, 2008). There is also a tendency for sampling variation to mask true conditional dependencies with these latter methods. To our knowledge, a limitation for both categories is there has been little to no theoretical work concerning the sufficient set of perturbations to allow inference of arbitrary directed graphs, which maximally limits the set of equivalent models. Limiting the set of equivalent models is of particular concern in cases where the true network has cyclic structure, where the set of statistically indistinguishable models may include drastically different topologies.

In this paper, we present theoretical results concerning a minimally sufficient set of perturbations to infer a maximally limited equivalent set of network architectures. We demonstrate that for a specific conditional independence graph (the interaction graph or moral graph (Lauritzen, 1996)), which can be efficiently estimated with many different approaches (Meinshausen and Buhlmann, 2006; Friedman *et al.*, 2008; Kraemer *et al.*, 2009; Anjum *et al.*, 2009; Schafer and Strimmer, 2005), the structure of this graph including an appropriate set of perturbations reflects the structure of the true regulatory graph. A consequence of this result is that, when such perturbation conditions apply, we do not have to consider the problem of identifying the Markov blanket or separating set for each node in the directed graph, to infer the set of edges (Pearl, 2000; Meinshausen and Buhlmann, 2006). In such cases, we can therefore avoid a computationally demanding step in the class of approaches which test for conditional independence and dependence (Friedman *et al.*, 2008). Our theoretical results are derived under the assumption that there are independent perturbations of the network, an assumption which seems reasonable, given recent biological observations of strong local polymorphism associations with gene expression (eQTL) which are often not in linkage disequilibrium (Stranger *et al.*, 2007; Doss *et al.*, 2005; Schadt *et al.*, 2003; Lum *et al.*, 2006).

Using our theory results, we develop a computationally efficient algorithm that is applicable to sample sizes that are typical of genome-wide gene expression data. Our algorithm includes three-steps. First, an association analysis is carried out to identify local (*cis*-eQTL) perturbations of gene expression. Second we pre-process to identify expression phenotypes with strong conditional dependencies using the 'glasso' undirected inference approach (Friedman *et al.*, 2008), which uses a lasso type penalty (Tibshirani, 1996) to estimate the conditional dependencies among the expression phenotypes. In the third step, the effects of *cis*-eQTL are used to further resolve the structure of the undirected graph, again using a lasso-type penalty. Based on our theory results, we can transform this second undirected graph directly into a system of regulatory relationships where we capture the direction of regulation. Our approach therefore mirrors directed network inference approaches that seek to identify conditional independencies (Kalisch and Buhlmann, 2007; Richardson, 1996; Neto *et al.*, 2008; Chu *et al.*, 2009). Given the components of our algorithm, we have named it EXPLoRE (EXpression Penalized Likelihood Regulation Extractor).

The effectiveness of our algorithm depends on the assumption of sparsity (Friedman *et al.*, 2008), as well as the presence of independent perturbations of the network. Sparsity is an essential assumption that is implicit in algorithms for both directed and undirected network inference algorithms (Margolin *et al.*, 2006; Kalisch and Buhlmann, 2007). The output of our algorithm is a directed graphical model that falls in a class known as structural equation models (SEMs), which assume Gaussian errors and allow for directed cyclic regulation among genes (Bollen, 1989).

2 THE DIRECTED GRAPHICAL MODEL

For p measured gene expression phenotypes and m loci for which we have genotypes, the directed graphical model of the network has $p + m$ nodes and $(p(p-1) + pm)$ possible edges, representing $p(p-1)$ possible regulatory relationships among the genes, and pm possible perturbation effects of loci (eQTL) on each of the expression phenotypes. Written in matrix notation, the network model for a sample of n individuals can be represented as:

$$\mathbf{Y}_{nxp}\mathbf{\Lambda}_{pxp} = \mathbf{X}_{nxm}\mathbf{B}_{mxp} + \mathbf{E}_{nxp}, \quad (1)$$

where $\mathbf{Y}$ is a matrix of gene expression measurements, $\mathbf{\Lambda}$ is a matrix of regulatory effects, $\mathbf{X}$ is a matrix of observed perturbations, $\mathbf{B}$ is a matrix of perturbation effects, and $\mathbf{E} \sim N(\mathbf{0}, \mathbf{R})$. Non-zero elements of $\mathbf{\Lambda}$ and $\mathbf{B}$ are edges representing regulatory relationships and eQTL effects, respectively, where the size of the parameter indicates the strength of the resulting relationship. We make the assumption that in the true network model, $\mathbf{\Lambda}$ and $\mathbf{B}$ are sparse. In addition, we assume that $\mathbf{R}$, the error covariance matrix of expression products, is diagonal, and $\mathrm{diag}(\mathbf{\Lambda}) = \mathbf{1}$, where the constraint on the diagonal of $\mathbf{\Lambda}$ ensures model identifiability. This constraint corresponds to a lack of self-loops, since the parameters representing self-loops are confounded with the error variance parameters specified by $\mathbf{R}$. These latter assumptions on $\mathbf{R}$ and $\mathbf{\Lambda}$ (i.e. no error covariance or self-loops) are standard, and used by all popular graphical network inference algorithms, directed and undirected, proposed to date (Friedman *et al.*, 2000; Schadt *et al.*, 2005; Li *et al.*, 2006; Liu *et al.*, 2008; Spirtes *et al.*, 2001; Kalisch and Buhlmann, 2007; Richardson, 1996; Neto *et al.*, 2008; Margolin *et al.*, 2006). The model depicted by Equation (1) is a completely observed structural equation model (SEM) (Bollen, 1989).

For a biological network, the model in Equation (1) is a reasonable representation, if we assume that gene expression measurements are taken from independent and identically distributed (*iid*) samples that have reached a steady-state (i.e. homeostasis), the specific system of regulatory relationships can be represented using a system of sparse linear equations, and the distribution of expression traits across samples is well modeled with a multivariate normal distribution. Given these assumptions of stationarity and normality, a linear representation of the system of regulatory relationships in Equation (1) is reasonable, since a linear transformation preserves multivariate normality. The representation of eQTL in Equation (1) is also reasonable, if we similarly assume that the set of perturbations from eQTL of unknown effect that are feeding into the gene expression phenotypes can also be modeled using a system of linear equations.

3 LIKELIHOOD AND EQUIVALENCE

The conditional log-likelihood of the model defined by Equation (1) can be written as:

$$l\left(\mathbf{Y}|\mathbf{X};\mathbf{\Lambda},\mathbf{B},\mathbf{R}\right)=\frac{1}{2}\log\left\{\det\left(\mathbf{\Sigma_{yy}}\right)\right\}-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma S}\right), \quad (2)$$

where the full precision matrix $\mathbf{\Sigma}$ and empirical covariance matrix $\mathbf{S}$ are:

$$\mathbf{\Sigma}=\left[\begin{array}{cc}\mathbf{\Sigma_{yy}} & \mathbf{\Sigma_{yx}}\\ \mathbf{\Sigma_{yx}^T} & \mathbf{\Sigma_{xx}}\end{array}\right]=\left[\begin{array}{cc}\mathbf{\Lambda R^{-1}\Lambda^T} & \mathbf{\Lambda R^{-1}B^T}\\ \mathbf{B R^{-1}\Lambda^T} & \mathbf{B R^{-1}B^T}\end{array}\right] \quad (3)$$

$$\mathbf{S}=\frac{1}{n}\left[\begin{array}{cc}\mathbf{Y^TY} & \mathbf{Y^TX}\\ \mathbf{X^TY} & \mathbf{X^TX}\end{array}\right], \quad (4)$$

with the data matrices $\mathbf{Y}$ and $\mathbf{X}$ re-centered.

We can now define a fully parameterized model matrix $\mathbf{\Gamma}$ as follows:

$$\mathbf{\Gamma}=\left[\begin{array}{c}\mathbf{\Lambda R^{-\frac{1}{2}}}\\ \mathbf{B R^{-\frac{1}{2}}}\end{array}\right], \quad (5)$$

since by definition $\mathbf{R}>\mathbf{0}$, and $\mathrm{diag}\left(\mathbf{\Lambda}\right)=\mathbf{1}$, both $\mathbf{\Lambda}$ and $\mathbf{B}$ can be rescaled by the positive square root of the error precision matrix $\mathbf{R^{-1}}$.

From Equation (3) and Equation (4) the relationship between the fully parameterized model matrix $\mathbf{\Gamma}$, and the full precision matrix $\mathbf{\Sigma}$ is:

$$\mathbf{\Gamma\Gamma^T}=\mathbf{\Sigma}. \quad (6)$$

This defines a system of homogenous polynomials of degree two which exactly specifies the relationship between the directed graph $\mathbf{\Gamma}$, which may contain no cycles (a directed acyclic graph or DAG) or may contain cycles (a directed cyclic graph or DCG), and the moralized undirected graph $\mathbf{\Sigma}$.

A potential pitfall of modeling expression traits using directed networks is the problem of likelihood equivalence between models. Figure 1 presents a simple example that illustrates the problems raised by equivalence for network inference. In this example, the true model, which is a linear pathway between four genes $w \rightarrow x \rightarrow y \rightarrow z$, is probabilistically indistinguishable from three other equivalent models. Each of these equivalent models has a very distinct implication for regulatory relationships among these genes but they are indistinguishable, regardless of the sample size. To be able to distinguish between these models, one needs to either collect time-course data to determine the temporal sequence in which regulation occurs, or alternatively, perturb the expression level of these genes in some fashion. A definition of equivalence based on the form of Equation (2) follows from (Pearl, 2000):

Definition of equivalence: Two sparse directed cyclic graphs specified by the model in Equation (1), with parameterizations $\mathbf{\Gamma_1}$ and $\mathbf{\Gamma_2}$, are equivalent in distribution iff for all parameterizations $\mathbf{\Gamma_1}, \exists \mathbf{\Gamma_2} : \mathbf{\Gamma_2\Gamma_2^T}=\mathbf{\Gamma_1\Gamma_1^T}$ and for all parameterizations $\mathbf{\Gamma_2}, \exists \mathbf{\Gamma_1} : \mathbf{\Gamma_1\Gamma_1^T}=\mathbf{\Gamma_2\Gamma_2^T}$.

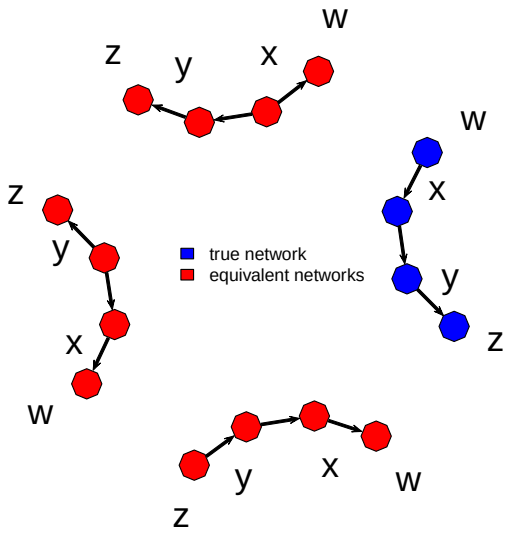

Fig. 1. Example of an equivalence class when determining regulatory relationships without perturbations (eQTL). In this case, the true regulatory network connecting the four genes (blue) has the same sampling distribution as the other three incorrect models (red), and these are therefore indistinguishable.

4 SUFFICIENT PERTURBATIONS

Given the importance of having as small a set of equivalent models as possible for making meaningful inference, and the necessity of perturbations for minimizing equivalence classes, it is of interest to know what will constitute a sufficient set of perturbations, i.e. to shrink the size of arbitrary equivalence classes as much as possible. In this section we motivate a minimal class of perturbation architectures that define a unique directed acyclic graph (DAG), or unique equivalence class of directed cyclic graphs (DCG), that can be reconstructed from the empirical covariances in Equation (4). We do this using three theorems. Theorem 1 is used to define a linear operator that allows transformations between models that produce the same precision (inverse covariance) structure, and can include transformations between models which are not faithful (i.e. models that have pathological parameterizations, where a richly parameterized model behaves like a reduced parameterized model). This equivalent model operator is used in Theorem 2, which demonstrates that with a sufficient set of perturbations, there are no DAGs that have equivalent models, and the equivalence classes for DCGs only contain models with reversed directed cycles. This theorem defines a sufficient set of perturbations as including at least one independent perturbation per expression phenotype (i.e. in genetic terms, this means no pleiotropy). It is important to note that a sufficient set can contain additional perturbations of these phenotypes, as long as there exists at least one perturbation for each

phenotype that is not pleiotropic.

Theorem 3 demonstrates how the set of equivalent DCGs can be recovered from the precision matrix between expression phenotypes and loci (the matrix $\mathbf{\Sigma_{yx}}$). This last result is incorporated into our EXPLoRE algorithm for inferring sparse network structure with a sufficient perturbation (eQTL) set. Note that while the EXPLoRE algorithm depends on sparsity for efficient network recovery, the results of these theorems are general and do not require such a constraint. In addition, we note in a further **Lemma** that even in the case of directed cycles, if we know which phenotype a perturbation feeds into, we can further reduce the size of the equivalence class to a unique directed cyclic graph.

Theorem 1: Given two distribution equivalent directed cyclic graphs, with equivalent parameterizations $\mathbf{\Gamma_1}$ and $\mathbf{\Gamma_2}$, any matrix $\mathbf{A}$ which satisfies $\mathbf{\Gamma_1 A = \Gamma_2}$, must be orthonormal (i.e. $\mathbf{AA^T = I}$).

Proof: Since $\mathbf{\Gamma_1 A A^T \Gamma_1^T = \Gamma_2 \Gamma_2^T}$, and from the definition of equivalence, if $\mathbf{\Gamma_1}$ and $\mathbf{\Gamma_2}$ are equivalent, then $\mathbf{\Gamma_1 \Gamma_1^T = \Gamma_2 \Gamma_2^T}$. Therefore, $\mathbf{\Gamma_1 A A^T \Gamma_1^T = \Gamma_1 \Gamma_1^T}$. Left multiply by $\mathbf{\Gamma_1^T}$ and right multiply by $\mathbf{\Gamma_1}$, then $\mathbf{CAA^TC = CC}$, where $\mathbf{C = \Gamma_1^T \Gamma_1}$ is a positive definite invertible matrix of rank p. Left and right multiply by $\mathbf{C^{-1}}$, and $\mathbf{AA^T = I}$.

The matrix $\mathbf{A}$ can be thought of as a linear operator that allows transformations between models which produce the same covariance structure (even between models which are not faithful). We use this operator to prove the following theorem after rescaling the network and perturbation parameters as in Equation (5): $\mathbf{\Lambda_i = \Lambda_i R_i^{-\frac{1}{2}}}$, $\mathbf{B_i = B_i R_i^{-\frac{1}{2}}}$:

Theorem 2: If there exists an ordered set $S = \{s_1, s_2, \ldots, s_p\}$ of rows of the perturbation graph parameterized by $\mathbf{B_1}$ such that $\mathbf{L_1 = B_1^{(S)} P_1}$, where $\mathbf{L_1}$ is a diagonal matrix of rank p and $\mathbf{P_1}$ is a signed permutation matrix, then 1) if $\mathbf{\Lambda_1}$ parameterizes a DAG, then for any parameterization $\mathbf{\Lambda_1}$ of any DAG, there does not exist an alternative equivalent DAG or DCG , and 2) if $\mathbf{\Lambda_1}$ parameterizes a DCG, then for any parameterization of any DCG, there exists a finite set of equivalent DCGs, where each equivalent DCG contains a reversed directed cycle with reference to the original DCG.

Proof: Given $\mathbf{L_1}$ exists, assume there exists an alternative equivalent model parameterized by $\mathbf{B_2}$ and $\mathbf{\Lambda_2}$. Then, by Theorem 1, there exists an orthogonal matrix $\mathbf{A}$ where $\mathbf{\Lambda_1 A = \Lambda_2}$, $\mathbf{B_1 A = B_2}$, and $\mathbf{L_1 A = L_2}$. Because $\mathbf{L_1}$ and $\mathbf{L_2}$ are invertible, we have: $\mathbf{A = L_1^{-1} L_2}$. This implies that $\mathbf{L_1 L_1^T = L_2 L_2^T}$. Since $\mathbf{L_1}$ is diagonal for any parameterization $\mathbf{B_1}$, $\mathbf{L_1 L_1^T}$ and $\mathbf{L_2 L_2^T}$ must also be diagonal for all equivalent parameterizations $\mathbf{L_1, L_2}$. If $\nexists$ a signed permutation matrix $\mathbf{P_2}$ such that $\mathbf{F = L_2 P_2}$, with $\mathbf{F}$ diagonal, then there always exists a parameterization of $\mathbf{L_2}$ where $\mathbf{L_2 L_2^T}$ is not diagonal, and therefore not equivalent (since all non-zero elements of $\mathbf{L_2}$ are free to vary). Therefore $\mathbf{A = P_2^T}$ is either an identity matrix or a signed permutation matrix. Now consider $\mathbf{\Lambda_1 A = \Lambda_2}$. Because in this parameterization, $\mathrm{diag}\,(\mathbf{\Lambda}) = \mathrm{diag}\left(\mathbf{R^{\frac{1}{2}}}\right)$, the only allowable equivalent model transformations must have positive non-zero elements along the entire diagonal. Therefore, if $\mathbf{\Lambda}$ parameterizes a DAG, then $\mathbf{A = I}$, and if $\mathbf{\Lambda}$ parameterizes a DCG, then $\mathbf{A = P}$ where $\mathbf{P}$ is any signed permutation matrix which ensures non-zero positive elements along the diagonal of $\mathbf{\Lambda}$. This corresponds directly to reversing the order of any set of directed cycles in the graph.

We now define the set of parents of a particular node, y_i, from the directed graph as $pa(y_i)$, and the set of all nodes in an undirected graph $\mathbf{\Sigma}$ that have edges to node z as $adj(\mathbf{\Sigma}, z)$.

Theorem 3: If in $\mathbf{\Sigma}$ there exists an independent perturbation vertex set $x = (x_1, \ldots, x_q)$ and a response vertex set $y = (y_1, \ldots, y_q)$ where $\forall i, |adj(\mathbf{\Sigma_{yx}}, y_i)| \geq 1$ and $\exists x_j \in pa(y_i)$, then the only equivalent directed cyclic graphs among y contain permutations of cycles, and can be recovered from $\mathbf{\Sigma_{yx}}$.

Proof: The existence of an independent perturbation vertex set and response vertex set that satisfies these conditions corresponds directly to a perturbation topology and parameterization specified by $\mathbf{L_1}$. Given this observation, Theorem 2 ensures the constraint on possible equivalent models. Finally, the reason the structure can be recovered from $\mathbf{\Sigma_{yx}}$ is apparent from Equation (3) and (5), where $\mathbf{\Sigma_{yx} = \Lambda B^T}$, and therefore $\mathbf{\Sigma_{yx}^{L_1} = \Lambda L_1^T}$ Since $\mathbf{L_1^T}$ is diagonal it won't change which elements of $\mathbf{\Sigma_{yx}^{L_1}}$ are zero or non-zero.

Lemma: If the underlying perturbation topology, $\mathbf{B_1}$, is known, then the cardinality of all directed cyclic equivalence classes is reduced to one.

Proof: This further reduction of the equivalence relationships is apparent when one considers that each equivalent perturbation topology specifies exactly one member of the equivalence class (from Theorem 3). Therefore, if one knows the true perturbation topology, then one knows the true regulatory model.

5 THE EXPLoRE ALGORITHM

Our algorithm EXPLoRE is constructed by making direct use of the implications of Theorem 3. This theorem indicates that for hypothetical studies of unrestricted sample size, we could develop an algorithm that simultaneously incorporates all expression phenotypes with putative cis-eQTL from a genome-wide gene expression and genotype data set, and directly generate a regulatory network for thousands of expression traits and genotypes. For realistic sample sizes, we have developed EXPLoRE, which scales efficiently to 30-50 phenotypes.

We begin with a screening process to identify a set of expression traits with putative *cis*-eQTL (Step 1). We then make use of the R package 'glasso' Friedman *et al.* (2008) in Step 2 to identify genes with strong undirected conditional dependencies. For step 3, we use the 'cvx' package (Grant *et al.*, 2009) to solve the convex optimization problem for the lasso penalized conditional Gaussian-Graphical Model that includes both genotypes and phenotypes. This package was used because additional constraints on the structure of the full undirected graph $\mathbf{\Sigma}$ (described below) are incorporated to ensure the assumptions required by Theorem 3 are enforced. The advantage of this approach is it defines a convex optimization procedure for shrinkage estimation, which will efficiently filter

out conditional dependencies of weak or zero-effect. This leads to higher power to detect true non-zero dependencies, when assuming a sparse graph and small sample sizes (Friedman *et al.*, 2008; Kraemer *et al.*, 2009). While we could make use of any undirected inference approach that infers the conditional independence graph (Kalisch and Buhlmann, 2007; Richardson, 1996; Neto *et al.*, 2008; Chu *et al.*, 2009) for Steps 2 and 3 we use a lasso penalized Gaussian-Graphical Model (GGM) (Friedman *et al.*, 2008).

The entire EXPLoRE algorithm includes three steps:

Step 1: Selection of expression phenotypes: A standard genome-wide association analysis is performed on each expression trait, focusing on genetic polymorphisms in a *cis*-window around a gene (e.g. a 2Mb window) (Stranger *et al.*, 2007). Each marker is tested individually using either a linear statistical model or non-parametric test statistic (e.g. Spearman rank-correlation), with a correction for multiple tests using either a control of false discovery rate (Benjamini and Hochberg, 1995), a conservative Bonferroni correction (i.e. α/n, where α is the significance level and n is the number of tests), or through a permutation approach to compute significance based on the empirical distribution of test statistics after shuffling the data, as in Stranger et al. (Stranger *et al.*, 2007).

Step 2: Expression interaction network reconstruction: Once the set of expression phenotypes are identified, we estimate a sparse interaction graph among phenotypes $\mathbf{\Theta} = \mathbf{\Sigma_{yy/xx}}$. This corresponds to learning the moralized graph among phenotypes without conditioning on genotypes: $\mathbf{\Sigma_{yy/xx}} = \mathbf{\Lambda^T Q^{-1} \Lambda}$ where $\mathbf{Q} = \frac{1}{n}\mathbf{B^T X^T X B} + \mathbf{R}$ is a covariance matrix combining the effects of shared genetic architecture between expression phenotypes and error covariance among phenotypes (which is assumed to be zero). Assuming that the eQTL effects are not correlated (i.e. $\mathbf{X^T X}$ is diagonal), then both $\mathbf{Q}$ and $\mathbf{Q^{-1}}$ will be diagonal, and therefore the set of interactions observed in the moral graph among phenotypes will be the same in $\mathbf{\Sigma_{yy}}$ and $\mathbf{\Sigma_{yy/xx}}$. Studies of genome-wide gene expression and genotypes tend to have severely constrained sample sizes, such that even when selecting a set of expression phenotypes in Step 1, the number of expression phenotypes being considered will often greatly exceed the sample size, or $p >> n$. Fortunately, a new class of estimators for precision matrices have been proposed when $p >> n$, which use a penalty on the precision matrix itself (Friedman *et al.*, 2008; Kraemer *et al.*, 2009).

$$\arg\max_{\mathbf{\Theta \geq 0}} : \log\left\{\det\left(\mathbf{\Theta}\right)\right\} - \mathrm{Tr}\left(\mathbf{\Theta S_{yy}}\right) - \lambda||\mathbf{\Theta}||^1. \quad (7)$$

The objective function defined in Equation (7) is a convex semi-definite program which can be solved using interior point methods (Vandenberghe *et al.*, 1998; Friedman *et al.*, 2008). We use 'glasso', a package in R (Ihaka and Gentleman, 1996), which is based on the cyclic descent algorithm proposed by (Friedman *et al.*, 2008). The hyperparameter λ is chosen based on five-fold cross validation.

Step 3: Regulatory network reconstruction: Once a subset $E = \{e_1, e_2, \ldots, e_k\}$ of k expression phenotypes with strong interactions is identified based on the non-zero structure of $\mathbf{\Theta} = \mathbf{\Sigma_{yy/xx}}$, the final step in the EXPLoRE algorithm is to infer a regulatory network among these expression traits. To do this we use the following constrained convex semi-definite program for the full phenotype-genotype precision matrix (as in Equation (3)), with the subset E of phenotypes and corresponding set S of *cis*-eQTL: $\mathbf{\Sigma} \equiv \mathbf{\Sigma^{E,S}}$:

$$\arg\max_{\mathbf{\Sigma \geq 0, \Sigma_{xx} \equiv D}} : \log\left\{\det\left(\mathbf{\Sigma_{yy}}\right)\right\} - \mathrm{Tr}\left(\mathbf{\Sigma S}\right) - \lambda||\mathbf{\Sigma_{yx}}||_1, \quad (8)$$

with $\mathbf{D}$ a diagonal matrix. Within EXPLoRE, we constrain $\mathbf{B}$ to be diagonal, which forces the eQTL to be direct perturbations of single expression traits to which the are *cis*-eQTL. This is the same as writing $\mathbf{\Sigma_{xx}}$ as a diagonal matrix and ordering the phenotypes such that they match the order of their respective cis-eQTL genotypes. This constraint directly represents the assumption that all effects of *cis*-eQTL must feed through a regulatory pathway (i.e. none of these genetic markers have *trans* effects that are not explainable by a regulatory pathway). As with the previous expression interaction network, $\mathbf{\Theta}$, the tuning parameter λ is chosen based on five-fold cross validation from Step 2. To solve this program we us the convex optimization package 'cvx' (Grant *et al.*, 2009), implemented in MATLAB. The best-fitting (from cross-validation) non-zero structure of $\mathbf{\Sigma_{yx}}$ represents the the regulatory network structure that best explains the propagation of the effect of the *cis*-eQTL through the network, as shown in Theorem 3.

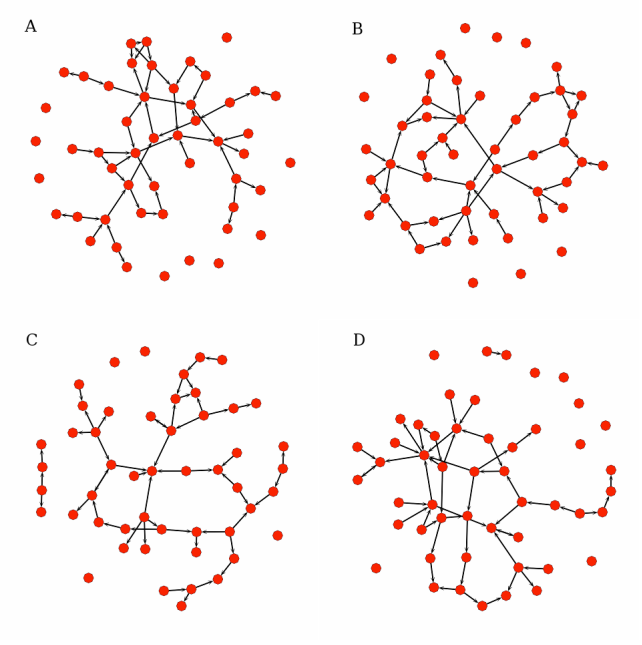

Fig. 2. Network topologies used to simulate gene expression data. Nodes represent expression levels of genes and the directed edges represent regulatory (conditional) relationships among genes, where the strength of the relationships were determined by sampling from $N(0,1)$, and the error variance sampled from $\Gamma^{-1}(1,2)$. Each phenotype (node) has a unique, independent cis-eQTL feeding into into it (not shown), with effects sampled from $\Gamma(1,2)$

6 SIMULATION ANALYSES

To assess the performance of EXPLoRE for sample sizes typical of genome-wide gene expression and genotype studies, we simulated a set of 50 expression phenotypes with 50 edges for $n = 60, 120, 300, 1000$, for the four sparse regulatory networks shown in Figure 2. We simulated weak to strong known unique *cis*-eQTL sampled from $\Gamma(1,2)$, weak to strong regulatory effects sampled from $\mathrm{N}(0,1)$, and independent error variances sampled from $\Gamma^{-1}(1,2)$. Five replicate simulations were performed for each simulated network topology.

When considering all non-zero parameters returned by our analyses, since the lasso can shrink parameters to exactly zero (Tibshirani, 1996), the observed false-discovery rate was unacceptably high, even for very harsh shrinkage (>60% for $\lambda >> 0$, results not shown). This is not surprising given that the graphical lasso penalty is not model selection consistent for arbitrary interaction network topologies and parameterizations (Meinshausen and Buhlmann, 2006). Since Fisher's z-transform of partial correlations (Kalisch and Buhlmann, 2007) cannot be applied, because the shrinkage estimation violates the assumptions required for this test statistic, we incorporated a thresholding procedure on the empirical partial correlations for both $\mathbf{\Theta}$ and $\mathbf{\Sigma_{yx}}$ to control the false-discovery rate. The empirical partial correlations computed for a generic positive semi-definite precision matrix $\mathbf{\Pi}$ are defined as: $\mathbf{\Pi} \odot l^{\mathrm{T}} l, l = \sqrt{\mathrm{diag}\,(\mathbf{\Pi})}$, with $\odot$ denoting the Hadamard product, or element-wise matrix multiplication. By thresholding we were able to control the false discovery rate to an arbitrary level, while ensuring acceptable power to detect regulatory relationships.

We show Receiver Operator Characteristic (ROC) curves demonstrating the performance of Steps 2 and 3 of EXPLoRE in Figure 3 for the topology shown in Figure 2a, which was a typical result across all topologies simulated. Figure 3a illustrates the reconstruction of the interaction network (Equation (7)) as a function of the empirical partial correlation threshold for multiple sample sizes (the shrinkage parameter λ was determined by cross-validation and set to 0.20). In Figure 3b, we show the ROC curves for the reconstruction of the regulatory network for multiple sample sizes, again using the cross-validated estimate of λ, where the set of phenotypes was determined by the nodes in the interaction graph $\mathbf{\Theta}$ that had significant edges as estimated using 'glasso' (Friedman *et al.*, 2008) (Step 2 of EXPLoRE).

As sample size grows, the performance of Step 3 increases drastically. However, even with smaller sample sizes, the method is still able to identify edges in the graph which have strong effects for moderate sample sizes, while adequately controlling for false-positives. This is shown in Figure 4, which presents the percentage of true positives identified from the true regulatory network for different corresponding absolute empirical partial correlations (from $\mathbf{\Sigma_{yx}}$), while controlling the false discovery rate for each simulation to 5%. As illustrated in Table 1, the total power is fairly low for $n = 60$ and $n = 120$, again when controlling the false discovery rate to 5% or 10%, but as sample size increases, the total power increases appreciably.

Table 1. Performance of EXPLoRE for simulated data, controlling false discovery rate (FDR)

	FDR=5%		FDR=10%	
sample size	power	ρ_e cutoff	power	ρ_e cutoff
n=60	0.084	0.648	0.112	0.608
n=120	0.172	0.675	0.328	0.486
n=300	0.504	0.344	0.516	0.325
n=1000	0.684	0.158	0.684	0.146

The empirical partial correlation ρ_e represents a scaled measure of the strength of the regulatory effect and its associated *cis*-eQTL effects (see text for details), with the cutoff used to control the FDR. Power represents the total percentage of true edges found in the full graph (Figure 2a) after Step 2 and Step 3 of EXPLoRE.

7 HAPMAP NETWORK ANALYSIS

To illustrate the usefulness of EXPLoRE, we analyzed genome-wide gene expression levels measured in eternal lymphoblastoid cell lines generated from the 270 individuals of Phase II of the International HapMap project (Stranger *et al.*, 2007). There are four distinct populations in this sample, Caucasian with European origin (CEU), Chinese from Beijing (CHB), unrelated Japanese from Tokyo (JPT), and Yoruba individuals from Ibadan Nigeria (YRI) (Frazer *et al.*, 2007). We focus on a subset of 858 phenotypes for 210 unrelated individuals, which Stranger et al. identified as having a *cis*-eQTL in at least one population. These *cis*-eQTL were identified by limiting the analysis to SNPs in a 2 Mb window around the transcription start site of each gene to control for multiple testing.

Figure 5 and Table 2 show the best expression regulatory networks, as determined by cross-validation, with $\lambda = 0.7$ and an empirical cutoff of the partial correlation of 0.20, for a sub-network of 15 genes. These genes were physically distant (i.e. in low linkage disequilibrium between *cis*-eQTL), had statistically significant *cis*-eQTL of strong effect, and were identified as having robust interactions from the initial screening using 'glasso'. We chose an empirical cutoff for the partial correlation of 0.20 based on our simulation results, which suggests that this is a reasonable choice given the sample size. The simulation study may not represent the exact conditions in the observed data (i.e. multiple populations with different error covariance structures and regulatory relationships), so this inferred network should be interpreted with caution.

Despite the preliminary nature of this analysis, the results are interesting and suggestive. There is a high representation of genes involved in cancer, cell-signaling, and apoptosis, as shown in Table 2. In particular, both RAB31 and RASSF6 are genes in the RAS oncogene family of genes (Bos, 1989), and they have a directed edge between them, though there is no previously demonstrated direct interaction between these particular genes. In addition, there are some major regulatory genes, including the RAS genes, PPARG, TNFRSF18, IAN4L1, and PASK, which is encouraging, since the goal of this analysis is to recover regulatory relationships among genes.

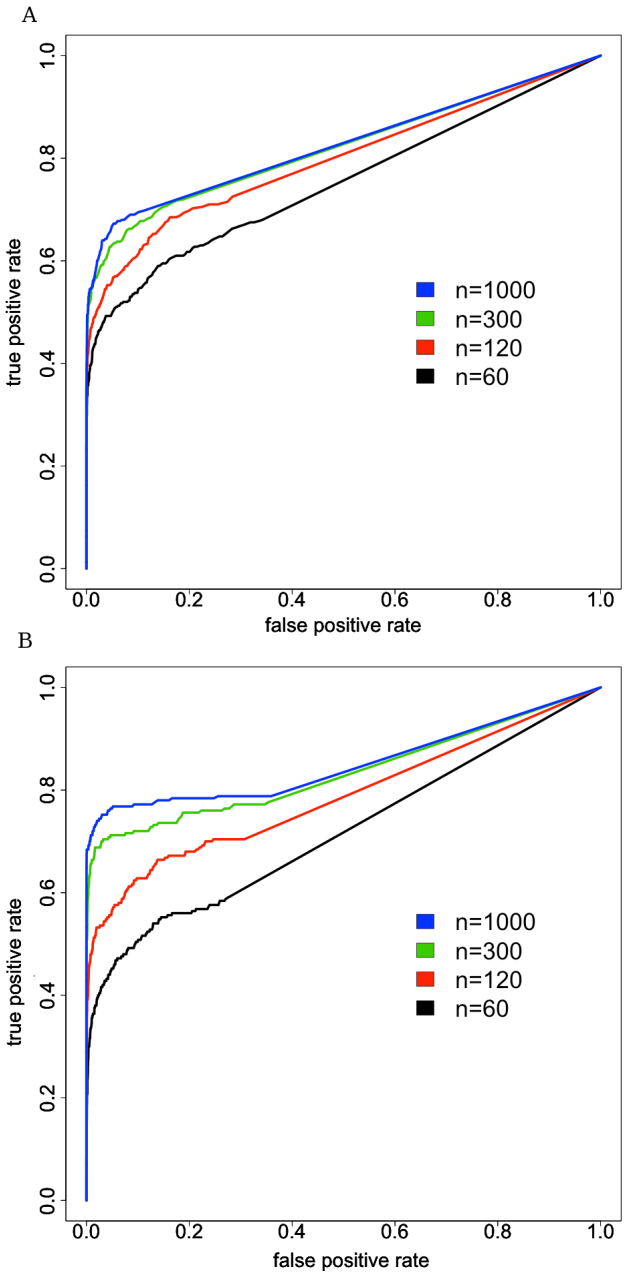

Fig. 3. Receiver Operator Characteristic (ROC) curves plotting the true positive rate (y-axis) versus the false positive rate (x-axis) for different sample sizes n for A) Step 2 of EXPLoRE inferring edges of the the interaction graph and B) Step 3 of EXPLoRE, inferring directed edges in the regulatory network reconstruction. These curves were generated by applying EXPLoRE to gene expression and genotype data simulated using the network models in Figure 2a.

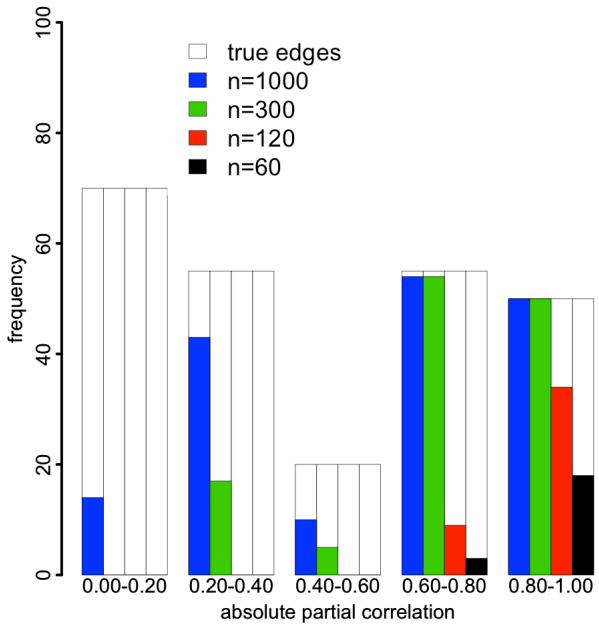

Fig. 4. Proportion of correctly identified regulatory edges as a function of absolute partial correlation ρ_e of the corresponding edge in $\mathbf{\Sigma_{yx}}$ identified by EXPLoRE for data simulated using the regulatory networks in Figure 2a while controlling FDR to 5%.

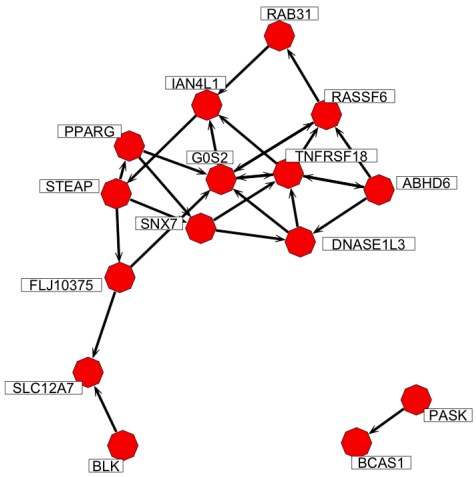

Fig. 5. A putative regulatory network operating in the immortalized cells harvested from individuals in HapMap. This network was inferred by applying EXPLoRE to gene expression data collected on these cell lines. Additional details concerning gene function and strength of the network relationships are presented in Table 2.

8 DISCUSSION

EXPLoRE is a novel methodology for identifying signatures of cyclic regulation from genome-wide gene expression and genotype data. This is the first algorithm that makes use of sufficient sets of *cis*-eQTL to infer unique directed cyclic networks from gene

Table 2. Biological functions and the strength of conditional relationships among genes in the network extracted from HapMap gene expression (Figure 5).

gene symbol	biological functions	children(ρ_e)
BCAS1	candidate oncogene	-
PASK	intracellular signaling regulator	BCAS1(-0.266)
RAB31	vesicle and granule targeting	IAN4L1(-0.241)
RASSF6	growth inhibitor and tumor suppressor	G0S2(-0.301),RAB31(-0.233)
TNFRSF18	immunological self tolerance and apoptosis regulator	RASSF6 (0.53), G0S2(-0.2634),ABHD6(0.293)
ABHD6	lipase and hyrdolase activity	RASSF6(-0.204), DNASE1L3(-0.252)
DNASE1L3	DNA hydrolase and apoptosis	TNFRSF18(0.312), G0S2(-0.305)
SNX7	intracellular trafficking	TNFRSF18(0.336), DNASE1L3(-0.323)
STEAP	prostate cancer antigen	PPARG(0.251),SNX7(-0.266) FLJ10375(-0.346)
PPARG	adipocyte differentiation regulator,tumorigenesis, metabolism	SNX7(-0.252),G0S2(-0.226)
FLJ10375	ion transport	G0S2(0.368), SLC12A7(-0.204)
SLC12A7	ion transport cell volume homeostasis	-
G0S2	cell cycle switch apoptosis	TNFRSF18(0.374), IAN4L1(0.583), RASSF6(-0.223)
IAN4L1	apoptosis, T-cell differentiation	STEAP(0.281)
BLK	protein kinase cascade	SLC12A7(0.332)

Gene symbols and biological functions accessed from http://www.ncbi.nlm.nih.gov/gene. The children are genes which are directly regulated based on the directed graph (Figure 5). The strength of the conditional relationship is the empirical partial correlation (ρ_e), a scaled measure of the strength of regulation and the strength of the *cis*-eQTL effect (see text for details).

expression data. EXPLoRE also provides an efficient strategy for network inference by learning the interaction network structure through the lasso procedure, which can then be transformed into a unique directed cyclic regulatory network. Since the underlying optimization program is convex, EXPLoRE provides significant efficiency advantages over many previously proposed algorithms for genome-wide cyclic regulatory network reconstruction (Liu *et al.*, 2008; Neto *et al.*, 2008). These previous approaches require either heuristic searches through regulatory network space with no guarantee to reach networks with the strongest evidence given the data, or lack sufficient perturbations to allow unambiguous regulatory inference.

A number of assumptions concerning biological networks are made when applying EXPLoRE. These include assumptions that are common to most graphical modeling techniques, such as sparsity, linearity of regulatory relationships, and normally distributed error, as well as an assumption that is specific to EXPLoRE: the presence of known, independent perturbations from *cis*-eQTL. The common assumptions are reasonable when constructing a first approximation to regulatory network structure. Regulatory relationships are not linear, but linearity is the simplest approximation that provides biologically relevant information, i.e. there is a detectable relationship between two genes, or no relationship. Given an observed covariance structure, normal distributions have maximum entropy (Wainwright and Jordan, 2008). A normal assumption is therefore conservative in terms of being the most 'random' distribution that could have generated the data. Given the absence of knowledge about the specific biological process generating the distribution of expression measurement error, and barring any clear evidence of non-normality in data, such a conservative approximation seems appropriate.

The assumption of independent, detectable *cis*-eQTL effects is perhaps the most restrictive assumption of EXPLoRE. We do not expect this to be a good approximation for all regulatory modeling situations. However, in the case of gene expression, a number of studies have found that eQTL with the largest (detectable) effects on expression, tend to be local to the gene they are affecting, i.e. *cis* (Brem and Kruglyak, 2005; Doss *et al.*, 2005). What's more, due to the structure of linkage disequilibrium in populations (the correlation structure among genotypes) it is often possible to identify a large set of *cis*-eQTL that are uncorrelated that each have unique expression phenotypes. For example, a set of eQTL that are present on different chromosomes or are far away from one another in terms of genetic map distance (Stranger *et al.*, 2007). The organization of eQTL may not therefore be badly approximated by a model of independent, local perturbation and, given the resolution advantages when these assumptions are met, an EXPLoRE analysis can be of value. We do note that, as with any graphical model analysis where the true nature of the regulatory network is unknown, the inferences extracted by EXPLoRE should be interpreted with caution.

As a final comment, the theory of sufficient perturbations that maximize regulatory resolution, which is used as the foundation of EXPLoRE, is actually far more general. Theorem 3 defines a class of perturbation architectures where there is a direct isomorphism between two very different types of graphs: the undirected graph with perturbations and a directed cyclic graph representing a regulatory network. The theory does not require perturbations to be *cis*, just that there be an appropriate set of perturbations that provide resolution. More complex perturbation sets, which include sufficient perturbations as a subset, can also provide maximum resolution. One could therefore construct algorithms similar to EXPLoRE without the *cis*-restriction. Moreover, the specific topology of eQTL effects need not be known, if one is willing to accept the cost of larger network equivalence classes and therefore less total regulatory resolution. With this restriction lifted, it would be possible to jointly infer the genetic perturbation architecture simultaneously with regulatory architecture, although such a joint reconstruction would require much larger sample sizes. We are currently working on these extensions for the EXPLoRE algorithm. In addition, we are working on a version of the EXPLoRE algorithm which uses the cyclic descent method of Friedman et al. (Friedman

et al., 2008; Kraemer *et al.*, 2009) to improve the scaling property of Step 3.

9 CONCLUSION

We have presented theory concerning the sufficient set of perturbations necessary to limit equivalence classes of directed cyclic graphs (DCG) to a unique network. This theory, in combination with existing penalized likelihood approaches for extracting sparse undirected graphs, is a strategy for unambiguously inferring regulatory relationships from gene expression and genotype data. We implement this strategy in the algorithm EXPLoRE. This is the first algorithm for directed network inference that does not rely on exhaustive scoring techniques, when leveraging sufficient perturbations to uniquely identify cyclic regulation. Our simulation and data analyses indicate that EXPLoRE performs well for realistic sample sizes and has the potential to extract regulatory networks responsible for observed patterns of gene expression.

10 ACKNOWLEDGEMENTS

We thank Larsson Omberg and Gabriel Hoffman for discussion and for their comments on this manuscript. This work was supported by a fellowship from the Center for Vertebrate Genomics at Cornell University and by National Science Foundation Grant DEB0922432.

REFERENCES

Anjum, S., Doucet, A., and Holmes, C. (2009). A Boosting Approach to Structure Learning of Graphs with and without Prior Knowledge. *Bioinformatics*, **25**(22), 2929.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Bollen, K. (1989). *Structural equations with latent variables*. Wiley. NY.

Bos, J. (1989). Ras oncogenes in human cancer: a review. *Cancer research*, **49**(17), 4682.

Brem, R. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(5), 1572.

Chen, Y., Zhu, J., Lum, P., Yang, X., Pinto, S., MacNeil, D., Zhang, C., Lamb, J., Edwards, S., Sieberts, S., *et al.* (2008). Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**(7186), 429.

Chu, J., Weiss, S., Carey, V., and Raby, B. (2009). A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Systems Biology*, **3**(1), 55.

Doss, S., Schadt, E., Drake, T., and Lusis, A. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research*, **15**(5), 681.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G., Gunnarsdottir, S., *et al.* (2008). Genetics of gene expression and its effect on disease. *Nature*, **452**(7186), 423–428.

Frazer, K., Ballinger, D., Cox, D., Hinds, D., Stuve, L., Gibbs, R., Belmont, J., Boudreau, A., Hardenbol, P., Leal, S., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**(3-4), 601–620.

Grant, M., Boyd, S., and Ye, Y. (2009). CVX: Matlab software for disciplined convex programming. *Available at http://www. stanford. edu/boyd/cvx*.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Kalisch, M. and Buhlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, **8**, 636.

Kraemer, N., Schaefer, J., and Boulesteix, A. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *Arxiv preprint arXiv:0905.0603*.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press. NY.

Li, R., Tsaih, S., Shockley, K., Stylianou, I., Wergedal, J., Paigen, B., and Churchill, G. (2006). Structural model analysis of multiple quantitative traits. *PLoS Genet*, **2**(7), e114.

Liu, B., de la Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**(3), 1763.

Lum, P., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., Drake, T., Lusis, A., and Schadt, E. (2006). Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *Journal of Neurochemistry*, **97**, 50.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**(Suppl 1), S7.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, **34**(3), 1436.

Neto, E., Ferrara, C., Attie, A., and Yandell, B. (2008). Inferring causal phenotype networks from segregating populations. *Genetics*, **179**, 1089–1100.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press. Cambridge, UK.

Peer, D., Regev, A., Elidan, G., and Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **1**(1), 1–9.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 454–61.

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

Schadt, E., Monks, S., Drake, T., Lusis, A., Che, N., Colinayo, V., Ruff, T., Milligan, S., Lamb, J., Cavet, G., *et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**(6929), 297–302.

Schadt, E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S., Monks, S., Reitman, M., Zhang, C., *et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, **37**(7), 710–717.

Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, prediction, and search*. The MIT Press. Boston, MA.

Stranger, B., Forrest, M., Dunning, M., Ingle, C., Beazley, C., Thorne, N., Redon, R., Bird, C., de Grassi, A., Lee, C., *et al.* (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**(5813), 848.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Vandenberghe, L., Boyd, S., and Wu, S. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM journal on matrix analysis and applications*, **19**, 499–533.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, **1**(1-2), 1–305.